EPO Academic Research Programme 2018

Approximating the Standard Essentiality of Patents –

A Semantics-Based Analysis

Final version from June 12, 2020.

Lorenz Brachtendorf
Fabian Gaessler
Dietmar Harhoff

Max Planck Institute for Innovation and Competition, Munich, Germany

Contents

Executive Summary

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 report introduces a semantics-based method for approximating the standard essentiality of patents. We briefly discuss the current state of the literature on semantic algorithms applied to patent text data and explain the peculiarities when using such algorithms to standards. We then provide details on the mechanics of our approach and the measures of semantic similarity between patent and standard texts. We assemble data on patent-standard pairs (either specifically declared or determined by our similarity measure) for three leading standard-setting organizations (SSOs) in the ICT industry: ETSI, IEEE, and ITU-T. We describe the content and structure of the generated database, which will be made publicly available, and present selected descriptives.

We demonstrate the method's internal and external validity through several exercises:

- 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 throughout a significantly higher semantic similarity for standard-patent pairs defined by SEP declarations.

- Second, we correlate our measure with different patent characteristics. In line with the general notion that truly standard-essential patents are of considerably high value, we find a strong and significant correlation between our measure of semantic similarity and established patent value indicators.

- Third, we exploit information on manual essentiality assessments for a sample of patents declared essential to either ETSI or IEEE standards. Again, we find strong and significant correlation between the experts' decisions on standard essentiality and our measure of semantic similarity.

In a first empirical application, we demonstrate that the similarity measure can be used to estimate the share of (presumably) true SEPs in firm patent portfolios. Doing so, we find statistically and economically substantial differences between firms. We further illustrate that our measure can be used to shed light on the number and identity of SEPs in those cases, where firms filed only blanket (i.e., unspecific) declarations.

Despite the method's limited accuracy, we see various possible use cases in the academic as well as practical sphere. Most importantly, the method may facilitate the large-scale assessment of declared SEPs and the search for relevant, but (so far) undeclared patents, rendering it a potentially valuable tool for SSOs, regulators, and firms alike.

Chapter 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. However, the successful development and adoption of standards depends on ex ante coordination among technology contributors and implementers – in particular, if proprietary technologies are to be incorporated (Lerner and Tirole, 2015). Standard-essential patents (SEPs) protect inventions that are part of technical standards and are by definition infringed whenever the respective standard is implemented. However, due to the vast amount of patents and uncertain patent scope, the identification of SEPs poses a considerable challenge to potential implementers. Standard-setting organizations (SSOs) rarely conduct searches for SEPs on their own. Instead, they demand from their members to timely disclose SEPs through declaration. The 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, in fact, protect a relevant contribution to the selected technological solution, i.e., are truly standard-essential. While there is empirical evidence suggesting that declared SEPs are relatively more valuable (Rysman and Simcoe, 2008), there are several factors beyond technical merit that may influence whether a patent is declared standard-essential.[1] Most notably, there are concerns that patents are declared to be SEPs due to strategic incentives of their holders, irrespective of the underlying technical quality and the relevance to the respective standard (Dewatripont and Legros, 2013).[2] 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, 2018, for an overview). In fact, standard essentiality frequently fails to survive scrutiny if the patent is disputed in court (Lemley and Simcoe, 2018). Uncertainty about the true relevance of a patent to a standard may introduce legal and contractual frictions, as it creates considerable transaction costs during the standardization process and subsequent licensing negotiations. Ensuring a fair and efficient framework to foster the devel-

[1] In this report, we focus on *technical* standard essentiality. We discuss different essentiality definitions in Chapter 2.

[2] 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.

opment and adoption of technical standards is a key goal of SSOs, which puts current intellectual property (IP) policies, particularly essentiality checks, into regulatory focus (EC, 2017).[3]

This report describes a semantics-based method to approximate the standard essentiality of patents, which facilitates the identification of systematic discrepancies between the declared and true standard essentiality of patents. This method relies on a novel measure of semantic similarity between patents and standards. In recent years, text-based measures have proven to be useful for the empirical assessment of patent similarity and technological relatedness (e.g., Arts et al., 2018; Natterer, 2016; Younge and Kuhn, 2016). So far, these large-scale applications have focused on texts within the patent universe. In contrast, we propose a method for a semantics-based comparison of patent texts and standard specifications. In several validation exercises, we show that the calculated similarity serves as 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 SEP declarations. Second, we find that semantic similarity strongly correlates with common patent value indicators, which constitutes additional support as true SEPs are considered to be high value. Finally, we benchmark our results against manually examined SEPs for several ICT standards (GSM, UMTS, LTE, 4G, WiFi, etc.). Based on these data, we confirm the predictive power of our similarity measure on patent level, and illustrate the generalizability of our method across different technologies.

As recent legal disputes have exemplified, the calculation of licensing fees for standard technologies often involves not just one SEP but whole portfolios. This demands scalable approaches to assess standard essentiality. 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.*" We therefore estimate, in a first empirical application of our method, the share of presumably true SEPs in firm patent portfolios for ETSI, IEEE, and ITU-T standards. We provide evidence for the high accuracy of our approach when predicting standard essentiality on an aggregate level. Our results show considerable firm-level differences in the estimated share of presumably true SEPs. These differences are statistically significant and economically substantial. Among all ETSI 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.

So far, economic and legal analyses regarding the relationship between patents and standards have had little choice but to take SEP declarations at face value.[4] Therefore, in introducing a new method to approximate standard essentiality, this report makes various contributions of academic as well as practical relevance. First, we illustrate how a semantics-based tool can be used to measure the essentiality of patents to specific technical standards. Second, while computationally demanding,

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

[4] Notable exceptions are the case studies of Goodman and Myers (2005) and, most recently, Stitzing et al. (2017), both drawing on manual assessments of declared SEPs by patent attorneys. 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).

this method is scalable, objective and replicable – opening up new avenues of empirical research in the context of 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 report is structured as follows: Chapter 2 surveys the prior literature and describes the relationship between patent rights and standards. Chapter 3 details the methodology of our semantics-based approach. Chapter 4 describes the database that is used in the subsequent analyses and will be made publicly available. Chapters 5, 6, and 7 each provide descriptive results validating the method for ETSI, IEEE, and ITU-T standards. A brief discussion and outlook on future use cases of our essentiality measure conclude the report.

Chapter 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, 2018). SSOs differ in various dimensions such as their 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 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 (binding) aspects, such as upfront patent searches, the disclosure content, as well as the disclosure timing, and may or may not be binding. For instance, some SSOs *demand* from 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 major 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. However, in practice, 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.[5] 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.[6] Consequently, the standard essentiality of a patent needs to be understood (and ultimately 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 patent grant, the patent's scope of protection may be narrowed as a result of patent 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, SEP declarations may represent a poor signal of true standard essentiality. The true standard essentiality of a patent typically remains private information held by the respective rights holder. Occasionally, however, a patent's true standard essentiality becomes public knowledge. First, results of standard essentiality assessments in the context of legal disputes are disclosed through court decisions.[7] SEP litigation usually deals with selected subsets of SEPs rather than with entire SEP portfolios or, let alone, all SEPs for a particular standard.[8] Second, true standard essentiality of patents can be inferred from SEP assessments by third parties, which do not occur within the context

[5] 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.

[6] 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.

[7] Although SEP litigation certainly takes place in Europe as well (cf. Contreras et al., 2017), the US remain the hotspot for SEP litigation. Lemley and Simcoe (2018) 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.

[8] 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.

of SEP lawsuits.[9] 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 standard essentiality, although 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 benefits. First and foremost, SEPs represent revenue-generating opportunities as all standard implementers become potential licensees. Furthermore, owning SEPs likely improves a firm's bargaining position in cross-licensing negotiations.[11] Hence, it seems reasonable to assume that firms follow various strategies to increase the chance of holding standard-relevant patents. In the first place, 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]

In the context of patent disclosure, firms usually enjoy some discretion in their decision whether they want to declare their patent as standard-essential (or not), irrespective of true essentiality. With no further assessment of SEP status, it stands to reason that an SEP declaration likely affects the patent's *perceived* essentiality for third parties. In this context, 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

[9] Notably, Stitzing et al. (2017) use a proprietary dataset on SEP assessments to study the characteristics of SEPs that were scrutinized and found to be standard-essential.

[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.

[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.

harsher punishments for patent holders if they do not disclose standard-essential patents rather than if they disclose 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 expectations 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 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). Depending on the rules of the SSO, firms may choose the level of detail concerning the disclosed information about relevant technologies and IP rights. Lerner et al. (2016) empirically investigate SEP declaration strategies and find that firms with major downstream businesses and low-quality patents prefer *blanket* (i.e., non-specific) disclosures.

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]

[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.

Chapter 3

Methodology

In this chapter, 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. Finally, we discuss several validation exercises.

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 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. Most 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.[17] In line with other text-mining algorithms, a vector space model is employed to calculate the similarity between two defined texts. The algorithm measures the semantic similarity between patents, but can also measure semantic similarity between patents and any other input text (such as scientific publications, wikipedia articles, etc.). The major advantage of this algorithm is the extremely efficient implementation which allows the comparison of any text to the patent universe and yields in a list with the most similar patents ranked by their similarity score.[18] 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.

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 3.1, 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.[19]

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 informative technology descriptions.[20] In Section 5.2, we provide evidence for the validity of patent-standard text similarity as a measure of technological similarity and ultimately standard essentiality.

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

[17] A non-exhaustive list of techniques incorporated in the algorithm includes part-of-speech tagging, spelling correction, n-grams, stop words, stemming techniques, entropy-based weighting, synonym dictionaries, and other relationships.

[18] 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.

[19] 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.

[20] 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.

Figure 3.1: 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})."

The used text-mining algorithm is proprietary, which renders some aspects of the similarity calculation non-transparent and 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 achieved with this open-source algorithm 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 Appendix C.

3.3 Similarity measures

In the subsequent analyses, we apply two different measures to approximate the true essentiality of a patent to a standard: 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*). Strongly correlated with each other, both measures can be used 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, the *similarity rank* provides the standard-specific order of the most similar patents. Both similarity measures are retrieved for the most similar 3,000 patent families for each standard document. Although this allows us to limit the amount of data, it also implies that we have to account for truncation (or censoring, respectively) when interpreting our results.

3.4 Validation strategies

In the following, we propose several validation strategies that we use to establish the explanatory value of our semantic similarity measure.

Control group comparison

We conduct several distinct validation exercises. 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.

The first step to validate our semantic approach involves a comparison of SEPs with patents describing technologies from the very same technology class. If our measure has any explanatory value, SEPs will be significantly more similar to the respective standard than the control patents. As discussed in previous sections, under scrutiny many declared SEPs may 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 the set of correctly declared and hence truly essential patents. The control group comparison with all SEPs thus renders the average difference in similarity a lower bound. We exploit the information that SEP declarations usually cite the respective standard. We call these predefined pairs of SEPs and standards simply *SEP declarations* and compare those to pairs of the same standard and undeclared patents from the same technology class and cohort.

Vice versa, to test the validity for the standard cited in the declaration, we keep the declared SEP fixed and compare the associated standard document to another randomly chosen standard document from the same standards project[21] and the same publication year as the focal standard.

Correlation with patent characteristics

Given that the underlying technologies are implemented into various products, standard-essential patents are considered to be of particularly high economic and technological value. We therefore estimate multivariate regressions of our semantic similarity measure on various patent characteristics that prior literature has established as meaningful value proxies. If the semantic similarity measure identifies truly essential patents, we would expect a positive correlation between the measure and these values proxies.

Benchmark against manual SEP assessments

Finally, we benchmark the similarity measure with manually examined SEPs to test the predictive power of the similarity measure to determine true standard essentiality.

For ETSI SEPs, we can draw on secondary data created in the context of a recent legal dispute on SEP royalties. 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.[22] The case concerned the calculation of royalty fees for SEPs, but also addressed the question 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

[21] We classify standard documents based on keywords occurring in the title of the standard document.

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

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 should be strongly correlated with true standard essentiality on an aggregate level.

We are not aware of any publicly or commercially available data including manual essentiality checks of IEEE (or ITU-T) SEPs. We therefore chose to create a benchmark dataset. We selected a random sample of all patent families declared to IEEE standards, mostly referencing WiFi or WiMAX standard specifications. In line with previous efforts to analyze SEPs in greater numbers (see Table B-1 in the Appendix for an overview), we rely on patent attorneys to assess standard essentiality. For this project, we recruited several Munich-based patent attorneys specialized in the fields of electrical engineering and computer science. Each patent attorney received a random set of at least 30 patent families declared to one or more standards.[23] This minimum threshold is necessary to minimize power concerns in the subsequent statistical analysis. To avoid selection issues, we further required that the patent attorneys complete the items in the order as stated in the provided evaluation sheet. All relevant patent and standard documents were provided as digital copy to the patent attorney. The task description reads as follows.

Instructions:

- *Please review the following patent documents and specify the relevant claims of the patent and the relevant sections of the corresponding standard specifications.*
- *Indicate whether a patent is, according to your judgment, standard-essential.*
- *We ask you to rate the respective patent's technical standard essentiality in the following probabilistic way: very likely essential, likely essential, unlikely essential, very unlikely essential.*
- *When assessing essentiality, please do not draw on any secondary information (court cases, reports, etc.), but stick to the documents provided.*

The patent attorneys conducted their assessments between November 2019 and January 2020. The final dataset includes 272 assessed patent-standard pairs based on 144 unique patent families.

[23] The patent attorneys received a fixed payment for each assessed item (i.e., the unique patent).

Chapter 4

Database Description

4.1 Standards and declared SEPs

ETSI

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 by ETSI or within the framework of the 3rd Generation Partnership Project (3GPP).[24] 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.[25]

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.[26] 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.

[24] 3GPP is a global network of seven standards organizations of which ETSI is one of the key organizations.

[25] 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.

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

The set of documents covers 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 at 194 (median: 84) is 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.[27] 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.

IEEE

The Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA) is a global standard-setting organization based in the US, which sets standards in various technology areas, such as telecommunications, robotics, health, power and energy. Some of the successful technologies that are applied worldwide and are used by billions of users are WiFi, WiMAX, WPAN, and Ethernet. These standards are set by both public and private organizations active in the ICT sector. The technical solutions supplied by the stakeholders are often, similarly to technologies contributed to ETSI standards, protected by IPR.

Due to the possibility to file blanket declarations, most relevant IPR is not disclosed and the number of specifically declared SEPs is relatively low, as compared to ETSI. We identify 961 patent families in the subset of specific declarations. In general, the information that can be retrieved from declarations at IEEE is much less comprehensive as compared to ETSI. Only about third of letters of assurance indicates patent numbers in the declaration. Although, declarants specify the relevant standardization projects for which they hold relevant IPR, this information is less detailed than the detailed information on technical specifications at ETSI. Nonetheless, we are able to relate declared SEPs to a subset of potentially referenced standard documents inferred from the overall standards project indicated in the declaration letter. The data on declarations for all IEEE standardization projects can be downloaded from the official IEEE-SA website.

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

We further received access to the full list of standard documents. The data comprise meta information as well as the full texts on 4,303 published as well as 4,003 draft standard documents. We focus on documents that have been published between 1922 and 2019. The documents are available as PDFs and most of them are directly machine-readable. Similarly to standards published by ETSI, these documents can be very large in size, comprising hundreds of pages. Whereas the average number of pages is 123 (median: 46), this number increases when focusing at the subset of WiFi related specifications. For these documents, the average number of pages is 535 (median: 177). Hence, we face similar issues as discussed for ETSI standards: traditional NLP approaches may not cope well with such large texts of technical descriptions. We therefore make use of the routine that we developed to automatically analyze the structure of ETSI standards documents in order to break down large texts into smaller parts, such as chapter, sections, subsections and so on. Although the structure of documents published at IEEE differs for some standards from ETSI documents, this approach works reasonably well. We obtain 37,911 unique chapters for 3,442 documents.

ITU-T

The standardization sector of the International Telecommunications Union (ITU-T) is responsible for international standards in the field of ICT. It is located in Geneva, Switzerland, and has developed numerous major technical standards since its establishment in 1865. These include JPEG, audio and video coding (e.g., H.264/MPEG-4) standards, VoIP, DSL, and many more. In general, the declaration policy is similar to some rules practiced at IEEE. Firms and other organizations often make use of blanket declarations, i.e., they contribute all of their developed technologies to the standardization process, but do not specify the particular patents. The data on standards and SEP declarations are publicly available. Information on IPR can be easily downloaded from the official ITU website. We identify 2,268 unique patent families. It should be noted that most relevant IPR is not disclosed in this database. Similarly to IEEE data, the information in declarations on particular referenced standard documents is not as detailed as in the case of ETSI.

To obtain information on ITU-T recommendations[28], we download all standard documents as PDFs and relevant meta information. We obtain 12,003 unique documents from the ITU-T website referring to recommendations published between 1958 and 2018. The average document size is smaller than for standards at other SSOs. Considering relevant standard specifications such as VDSL or H.264, however, these still comprise several hundreds of pages. The overall mean is 37 pages (median: 18). This is because extensions of existing specifications are published separately. We again apply our routine to obtain the structure of ITU-T recommendations. We end up with 73,445 unique chapters, which refer to 7,918 documents.

4.2 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

[28] Within ITU-T, standard specifications are referred to as recommendations.

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).[29] We retrieve information on patent families, technology classes, inventor team size, as well as detailed information on patent claims. We compute various forward and backward citation measures at patent family level that are needed for our validity checks.

4.3 Similarity data

For all three standard-setting organizations, we use the text of technical standards descriptions to calculate the semantic similarity between those standard documents and approximately 37 million patent documents from the patent database.

We identify all ETSI standards referenced in SEP declarations and end up with a set of 4,796 referenced standard documents. Using these data, 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 *document level*. The second dataset includes a more fine-grained comparison between patents and standards at *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 chapter texts results in 187,398,000 observations at patent-standard level.

Furthermore, we make use of all machine-readable and published IEEE standards and end up with the full text of 4,302 standard documents. Using the semantic algorithm, we generate a dataset with 12,906,000 pairs of patent families and IEEE standards. At *chapter* level, 3,371 documents are analyzed and separated into 37,746 unique chapters. This results in a dataset with 133,238,000 pairs of patent families and IEEE standards. Similarly, we use 11,117 out of 12,007 ITU-T documents to compute the text similarity between patents and the full text of standards. We obtain a dataset with 35,916,000 pairs of patent families and ITU-T standards. At the more fine-grained *chapter* level, we identify 66,127 chapters of 6,963 standard documents. In total, we obtain the two similarity measures for 210,509,452 patent-standard pairs.[30]

4.4 Entity relationship diagram

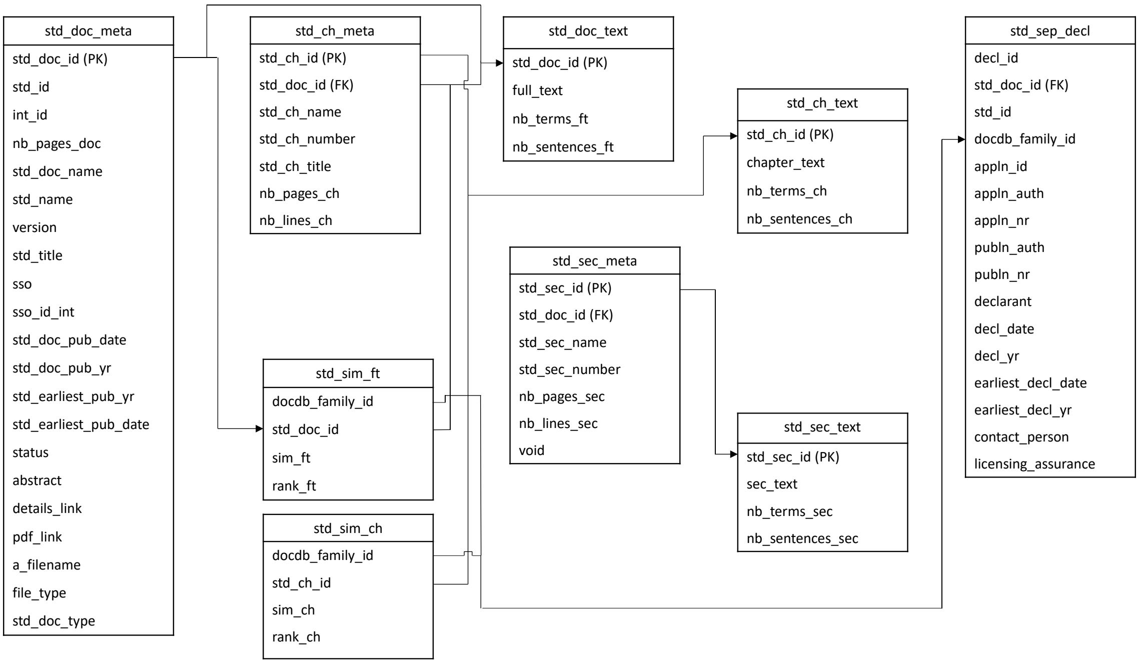
We create a database on patents and standards which is complementary to already existing databases on standards and standard-setting organizations. The main goal of this database is to provide text information on technical standards in an easily accessible and structured way. Data on all ETSI, IEEE,

[29] 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.

[30] For ITU-T recommendations and (to a lesser degree) for IEEE specifications, the structure of documents is not as stable as compared to ETSI documents. This leads to the lower share of successfully parsed documents for these SSOs.

and ITU-T standards are provided. The entity relationship diagram in Figure 4.1 shows variable names as well as the relationships between the tables. We provide detailed information on the variables in the Appendix. There are three tables that comprise meta information on standards at *document*, *chapter*, and *section* level.[31] Associated with these tables, there are three additional tables that contain the actual text of a standard. For standards at *document* and *chapter* level, we further provide the 3,000 semantically most similar patent families.[32] These tables can be linked to the actual standards to get further information on name, title, publication date, and other details. Apart from the sample of semantically similar patents, we further provide a table with information on declarations. For specific declarations, we identify, first, the patent family, which can be matched with the PATSTAT database, and second, the actual standard document, which can be linked to tables in this database.

Figure 4.1: Entity–relationship diagram



Notes: The table structure of standard documents and SEP declarations published at ETSI, IEEE and ITU-T is shown. The database includes meta information as well as the actual text on document, chapter and section level. Declaration data is included. For a large subsample of all texts, the most similar patent families are determined. Two tables on document and chapter level are provided.

[31] *Section* level refers to the most fine-grained level identified in the table of contents of a document.

[32] Due to the large number of texts, we restrict the sample of ETSI standards to those documents directly referenced in SEP declarations. This leaves us with 4,796 documents and 62,482 chapters. For the full text and chapters of IEEE and ITU-T standards, we provide the population.

Chapter 5

ETSI Descriptives and Estimates

In this chapter, we first describe the sample of patents relating to standard specifications published by ETSI and provide selected descriptive statistics. Moreover, we present several validation results and predictions for different telecommunication standards.

5.1 Sample description

In Table 5.1, 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 5.1a shows the similarity score distributions for all patents and the subset of all SEPs.[33]

In Figure 5.1b, 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 patents for the respective standard text. Similarly, in Figure 5.1c, the percentage of SEPs declared at ETSI is plotted against the rank reporting the samples of SEPs that are included in chapter as well as the full text datasets. 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. Notably, 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.

5.2 Validation results and predictions

In the following, we provide the results of several validation exercises.

[33] Likewise, Figure A-1 in the Appendix shows the similarity score distributions for all ESTSI standard documents.

Table 5.1: 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
SEP declarations	Score	285	92	0.941	69	720	9481
	Rank	877	871	8.945	1	3000	9481
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
SEP declarations	Score	339	100	0.815	74	735	15000
	Rank	877	896	7.316	1	3000	15000

Notes: Summary statistics for *similarity score* and *similarity rank* across three different datasets at document as well as chapter level. Minimum (maximum) possible score: 0 (1,000). Lowest (highest) possible rank: 3,000 (1).

Comparison of SEPs with control groups

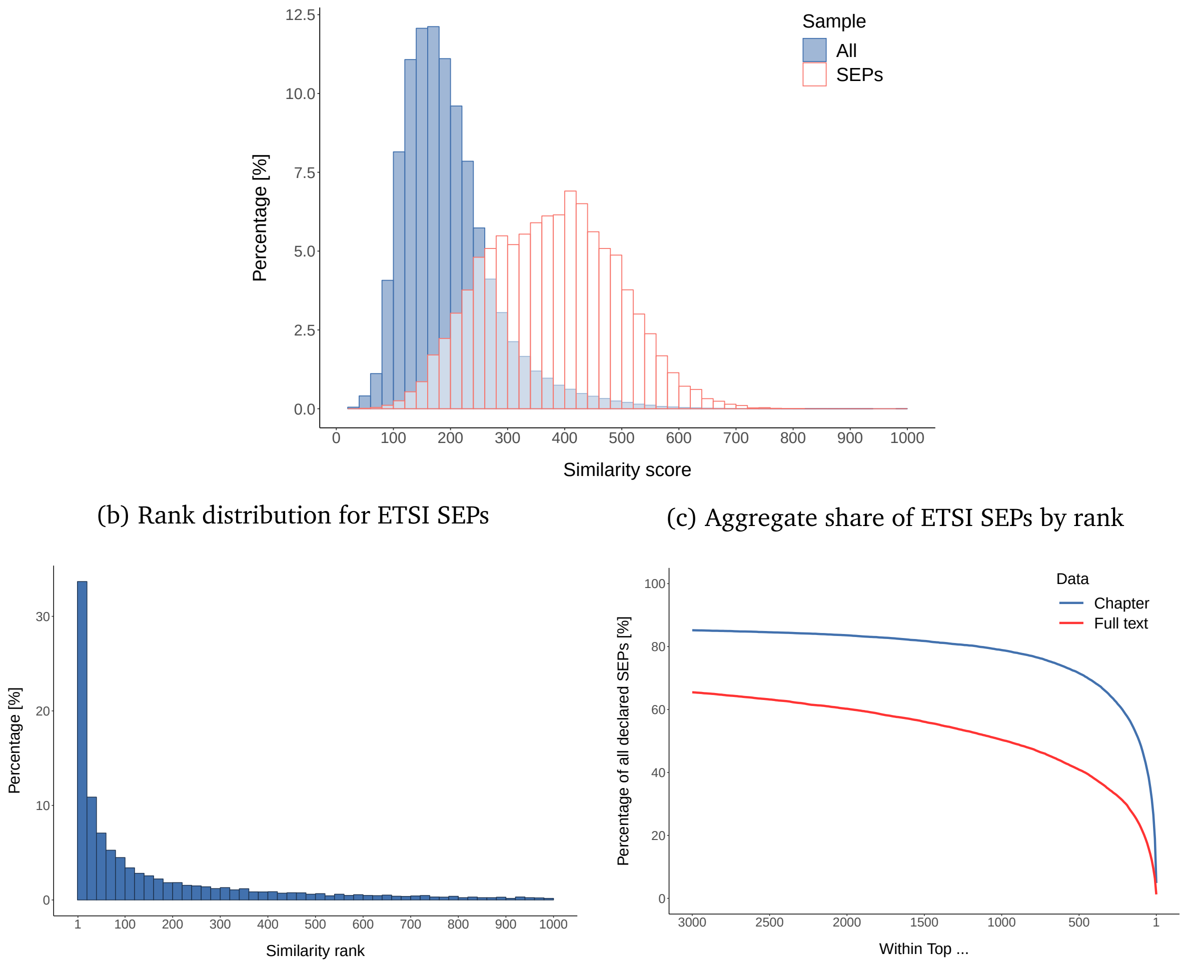
We first compare similarity scores between a given standard document and patents declared to be essential for the respective ETSI standard with scores of similarity between the standard and technologically similar, yet undeclared patents. To this end, we select patents with the same CPC-4 codes (e.g., one of the most common technology classes is the *H04W 72* class for local resource managements in wireless communications networks) and the 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.

As explained above, we observe the 3,000 most similar patent families for each chapter of each standard document cited in SEP declarations, meaning that we have to deal with either truncation or censoring. Using the most similar chapter for all standards to all patents, we observe 15,000 SEP-standard document pairs (*SEP declarations*) in our data. Considering the truncated dataset and additionally restricting the sample to 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 is not necessarily part of the dataset. Here, we conservatively assign the lowest similarity value for the given standard in the data to the control patent. This most likely results in a considerable overestimation of similarity scores for control patents.[34]

Figure 5.2 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 about 59 points.

[34] We obtain similar results when using censored data for both SEPs and controls. The results are reported in the Appendix.

Figure 5.1: Distribution of SEPs in similarity dataset



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

(b) Rank distribution for ETSI SEPs

(c) Aggregate share of ETSI SEPs by rank

Notes: This 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 ETSI SEPs at chapter level. The bottom right-hand graph shows the aggregate share of ETSI SEPs by *similarity rank* at chapter level (blue line) and document level (red line).

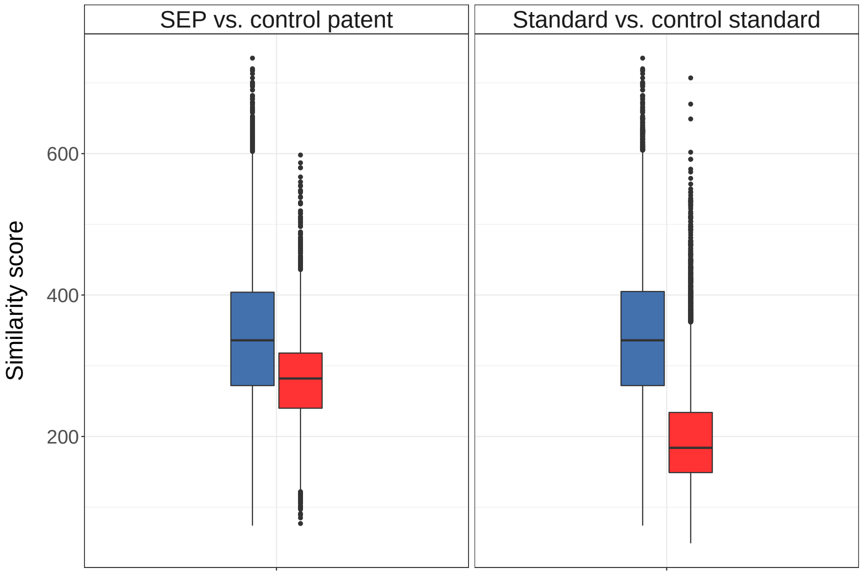
On the right-hand side, the standards referenced in the SEPs are compared with control standards. Here, the mean difference in similarity scores is about 135. All differences are statistically significant with t-values greater than 60.[35] To summarize, the results of our control group comparison strongly suggest that semantic approaches are appropriate to measure technological similarity between patents and standards.

Correlation of patent-standard similarity with patent characteristics

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.

[35] Table B-2 in the Appendix reports the corresponding t-statistics.

Figure 5.2: 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 similar 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 similar control standards (red).

Summary statistics are reported in Table 5.2. Secondly, we consider a subsample of declared SEPs in the dataset and examine correlations with various patent characteristics.

Table 5.2: Summary statistics (full sample)

	Mean	SD	Median	Min	Max	N
Similarity score	180.5000	72.7900	166	37	945	1762842
Similarity rank	1174.9000	883.7000	1021	1	3000	1765460
Granted US patent	0.4740	0.4990	0	0	1	1708537
# US fwd. cit. (5yrs)	11.9100	28.6800	4	0	3264	1320969
# Independent claims	3.1000	1.9060	3	1	19	914421
Length claim 1	151.2000	76.4500	138	0	399	892277
Patent family size	3.0580	3.1680	2	1	472	1708537
# Patent references	12.4500	22.5000	7	0	4148	1708537
# NPL references	3.5080	24.9600	0	0	12854	1708537
# Applicants	1.5540	1.3220	1	1	77	1687964
# Inventors	2.3780	1.6920	2	1	133	1700801
Priority Year	2003.7000	9.5470	2006	1950	2017	1707869

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 5.3, we correlate patent characteristics with the measure *Similarity score* in columns (1) and (2), and with 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 at document level. Looking at columns (1) and (3), we find significant and positive estimates for forward citations and patent family size, two established proxies for patent value in the literature. Furthermore, we find a negative relationship between patent grant and the similarity to a technical standard. We include claim characteristics 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 (i.e., less specific) claims are more similar to standards.

Table 5.3: Correlation of standards similarity with patent characteristics

	(1)	(2)	(3)	(4)
	Score	Score	Rank	Rank
# US fwd. cit. (5yrs)	0.0427^{***}	0.0391^{***}	-0.2915^{***}	-0.1736^{***}
	(0.002)	(0.002)	(0.029)	(0.033)
Granted US patent	-6.9068^{***}		123.4213^{***}	
	(0.112)		(1.679)	
Patent family size	0.5396^{***}	0.8246^{***}	-5.2433^{***}	-8.7013^{***}
	(0.016)	(0.020)	(0.240)	(0.294)
# Patent references	-0.0455^{***}	-0.0422^{***}	0.6337^{***}	0.5227^{***}
	(0.002)	(0.003)	(0.037)	(0.038)
# NPL references	0.0068^{**}	-0.0082^{***}	-0.1077^{**}	0.1208^{***}
	(0.002)	(0.002)	(0.035)	(0.036)
# Applicants	-0.1796^{***}	-0.3419^{***}	4.2723^{***}	6.3767^{***}
	(0.041)	(0.048)	(0.616)	(0.707)
# Inventors	0.0696^{*}	0.1054^{*}	0.2701	−0.0902
	(0.033)	(0.042)	(0.489)	(0.627)
# Independent claims		0.1194^{***}		0.0447
		(0.034)		(0.508)
Length claim 1		-0.0268^{***}		0.3332^{***}
		(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	1267993	717390	1270010	717519

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 5.4 reports the correlations of the similarity score with SEP characteristics, revealing some

Table 5.4: Correlation of standards similarity with SEP characteristics

	(1) Score	(2) Score	(3) Rank	(4) Rank	(5) Score	(6) Score
# US fwd. cit. (5yrs)	0.0281	−0.1598***	−0.0034	0.4380***	0.1104***	−0.1586***
	(0.020)	(0.031)	(0.102)	(0.157)	(0.028)	(0.044)
Granted US patent	0.5193	0.2809	31.3032***	31.8630***	2.4966	1.3933
	(2.224)	(2.218)	(11.304)	(11.299)	(4.832)	(4.784)
Patent family size	0.0545	0.1184	−0.2239	−0.3738	−0.4095**	−0.3081*
	(0.111)	(0.111)	(0.566)	(0.567)	(0.183)	(0.182)
# Patent references	−0.1927***	−0.1930***	1.1374***	1.1382***	−0.0660	−0.0815*
	(0.035)	(0.035)	(0.180)	(0.180)	(0.044)	(0.044)
# NPL references	0.0968***	0.0880***	−0.5064***	−0.4856***	0.0693*	0.0581
	(0.027)	(0.026)	(0.135)	(0.135)	(0.037)	(0.037)
# Applicants	−0.7366	−0.9278*	6.6527**	7.1019***	−0.5256	−0.7474
	(0.526)	(0.525)	(2.671)	(2.673)	(0.820)	(0.812)
# Inventors	−1.1188**	−1.2100**	5.1434*	5.3576**	−0.9763	−1.1521
	(0.524)	(0.523)	(2.662)	(2.661)	(0.858)	(0.849)
Section-specific declaration	1.5047	2.0092	−31.0910***	−32.2757***	11.0518***	11.4395***
	(1.918)	(1.913)	(9.745)	(9.745)	(3.391)	(3.357)
# SEP US fwd. cit. (5yrs)		1.5066***		−3.5383***		2.0076***
		(0.189)		(0.965)		(0.258)
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	0.46	0.47	0.17	0.17	0.57	0.58
Observations	12302	12302	12302	12302	3262	3262

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.

differences compared to the full sample of patents.[36] We include CPC-4 technology class and technical specification (TS) fixed effects. In column (1), we do not observe an effect of forward citations on similarity. Only after including SEP forward citations, we observe a statistically significant, negative effect of patent forward citations, whereas SEP forward citations are positively related to standards similarity. Also, patent grant and family size seem unrelated to semantic similarity for the subsample of SEPs. However, the relationship between the relative measure *similarity rank* and patent grant is highly significant suggesting that granted SEPs are relatively less similar to the standard. 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 reconstruct the sample used in

[36] Summary statistics for the SEP subsample can be found in Table B-7 in the Appendix.

Stitzing et al. (2017). This sample comprises 3,262 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 correlation between the forward citation measures Stitzing et al. (2017) used and their variable of true LTE standard essentiality. They also find that SEPs declared to specific technical specifications are more likely to be essential – a result that we find as well.

Benchmark against manual SEP assessments

To validate our measures of semantic similarity, we regress manual SEP assessments on semantic similarity measures using an array of different specifications.[37] Essentiality assessments are reported as a binary outcome with 1 being actually essential and zero 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.[38] The main variable of interest is the *similarity score*, which we report for pairs of patent families and the most similar standard in the sample. Additionally, we include several patent characteristics as controls. The number of forward citations is computed at US patent family level. *Length claim 1* refers to the 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 5.5, 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. The effect size for a one standard deviation increase in similarity score (roughly corresponding to 100 points in our data) is 7.8 pp with the specification in column (1) that includes no fixed effects. This effect is remarkably similar to the one of our full specification in column (4), which controls for patent priority year, declaration year, technology class, technical specification and firm fixed effects. This battery 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. Moreover, we 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 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-4 in the Appendix). Although the subsamples of patents evaluated by technical experts are considerably smaller, we again observe statistically significant correlations that compare well 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 score 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.

[37] Table B-3 in the Appendix provides summary statistics for the full sample of 2,541 evaluated patent families.

[38] 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%).

Table 5.5: 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 was deemed essential by the evaluators for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard in the full sample are selected for the regressions. 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 varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

To validate predictions of the semantic similarity measure, we consider the sample of LTE patents and employ a 10-fold cross validation for all of our predictions. Using weighted precision and recall metrics, we obtain precision and recall scores of 61% and 64% when only simple similarity scores are used. Once we control for patent characteristics, precision and recall scores increase to 63%

and 65%, respectively. The inclusion of additional patent characteristics contributes little to the prediction scores.[39] 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.[40] These test and training datasets are used in the subsequent SEP portfolio estimations.

5.3 Estimating SEP portfolio shares

We use the data from Section 5.2 to derive SEP portfolio shares, i.e., a firm's share of declared patents that are (presumably) truly standard-essential. Based on the logarithmic regression results, we compute the predicted probabilities of standard-essentiality for a given patent. We estimate the share of presumably true SEPs $\widehat{P}_F$ on firm-level with the following equation:

$$\widehat{P}_F = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{\hat{\beta}_0+\sum_{j=1}^{K}\hat{\beta}_j X_{ij}}}{1+e^{\hat{\beta}_0+\sum_{j=1}^{K}\hat{\beta}_j X_{ij}}}, \tag{5.3.1}$$

where n is the number of patents for a given firm F and X_{ij} represent the explanatory variables used in the logistic regression. To restrict the number of regressors K, we consider only those measures that have shown statistically significant correlations with true essentiality in the case of LTE standards: 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-6 in the Appendix.

To determine the error of our prediction at an aggregated level as a function of the number of patents in the portfolio, we draw random portfolios from the test dataset on LTE patents.[41] 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 in essentiality ratios for actual and predicted essentiality ratios for varying numbers of portfolio sizes. Figure 5.3 plots the mean 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 data. The following fitted function describes the error rate for LTE patents:[42]

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

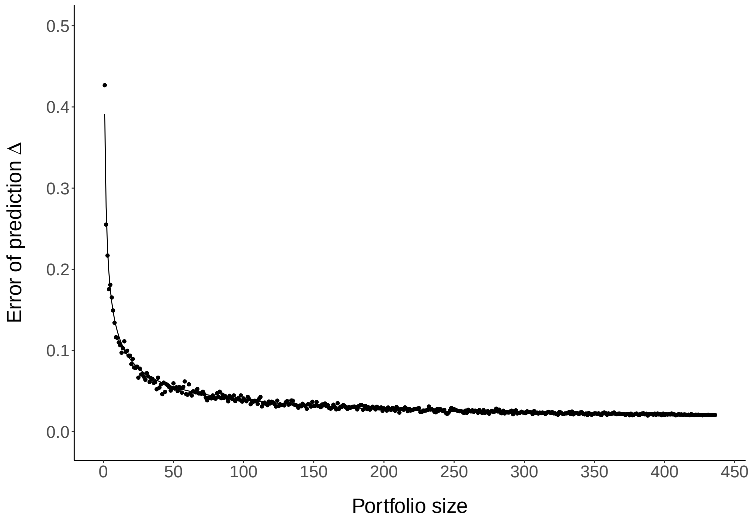
[39] We discuss regression results between various patent characteristics and the similarity score in the previous section.

[40] We report the confusion matrix for the test set of 402 SEPs for LTE standards in Table B-5 in the Appendix.

[41] 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.

[42] The error functions for UMTS and GSM standards are qualitatively very similar (see Figure A-5 in the Appendix).

Figure 5.3: The error of prediction as a function of portfolio size (LTE)



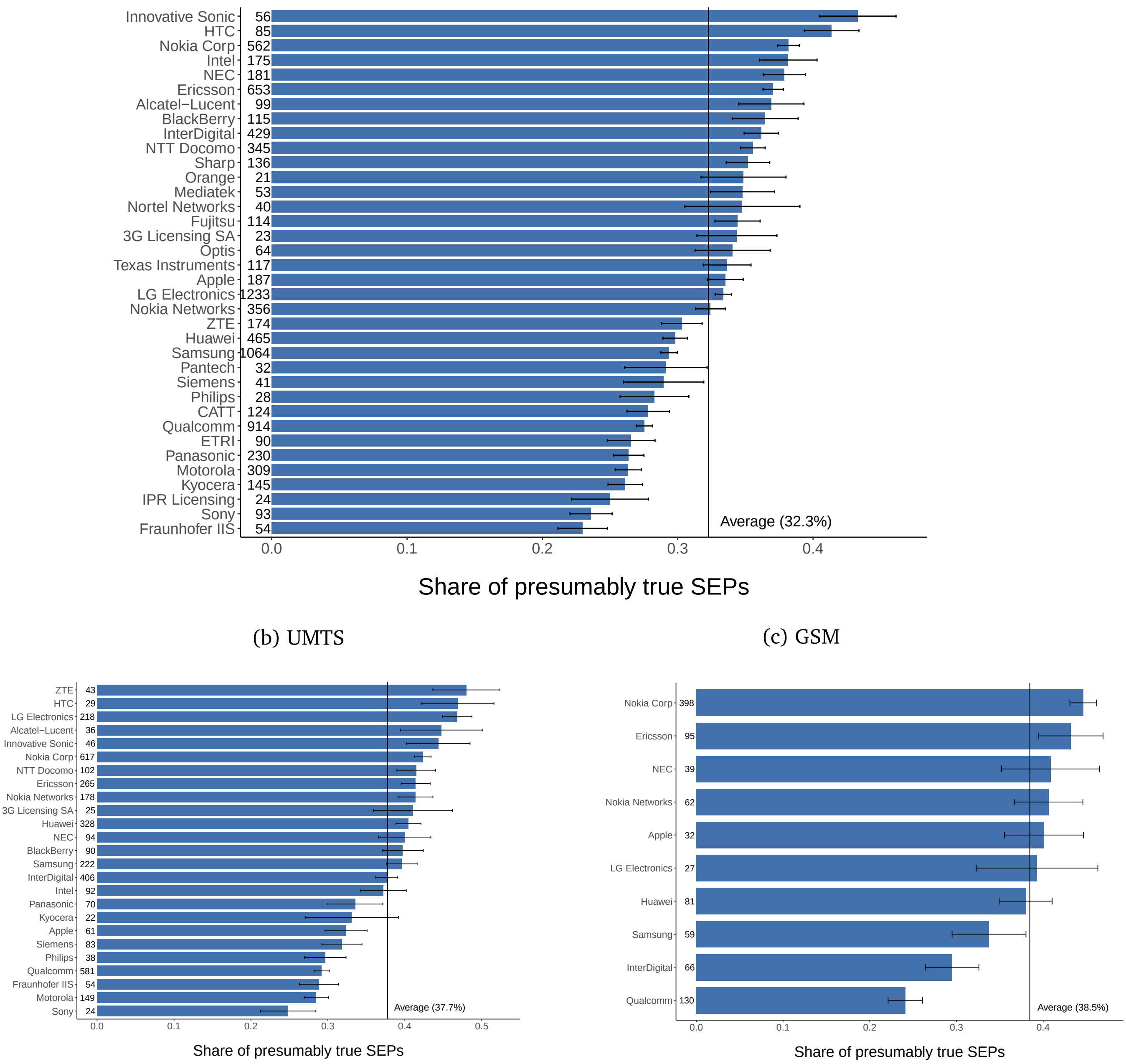
Notes: The error of prediction Δ is plotted as a function of portfolio sizes where portfolios are randomly drawn from the test sample. Additionally, a non-linear least squares fit is shown for the test sample of LTE patents. The fitted function is a power law function.

The left-hand side variable $\widehat{\Delta}$ is the difference in the share of presumably true SEPs for actual assessments and predictions and N the portfolio size, i.e., the number of patents for a given patent portfolio. We assume no additional constant in the power law function such that the function goes to 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 5.4, we present out-of-sample predictions for firm SEP portfolios for all three standard generations. In Figure 5.4a, 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 share of presumably true SEPs that is roughly twice as large as the one for the lowest-ranked firm. Notably, there seems no strong correlation between the share of true SEPs and portfolio size. In Figure 5.4b and Figure 5.4c, we present estimations at firm portfolio level 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).[43] We prefer to leave the question as to what causes this trend open for future work. However, the reason might be found in the changing composition of firms contributing technologies to standards. First, more and more firms hold a portfolio of at least 20 SEPs relevant to the younger generations of mobile telecommunication standards. Second, with non-practicing entities and implementers among them, the set of patent holders has become more diverse.

[43] Some late entrant firms, primarily known for being developers and implementers of recent standards such as UMTS and LTE, also made SEP declarations to later releases of the older GSM standard (GSM Phase 2+).

Figure 5.4: 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 patents. The lower left-hand graph shows the out-of-sample predictions on firm-level for UMTS patents. The lower right-hand graph shows predictions for GSM patents. The numbers on the left-hand side of the bars indicate the number of patent families declared to LTE/UMTS/GSM standards by the respective firm. Only results for firms with 20 or more declared patents reported. 95% confidence intervals are shown.

Chapter 6

IEEE Descriptives and Estimates

In this chapter, we first describe the sample of patents relating to standard specifications published by IEEE and provide selected descriptive statistics. Moreover, we present validation results and predictions for distinct standards.

6.1 Sample description

In Table 6.1, we report summary statistics for the two similarity measures (*similarity score* and *similarity rank*) based on the full text of all IEEE standard documents. Furthermore, we provide two comparisons between patents and standards. First, we compute the semantic similarity between patent and standards at document level. Second, we subdivide standard documents according to their table of contents into chapters and produce the similarity measures based on the comparison between patent and standard chapters. The measures reveal some distinct differences in similarity between different samples of patent-standard pairs. However, these are less pronounced as compared to the ones at ETSI. Notably, *similarity scores* are on average substantially lower compared to ETSI.[44] We provide statistics on all patents and declared SEPs, where patent-standard pairs are endogenously determined by the highest *similarity score*. Furthermore, we provide statistics on SEP declarations, where patent-standard pairs are predefined. We observe fairly small, yet statistically significant differences in the measured similarity as compared to the full sample. The average *similarity score* of declared SEPs to their most similar document is 263 whereas the average in the full sample of patent-document pairs is 196.[45] Figure 6.1a shows the similarity score distributions for all patents and the set of all SEPs.

In Figure 6.1b, the *similarity rank* distribution of all declared SEPs illustrates that this specific set of patents is among the highest ranked patent-standard pairs. About 28% of all SEPs declared at IEEE are among the top 100 for the corresponding standard text. Similarly, in Figure 6.1c, the percentage of SEPs declared at IEEE is plotted against the rank. We observe about 68% of declared SEPs within the top 3,000 patent families for the more fine-grained comparison between patents

[44] One reason may be the fact that many IEEE standards describe technologies that are not protected by intellectual property rights. Further in-depth analyses may shed light on this descriptive finding, but are beyond the scope of this report.

[45] We select the most similar document for each patent family in the data.

Table 6.1: Summary statistics: Similarity data (IEEE)

Sample	Variable	Mean	SD	SE	Min	Max	N
Document level							
All	Score	145	46	0.042	23	609	1204674
All	Rank	1485	898	0.819	1	3000	1204674
SEPs	Score	184	53	2.763	40	362	371
SEPs	Rank	955	937	48.639	1	2970	371
SEP declarations	Score	154	49	2.459	65	362	396
SEP declarations	Rank	1092	878	44.127	1	2973	396
Chapter level							
All	Score	196	67	0.033	27	893	4244135
All	Rank	1413	911	0.442	1	3000	4244135
SEPs	Score	263	82	3.206	79	582	650
SEPs	Rank	867	897	35.188	1	2970	650
SEP declarations	Score	231	78	2.782	93	582	786
SEP declarations	Rank	999	917	32.717	1	2996	786

Notes: Summary statistics for *similarity score* and *similarity rank* across three different datasets at patent family level. Minimum (maximum) possible score: 0 (1,000). Lowest (highest) possible rank: 3,000 (1).

and standards. This is a substantially larger share as compared to the full text comparison (about 40%). However, the difference remains less pronounced relative to the one found at ETSI. The curve decreases gradually and drops significantly for the highest ranked SEPs.

The specific disclosure of relevant IPR is not mandatory at IEEE. Consequently, many declarations are so-called blanket declarations, which do not inform about the number and identity of relevant patents. For illustration, we present firms that hold patents with a relatively high text similarity to IEEE standards. The left-hand side graph in Figure 6.2 lists the top patent applicants based on the number of similar patents. We restrict the sample to the 250 most similar patents for a given standard text.[46] In contrast, the right-hand side graph lists the top SEP declarants that specify the (supposedly) standard-essential patents. Similarly in Figure 6.3, we focus on IEEE 802.11 (WiFi) standard specifications only. The graphs reveal that some firms that file blanket declarations belong to the top patent applicants by portfolio size.

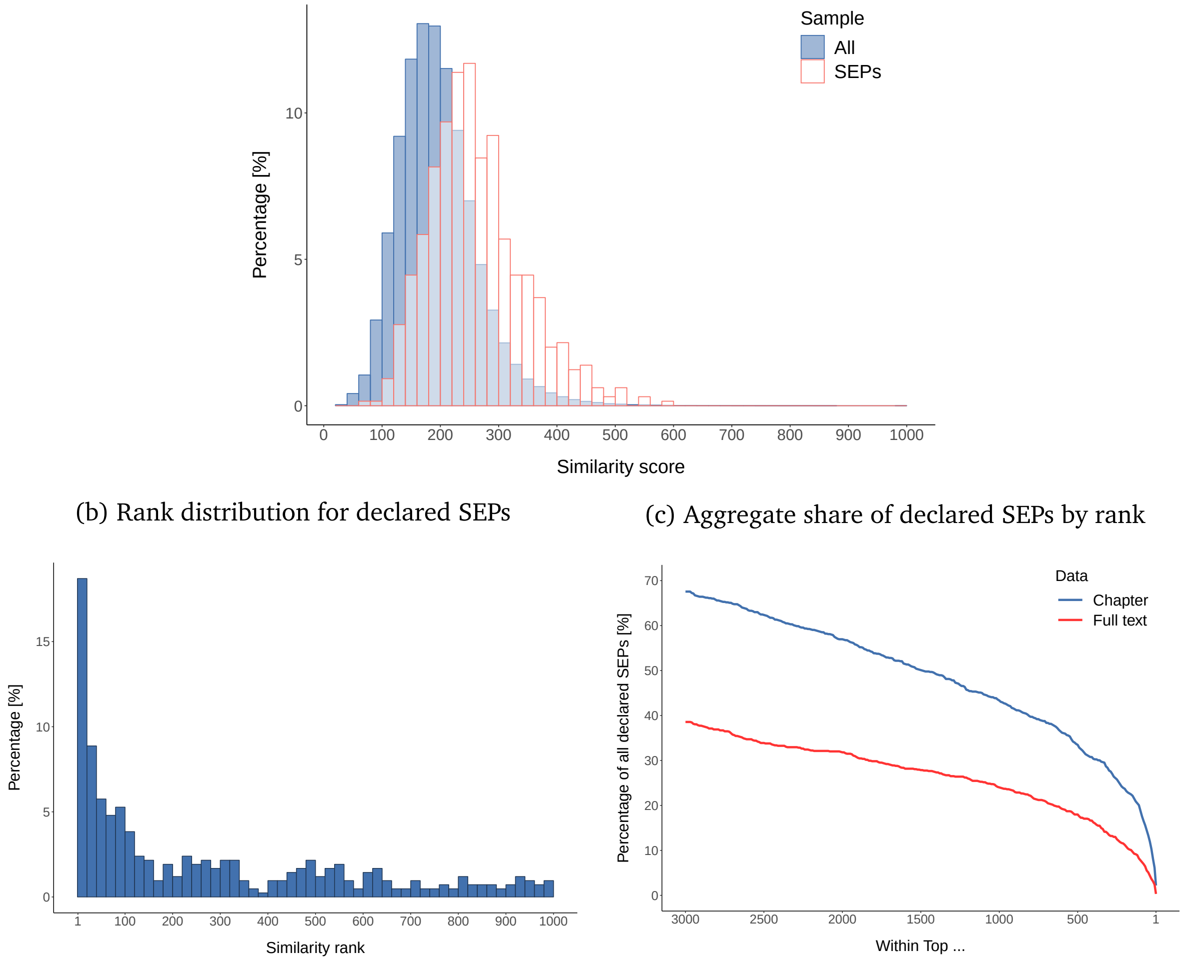
6.2 Validation results

We conduct three distinct validation exercises.[47] 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 estimate multi-

[46] In Figures A-6 and A-7 in the Appendix, we also report results for other thresholds. Namely, the top 100 and top 500 most similar patents are considered.

[47] For all subsequent analyses, we only report the results using the more fine-grained comparison between patent documents and chapters of standard documents.

Figure 6.1: Distribution of SEPs in similarity dataset (IEEE)



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

(b) Rank distribution for declared SEPs

(c) Aggregate share of declared SEPs by rank

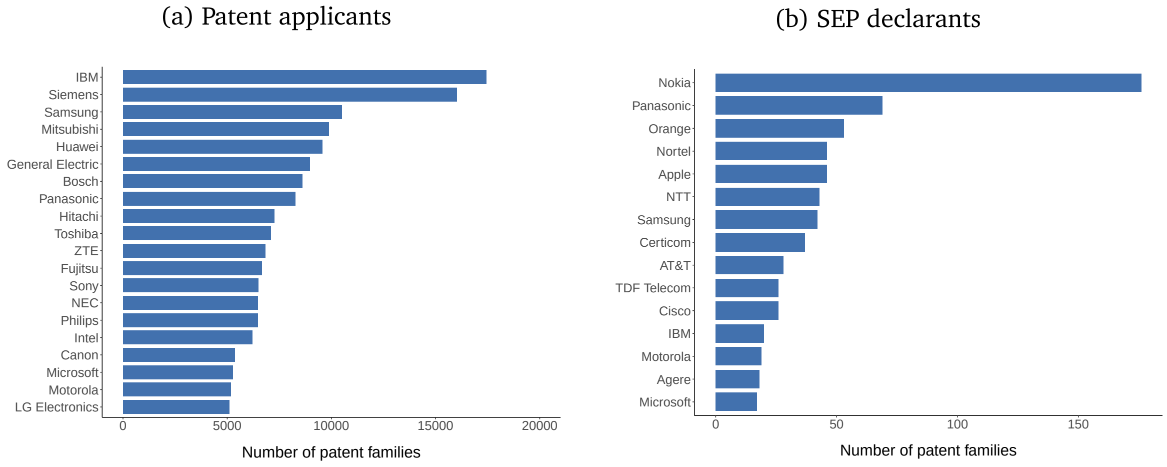
Notes: The top figure shows the similarity score distribution for two different sets of patents. All patents in the full sample (blue bars) are compared to the set of declared SEPs (white bars). The bottom left-hand graph shows the *similarity rank* distribution for declared SEPs at standard chapter level. The bottom right-hand graph compares the aggregate shares of declared SEPs by *similarity rank* at chapter and document level. The upper (blue) line refers to the more fine-grained comparison with chapters of standard documents, the lower (red) line represents the comparison on document level.

variate regressions of our semantic similarity measure on various patent characteristics. Third, we benchmark our results with a dataset of manually examined SEPs for various IEEE standard specifications. Based on these data, we test the predictive power of our semantics-based similarity measure to determine true standard essentiality.

Comparison of SEPs with control groups

The first step to validate our semantic approach for IEEE standards involves a comparison of declared SEPs with patents describing technologies from the very same technology class. If our measure has any explanatory value, SEPs will be significantly more similar to the referenced standards than the control patents. As discussed in Section 2, the majority of patents relevant for IEEE standards has

Figure 6.2: Patents by firm (IEEE, all standards)

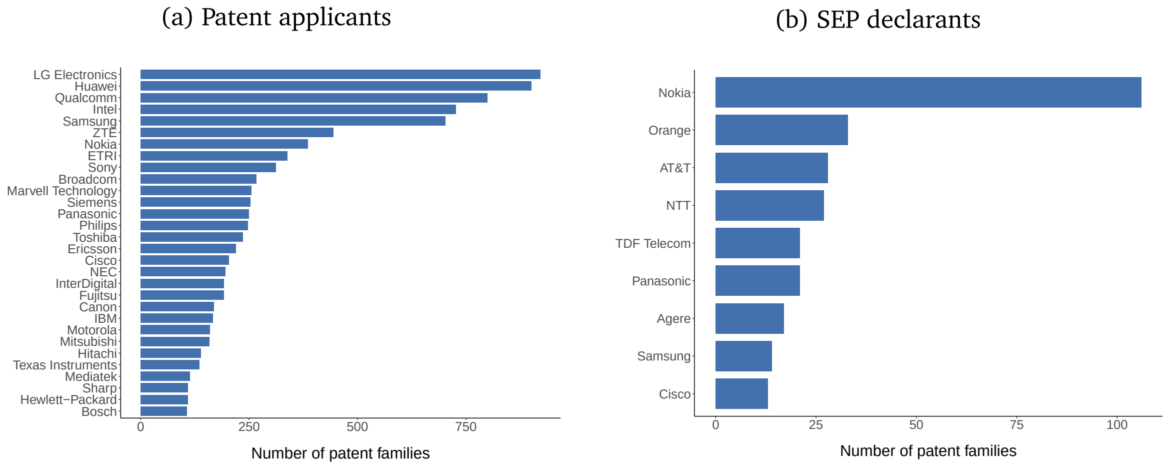


(a) Patent applicants

(b) SEP declarants

Notes: The graph on the left-hand side shows the number of patent families by patent applicant. Here, we exclude individual patent applicants and only consider patents that are among the 250 most similar patents for a given IEEE standard document. On the right-hand side, the number of declared SEPs (counted by the number of patent families) by declaring firm is shown.

Figure 6.3: Patents by firm (IEEE, WiFi standards)



(a) Patent applicants

(b) SEP declarants

Notes: The graph on the left-hand side shows the number of patent families by patent applicant. Here, we exclude individual patent applicants and only consider patents that are among the 250 most similar patents for a given standard document that relates to the 802.11 (WiFi) standards family. On the right-hand side, the number of SEPs which were declared to 802.11 specifications (counted by the number of patent families) by declaring firm is shown.

not been declared specifically. This implies that a text-based approach will not only identify declared SEPs as the most similar patents to standards, but also a considerable amount of relevant patents that are covered by blanket declarations. Notwithstanding this, we expect that the sample of declared SEPs is significantly more similar to the respective standards as compared to control patents. We note that the control group comparison with all SEPs renders the average difference in similarity a lower bound. We exploit the information that specific SEP declarations at IEEE refer to the respective standard. We call these predefined pairs of SEPs and standards simply *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 and same patent filing 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, to test the validity for the standard cited in the declaration, we keep the declared SEP fixed and compare the associated standard document to another randomly chosen standard document from the same IEEE standards family[48] as the focal standard.

As explained before, we only observe the 3,000 most similar patent families for each chapter of each standard cited in SEP declarations. This means that we have to account for either truncation or censoring in our analysis. Using the most similar document for all standards to all patents, we observe 786 SEP-standard document pairs (*SEP declarations*) in our data. Considering the truncated dataset, we obtain a total of 1,046 treated and control patents. Note that the control is not necessarily part of the dataset. Here, we conservatively assign the lowest observed similarity value for the given standard to the control patent. This most likely results in a considerable overestimation of similarity scores for control patents.[49]

Figure 6.4 illustrates 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 about 68 points. On the right-hand side, the standards referenced in the SEPs are compared with control standards. Here, the mean difference in similarity scores is about 76. All differences are statistically significant with t-values of 19.6 and 19.3, respectively. To summarize, the results of our control group comparison strongly suggests that semantic approaches are appropriate to measure technological similarity between patents and standards. The differences measured here are in a similar range as the ones in the control group comparison for ETSI SEPs and standards.

Correlation of patent-standard similarity with patent characteristics

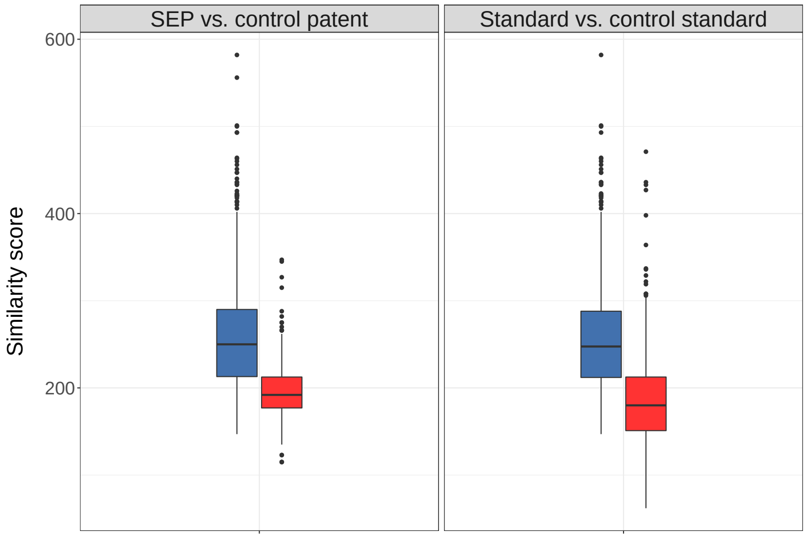
Mirroring the validation exercise for the ETSI sample in the previous chapter, we correlate our similarity measure with various proxies of patent value. We consider the full sample of patent families in our IEEE dataset. Summary statistics are reported in Table 6.2.

In Table 6.3, we correlate patent characteristics with the measure *similarity score* in columns (1) to (3), and with the relative measure *similarity rank* in columns (4) to (6). We include fixed effects for CPC-4 technology classes as well as for technical specifications at document level. Looking at columns (1) to (3), we find significant and positive effects for forward citations and patent family size. However, these findings are not consistent across all specifications with the relative similarity measure. In columns (2) and (5), we add a measure of SEP forward citations and find sizable and highly significant correlations with both measures of similarity. We further include claim characteristics in columns (3) and (6) and observe that fewer independent claims are associated with a higher standard similarity. Moreover, the length of the first claim is negatively correlated with simi-

[48] Examples for major standards families are IEEE 802.11 (WiFi), IEEE 802.3 (Ethernet), and IEEE 802.16 (WiMAX).

[49] When using censored data for both SEPs and controls, we find considerably smaller differences in similarity scores. This is may be caused by the presumably high share of unobserved SEP declarations. Nonetheless, the differences are statistically significant. The results using the censored sample are shown in Figure A-3 in the Appendix.

Figure 6.4: Comparison of SEP - standard pairs with control groups (IEEE)



Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and similar 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 similar control standards (red).

Table 6.2: IEEE: Summary statistics (full sample)

	Mean	SD	Median	Min	Max	N
Similarity score	196.3300	67.1330	189	27	893	4243422
Similarity rank	1413.3410	911.1640	1387	1	3000	4243422
# US fwd. cit. (5yrs)	4.1750	10.1200	1	0	1119	4243422
# SEP US fwd. cit. (5yrs)	0.0020	0.0610	0	0	23	4243422
# Independent claims	2.9560	1.9880	3	1	19	2054393
Length claim 1	128.8240	72.7840	114	0	399	2016577
Patent family size	3.1450	3.0360	2	1	427	4208290
# Patent references	13.6060	32.7370	8	0	6293	4208290
# NPL references	3.5260	30.3130	0	0	20508	4208290
# Applicants	1.4740	1.2240	1	1	77	4158351
# Inventors	2.3300	1.6690	2	1	133	4195591
Earliest filing year	2003.0450	10.5680	2005	1950	2019	4208290

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

larity suggesting that patents with broader claims are more similar to standards. These findings are consistent for both measures of similarity.

In Table 6.4, we report correlations of patent-standard similarity with SEP characteristics.[50] We

[50] Summary statistics for the SEP subsample can be found in Table B-8 in the Appendix.

Table 6.3: Correlation of standards similarity with patent characteristics (IEEE)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Score	Score	Score	Rank	Rank	Rank
Model	OLS	OLS	OLS	OLS	OLS	OLS
Patent family size	0.0390***	0.0359***	0.1471***	2.5289***	2.5596***	−1.0718***
	(0.0089)	(0.0089)	(0.0118)	(0.1598)	(0.1600)	(0.2025)
# Patent references	−0.0139***	−0.0136***	−0.0081***	0.3804***	0.3776***	0.1957***
	(0.0009)	(0.0009)	(0.0009)	(0.0193)	(0.0192)	(0.0166)
# NPL references	−0.0040**	−0.0042**	−0.0047***	−0.0512**	−0.0496**	0.0328*
	(0.0013)	(0.0013)	(0.0014)	(0.0172)	(0.0171)	(0.0165)
# Applicants	0.3050***	0.3069***	0.2129***	0.5637	0.5441	−0.5945
	(0.0230)	(0.0230)	(0.0252)	(0.4145)	(0.4144)	(0.4933)
# Inventors	−0.0471**	−0.0463**	−0.1231***	3.1661***	3.1582***	4.2210***
	(0.0169)	(0.0168)	(0.0214)	(0.2932)	(0.2932)	(0.4070)
# US fwd. cit. (5yrs)	0.1094***	0.1039***	0.0900***	0.4468***	0.5006***	0.1256*
	(0.0028)	(0.0028)	(0.0031)	(0.0495)	(0.0497)	(0.0565)
# SEP US fwd. cit. (5yrs)		12.1662***	13.2125***		−119.9265***	−125.1011***
		(0.9100)	(1.3385)		(9.5609)	(12.9913)
# Independent claims			−0.0738***			3.0171***
			(0.0175)			(0.3335)
Length claim 1			−0.0145***			0.2477***
			(0.0005)			(0.0091)
Earliest filing year	Yes	Yes	Yes	Yes	Yes	Yes
CPC-4 FE	Yes	Yes	Yes	Yes	Yes	Yes
Standard doc. FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.43	0.43	0.44	0.05	0.05	0.06
Observations	4,135,365	4,135,365	1,960,723	4,135,365	4,135,365	1,960,723

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.

include filing year and declaration year dummies as well as CPC-4 technology class fixed effects.

Similarly to the full sample of patents, we observe in columns (1) to (3) that forward citations are positively correlated with textual similarity to IEEE standards. Interestingly, this is not the case for SEP forward citations. The correlations with other patent characteristics are not stable across different specifications. This may be due to the relatively small sample size.

Table 6.4: Correlation of standards similarity with SEP characteristics (IEEE)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Score	Score	Score	Score	Score	Score
Model	OLS	OLS	OLS	OLS	OLS	OLS
Patent family size	0.4418	0.4856	0.4318	−0.1370	0.0018	0.4452
	(0.599)	(0.601)	(0.599)	(0.940)	(0.943)	(1.010)
# Patent references	0.1202	0.1220	0.0618	0.0903	0.0902	0.1173
	(0.131)	(0.131)	(0.141)	(0.163)	(0.162)	(0.176)
# NPL references	−0.0204	−0.0028	0.0085	−0.0088	0.0303	0.0175
	(0.057)	(0.061)	(0.060)	(0.075)	(0.080)	(0.083)
# Applicants	−0.2549	−0.3116	1.5509	1.3124	1.3303	1.5749
	(2.638)	(2.640)	(2.617)	(3.210)	(3.205)	(3.335)
# Inventors	0.4915	0.5563	−0.6333	−1.0742	−1.0101	−2.3693
	(1.987)	(1.989)	(2.071)	(2.816)	(2.812)	(3.021)
# US fwd. cit. (5yrs)	0.3620^{***}	0.3945^{***}	0.3200^{**}	0.2324	0.3194^{*}	0.3177
	(0.114)	(0.120)	(0.130)	(0.164)	(0.175)	(0.197)
# SEP US fwd. cit. (5yrs)		−1.4525	−0.7083		−3.2566	−3.0993
		(1.718)	(1.727)		(2.263)	(2.491)
# Independent claims				1.2606	1.1537	0.5346
				(1.813)	(1.812)	(2.091)
Length claim 1				0.0607	0.0583	−0.0315
				(0.063)	(0.063)	(0.072)
Earliest filing year	Yes	Yes	Yes	Yes	Yes	Yes
Earliest declaration year	Yes	Yes	Yes	Yes	Yes	Yes
CPC-4 FE	No	No	Yes	No	No	Yes
Adjusted R^2	0.19	0.19	0.33	0.19	0.19	0.32
Observations	649	649	648	379	379	378

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

Benchmark against manual SEP assessments

In Table 6.5, we present summary statistics for the sample of manually classified SEPs. The data are reported at patent-standard level. In total, we observe 272 pairs for 144 unique patent families. 25.7% of these can be considered as essential according to the assessment of our recruited patent attorneys. That is, the majority of patents is actually non-essential and several of these may not even be among the 3,000 most similar patents in our data. For unobserved patent-standard pairs, we therefore assign the lowest observed similarity score.[51]

Similarly to the analysis we conducted for the SEPs declared at ETSI, we test our method's validity for IEEE by regressing manual SEP assessments on semantic similarity measures using various specifications. Essentiality assessments are reported as binary outcome with one being actually essential and zero representing non-essential patent-standard pairs.

The main variable of interest is the *similarity score*, which we report for pairs of patent families and the most similar standard text in the sample. Additionally, several included patent characteristics are shown. The number of SEP forward citations is computed at US patent family level. *Length claim 1* refers to the number of words in the first independent claim.

In Table 6.6, we report logistic regression results for correlations between the similarity measure as independent variable and the manually assessed standard essentiality as dependent variable. Although the sample becomes very small when including the full set of fixed effects, we find positive and statistically significant correlations in all specifications. The effect size for a one standard deviation increase in similarity score (roughly corresponding to 75 points in our data) is 10.8 pp with the parsimonious specification in column (2) that includes no fixed effects. Even when we include fixed effects, the effect size does not become smaller than 9.3 pp.[52] Perhaps due to the small sample size, we do not observe statistically significant coefficients for the other covariates. However, the regressions show that SEP forward citations are generally positively related with standard essentiality. This is in line with previous findings for patents related to ETSI telecommunication standards.

[51] The lowest value for the similarity score is 80. Thus, we assign the value of 79 to all unobserved pairs. As a robustness check, we test various thresholds from zero to 100. Qualitatively, the results remain the same.

[52] Running the specification in column (2) at patent family level (Table B-9 in the Appendix) or only on WiFi related patent-standard pairs (Table B-10 in the Appendix) yields slightly larger coefficients. Considering the truncated sample of observed patent-standard pairs leads to a substantially smaller dataset. However, the coefficients remain statistically and economically significant with a 18.6 pp increase when the similarity score measure increases by 1 SD.

Table 6.5: Summary statistics (manual IEEE SEP assessments sample)

	Mean	SD	Median	Min	Max	N
Essential (y/n)	0.1950	0.3970	0	0	1	272
Similarity score	215.0480	74.3270	200	116	456	84
Similarity score (cens.)	121.0150	75.2150	79	79	456	272
Similarity rank	1080.8450	954.2820	736	4	2975	84
Similarity rank (cens.)	2408.0110	1033.8310	3001	4	3001	272
Patent family size	7.8010	11.1040	6	1	165	272
# Inventors	2.4630	1.6180	2	1	16	272
# Applicants	1.4710	1.4950	1	1	17	272
# Independent claims	4.2280	3.6420	3	1	42	272
Length claim 1	146.8930	109.3700	123	34	1362	272
# Patent references	23.4930	34.4660	16	0	227	272
# NPL references	21.4670	88.7020	3	0	723	272
# SEP US fwd. cit. (5yrs)	1.0850	2.4550	0	0	15	272
Earliest declaration year	2006.4410	4.9610	2007	1988	2017	272
Earliest filing year	1998.9190	6.0920	1999	1982	2014	272
IEEE 802.11 (WiFi)	0.4490	0.4980	0	0	1	272
IEEE 802.16 (WiMAX)	0.2720	0.4460	0	0	1	272

Notes: Summary statistics for the sample of declared patent families which were manually classified patents by technical experts. The sample is at patent-standard level.

Table 6.6: Logistic regressions: Standard essentiality (IEEE)

	(1)	(2)	(3)	(4)	(5)
Similarity score (cens.)		0.0014***	0.0019***	0.0012**	0.0032***
		(0.0003)	(0.0005)	(0.0005)	(0.0010)
# Independent claims	0.0110*	0.0119*	0.0157	0.0163	−0.0004
	(0.0066)	(0.0069)	(0.0106)	(0.0124)	(0.0195)
Length claim 1	−0.0001	−0.0001	−0.0001	−0.0002	0.0003
	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0006)
# Inventors	0.0084	0.0121	0.0087	0.0168	−0.0573
	(0.0173)	(0.0169)	(0.0250)	(0.0312)	(0.0614)
# Applicants	0.0287	0.0170	0.0224	0.0440	0.0977
	(0.0176)	(0.0172)	(0.0249)	(0.0323)	(0.0621)
Patent family size	0.0012	0.0027	0.0019	0.0048	−0.0445***
	(0.0020)	(0.0020)	(0.0027)	(0.0041)	(0.0171)
# Patent references	0.0008	−0.0004	−0.0019	−0.0024	−0.0137**
	(0.0012)	(0.0014)	(0.0020)	(0.0020)	(0.0062)
# NPL references	−0.0011	−0.0010	−0.0009	−0.0003	0.0011
	(0.0009)	(0.0009)	(0.0010)	(0.0010)	(0.0023)
# SEP US fwd. cit. (5yrs)	0.0128	0.0124	0.0191	0.0006	0.1716*
	(0.0132)	(0.0133)	(0.0185)	(0.0236)	(0.0898)
Earliest filing year	No	No	Yes	Yes	Yes
Earliest declaration year	No	No	No	Yes	Yes
CPC-3 FE	No	No	No	No	Yes
Pseudo R^2	0.05	0.14	0.20	0.28	0.37
AUC	0.67	0.74	0.79	0.84	0.87
Observations	272	272	215	181	132

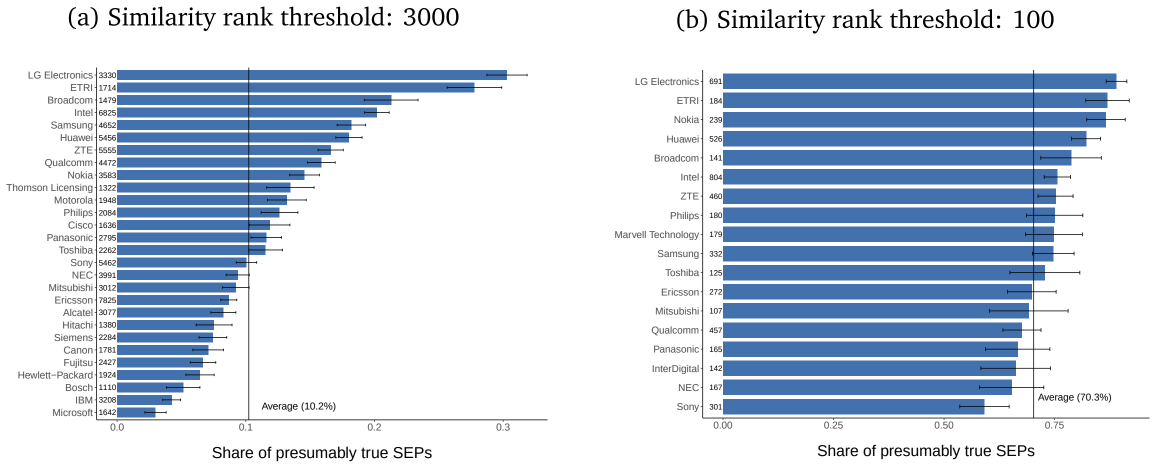
Notes: The dependent variable is a dummy equal to one if the patent family was deemed essential by the evaluators for IEEE standards. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard text for the standard specified in the SEP declaration are selected for the regressions. Marginal effects of one unit change are reported. The sample size varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

6.3 Estimating SEP portfolio shares

Analogously to the out-of-sample predictions for firms' LTE portfolios, we rely on the estimates from Section 6.2 to derive SEP portfolio shares for WiFi related technologies. Based on the logarithmic regression results, we compute the predicted probabilities of standard-essentiality for a given patent. As discussed above, SEPs often remain unidentified in the corresponding declaration. Due to these blanket filings, all patents from firms with at least one declaration letter qualify as standard-essential. We therefore compute predicted probabilities for all potentially relevant patents in our dataset.

In Figure 6.5, we present out-of-sample predictions for firm SEP portfolios. We observe substantial variation in standard essentiality between firms. We emphasize once more that some patents analyzed here are not declared specifically and therefore not claimed to be essential for IEEE 802.11 specifications. Consequently, the portfolio shares of presumably true SEPs can only be interpreted relative to other firms. The levels, however, depend on the sample presented in the graphs. Restricting the sample of patents in our dataset from the top 3,000 to the top 100 most similar patents increases the overall share of presumably true SEPs considerably. This is due to the fact that the similarity measure constitutes the main driver of the prediction. We present results for other thresholds in Figure A-8 in the Appendix.

Figure 6.5: SEP firm portfolio for IEEE 802.11 (WiFi) – out-of-sample predictions



(a) Similarity rank threshold: 3000

(b) Similarity rank threshold: 100

Notes: The graphs show the out-of-sample predictions at firm-level for patents that relate to the IEEE 802.11 (WiFi) standard. The numbers on the left-hand side of the bars indicate the patent family count. In the left-hand side graph, all patents within our dataset are considered whereas on the right-hand side, only patents which are among the 100 most similar patents for any WiFi related standard text. We presume patents with predicted probabilities greater than 0.5 as standard essential.

Chapter 7

ITU-T Descriptives and Estimates

In this chapter, we first describe the sample of patents relating to standard specifications published by ITU-T and provide selected descriptive statistics. Moreover, we present validation results and predictions for distinct standards.

7.1 Sample description

In Table 7.1, we report summary statistics for the two similarity measures (*similarity score* and *similarity rank*) based on the full text of all ITU-T standard documents. The measures reveal some distinct differences in similarity across different samples of patent-standard pairs. The differences are, however, less pronounced as compared to the case of ETSI. We provide statistics on all patents and declared SEPs, where patent-standard pairs are endogenously determined by the highest *similarity score*. Furthermore, we provide statistics on SEP declarations, where patent-standard pairs are predefined. Similar to the IEEE case, we observe small but statistically significant differences in the measured similarity as compared to the full sample. The average *similarity score* of declared SEPs to their most similar document is 287 whereas the average in the full sample of patent-document pairs is 199.[53] Figure 7.1a shows the similarity score distributions for all patents and the set of all declared SEPs.

In Figure 7.1b, the *similarity rank* distribution of all declared SEPs illustrates that this specific set of patents is among the highest ranked patent-standard pairs. Around 28% of all SEPs declared at ITU-T are among the top 100 for the corresponding standard text. Likewise, in Figure 7.1c, the percentage of SEPs declared at ITU-T is plotted against the rank reporting the samples of SEPs which are included in the similarity dataset. We observe 50% of declared SEPs within the top 3,000 patent families when considering the more fine-grained *chapter* level comparison. Based on the full text of a document, we only find 38% of all declared SEPs in our sample. This difference is less pronounced relative to the comparisons at ETSI and IEEE. In line with the results for the other two SSOs, the curve decreases gradually and drops significantly for the highest ranked SEPs.

As specific disclosure of relevant IPR is not mandatory at ITU-T and in fact many declarations are so-called blanket declarations, we present the applicants of all patents with a relatively high text

[53] We endogenously choose the most similar chapter text of any standard document for each patent in the data.

Table 7.1: Summary statistics: Similarity data (ITU-T)

Sample	Variable	Mean	SD	SE	Min	Max	N
Document level							
All	Score	187	58	0.055	37	746	1104496
All	Rank	1392	918	0.874	1	3000	1104496
SEPs	Score	248	74	2.537	84	655	857
SEPs	Rank	827	895	30.561	1	2992	857
SEP declarations	Score	208	79	3.342	72	655	558
SEP declarations	Rank	761	793	33.579	1	2961	558
Chapter level							
All	Score	199	75	0.039	22	898	3782982
All	Rank	1382	916	0.471	1	3000	3782982
SEPs	Score	287	98	2.918	32	669	1137
SEPs	Rank	859	911	27.013	1	2989	1137
SEP declarations	Score	276	84	2.972	83	669	803
SEP declarations	Rank	730	825	29.115	1	2994	803

Notes: Summary statistics for *similarity score* and *similarity rank* across three different datasets at document level. Minimum (maximum) possible score: 0 (1,000). Lowest (highest) possible rank: 3,000 (1).

similarity to ITU-T standards. The left-hand side graph in Figure 7.2 lists the top patent applicants according to the number of filed patents. We restrict the sample to the 250 most similar patents for a given standard document.[54] In contrast, the graph on the right-hand side lists the top firms that declared SEPs specifically. We focus on ITU-T's H.264 video compression standard in Figure 7.3. The graphs reveal that some firms with few – if any – specifically declared patents belong to the top patent applicants when considering all patents with high similarity to the H.264 standard.

7.2 Validation results

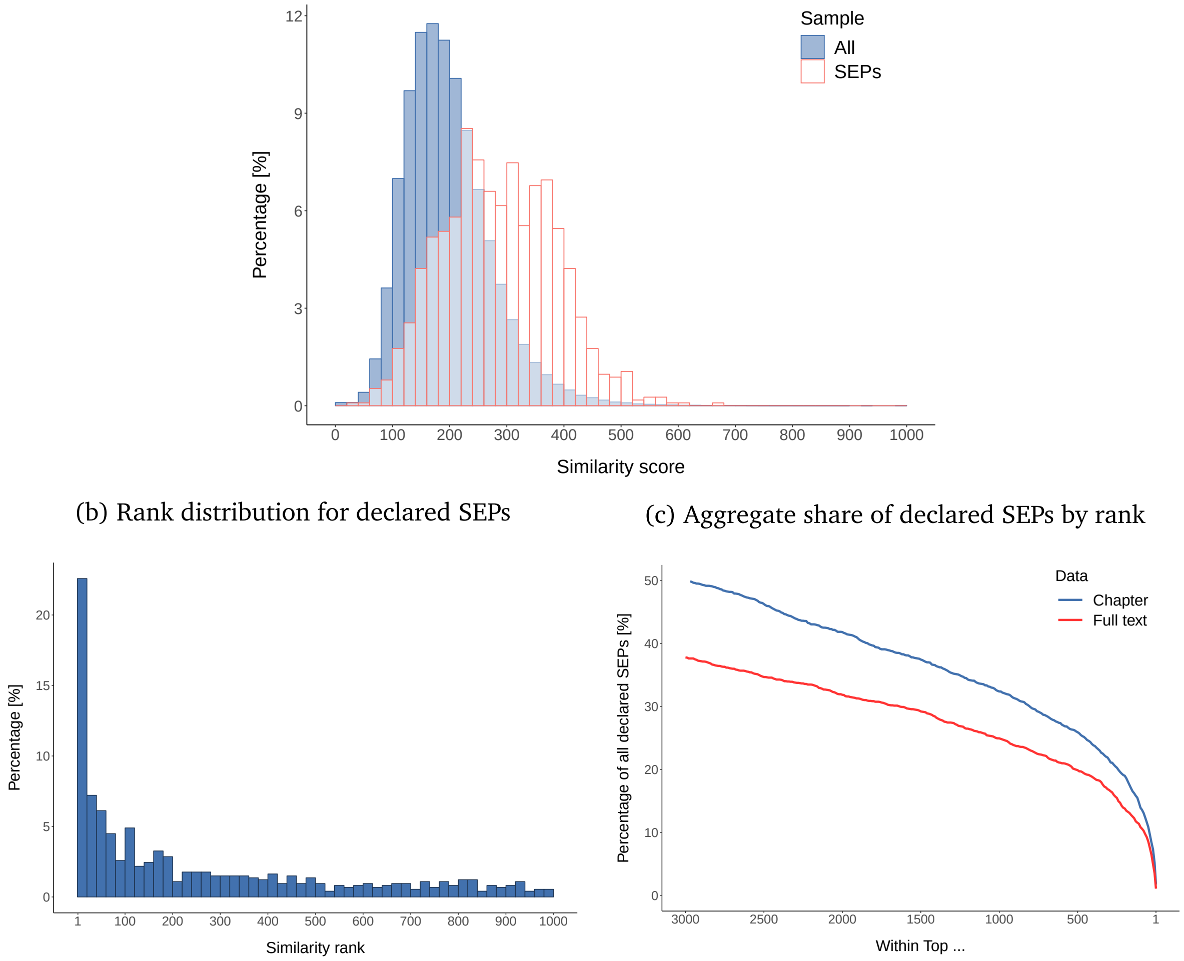
Comparison of SEPs with control groups

The first exercise to validate our semantic approach for ITU-T standards involves a comparison of declared SEPs with patents describing technologies from the very same technology class. If our measure has any explanatory value, SEPs will be significantly more similar to the referenced standards than the control patents. As discussed above, the majority of patents relevant for ITU-T standards has not been declared specifically. This means that a text-based approach will not only identify declared SEPs as the most similar patents to standards, but also patents that may have been declared in blanket declarations and are relevant for the referenced standard as well. In the same way as with IEEE, we still expect that the sample of declared SEPs is significantly more similar to the respective standards than control patents. Whenever the specific SEP declaration at ITU-T cites the

[54] In Figures A-9 and A-10 in the Appendix, we also report results for other thresholds. Namely, the top 100 and top 500 most similar patents are considered.

Figure 7.1: Distribution of SEPs in similarity dataset (ITU-T)

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



(b) Rank distribution for declared SEPs

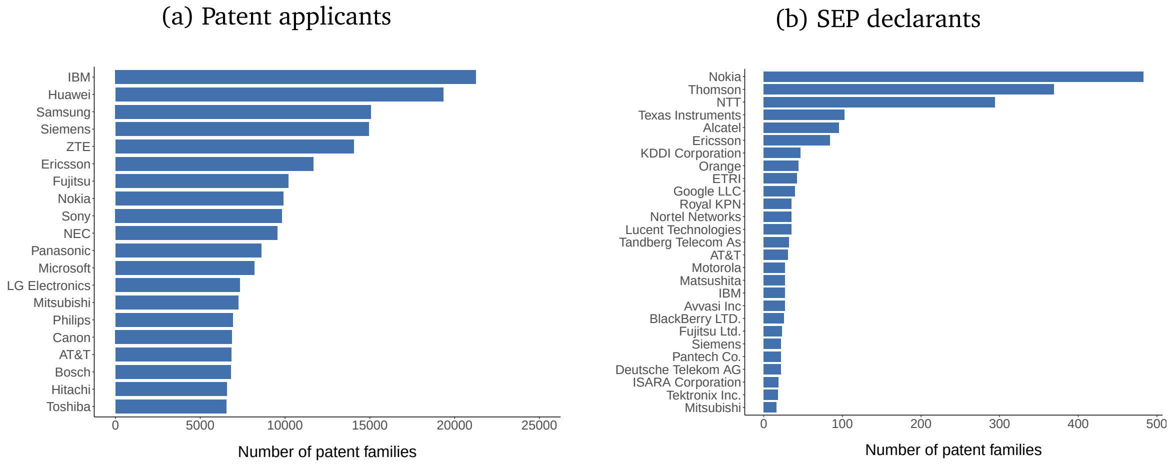
(c) Aggregate share of declared SEPs by rank

Notes: This 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 declared SEPs (white bars). The bottom left-hand graph shows the *similarity rank* distribution for declared SEPs at chapter level. The bottom right-hand graph compares the aggregate shares of declared SEPs by *similarity rank* at chapter and document level. The upper (blue) line refers to the more fine-grained comparison with chapters of standard documents, the lower (red) line represents the comparison at document level.

respective standard, we call these predefined pairs of SEPs and standards simply *SEP declarations*. We compare those to pairs of the same standard and undeclared patents from the same technology class and cohort.

Once again, we construct the control group out of patents with the same CPC-4 codes and patent filing 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, to test the validity for the standard cited in the declaration, we keep the declared SEP fixed and compare the associated standard document to another randomly chosen standard document from

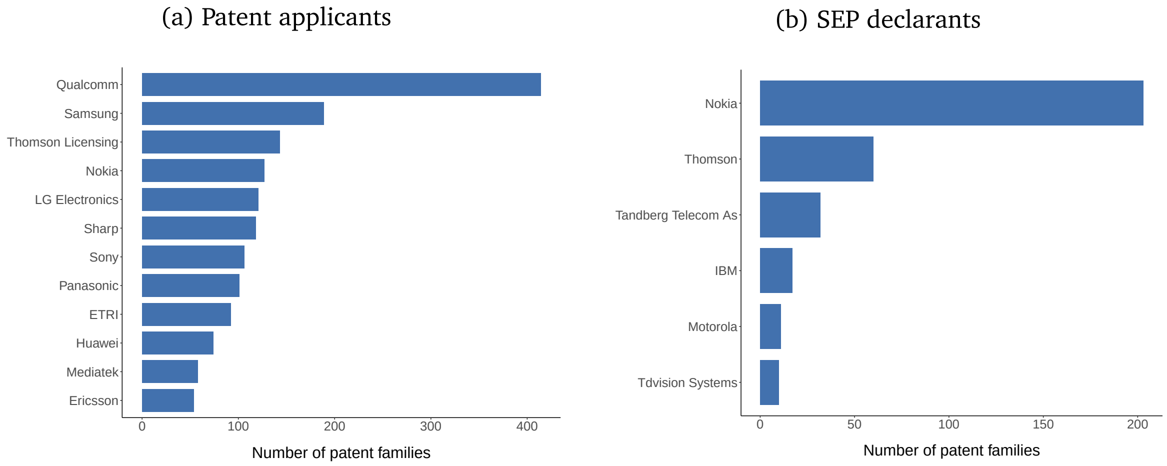
Figure 7.2: Patents by firm (ITU-T, all standards)


(a) Patent applicants

(b) SEP declarants

Notes: The graph on the left-hand side shows the number of patent families by patent applicant. Here, we exclude individual patent applicants and only consider patents that are among the 250 most similar patents for a given ITU-T standard document. On the right-hand side, the number of declared SEPs (counted by the number of patent families) by declaring firm is shown.

Figure 7.3: H.264 patents by firm (ITU-T, H.264 standards)


(a) Patent applicants

(b) SEP declarants

Notes: The graph on the left-hand side shows the number of patent families by patent applicant. Here, we exclude individual patent applicants and only consider patents that are among the 250 most similar patents for a given standard document that relates to the H.264 standards family. On the right-hand side, the number of SEPs which were declared to H.264 specifications (counted by the number of patent families) by declaring firm is shown.

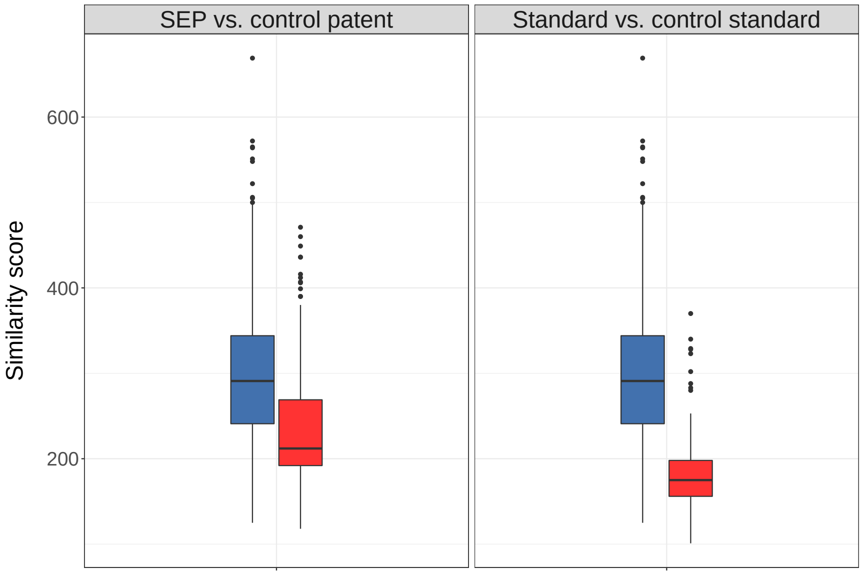
the same broad ITU-T recommendation series[55] as the focal standard. As explained before, we only observe the 3,000 most similar patent families for each standard document cited in SEP declarations. We therefore have to deal with either truncation or censoring. Using the most similar document for all standards to all patents, we observe 803 SEP-standard document pairs (*SEP declarations*) in our data. Considering the truncated dataset, we obtain a total of 1,272 treated and control patents.

[55] Examples for recommendation series are H ('*Audiovisual and multimedia systems*'), G ('*Transmission systems and media, digital systems and networks*') and V ('*Data communication over the telephone network*').

Note that the control is not necessarily part of the dataset. Here, we conservatively assign the lowest similarity value for the given standard in the data to the control patent. This most likely results in a considerable overestimation of similarity scores for control patents.[56]

Figure 7.4 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 about 73 points. On the right-hand side, the standards referenced in the SEPs are compared with control standards. Here, the mean difference in similarity scores is about 109. All differences are statistically significant with t-values of 20 and 32, respectively. To summarize, the results of our control group comparison strongly suggests that semantic approaches are appropriate to measure technological similarity between patents and standards. The differences in similarity are comparable to what we measured for ETSI standards and the corresponding declared SEPs.

Figure 7.4: Comparison of SEP - standard pairs with control groups (ITU-T)



Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and similar 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 similar control standards (red).

Correlation of patent-standard similarity with patent characteristics

We correlate our similarity measure with various bibliographic characteristics that capture patent value. We consider the full sample of patent families that appear in our dataset. Summary statistics are reported in Table 7.2.

In Table 7.3, we correlate patent characteristics with the measure *similarity score* in columns (1) to (3), and with the relative measure *similarity rank* in columns (4) to (6). We include fixed

[56] When using censored data for both SEPs and controls, we do not observe such large differences in similarity. This is most likely due to the high share of unobserved SEP declarations. The differences are, however, statistically significant. The results are shown in Figure A-4 in the Appendix.

Table 7.2: ITU-T: Summary statistics (full sample)

	Mean	SD	Median	Min	Max	N
Similarity score	198.7210	75.2340	188	20	898	3782312
Similarity rank	1382.5030	916.2260	1347	1	3000	3782312
# US fwd. cit. (5yrs)	4.2820	10.6360	1	0	1119	3782312
# SEP US fwd. cit. (5yrs)	0.0020	0.0650	0	0	14	3782312
# Independent claims	3.0460	1.9970	3	1	19	1837541
Length claim 1	129.7870	72.5420	115	0	399	1803843
Patent family size	3.0430	3.0000	2	1	427	3745472
# Patent references	13.0260	31.4160	8	0	4779	3745472
# NPL references	3.6120	31.0040	0	0	20508	3745472
# Applicants	1.4850	1.2390	1	1	100	3702209
# Inventors	2.3420	1.6830	2	1	99	3731100
Earliest filing year	2003.6920	10.3950	2006	1950	2019	3745472

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

effects for CPC-4 technology classes as well as for technical specifications at document level. In columns (1) to (3), we find significant and positive effects for patent forward citations. However, the correlation is not consistent with the relative similarity measure. Likewise, patent family size is not fully consistent across both measures of similarity. In columns (2) and (5), we add a measure of SEP forward citations and find very strong and statistically highly significant correlation for both measures. We include claim characteristics in columns (3) and (6) and find that fewer 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 7.4 reports correlations of patent-standard similarity with SEP characteristics.[57] We include filing year and declaration year dummies as well as CPC-4 technology class fixed effects. In line with the estimates for the full sample of patents, we observe in columns (1) to (3) that both forward citations and SEP forward citations are positively correlated with textual similarity to ITU-T standards. The correlation is much stronger for the latter. In this reduced sample, the size of the patent family is somewhat positively correlated with similarity to standards. We include information on patent claims in columns (4) to (6). We note that the sample size reduces considerably due to missing claim information. Nonetheless, we observe a positive relationship of the number of independent claims with textual similarity to standards.

[57] Summary statistics for the SEP subsample can be found in Table B-11 in the Appendix.

Table 7.3: Correlation of standards similarity with patent characteristics (ITU-T)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Score	Score	Score	Rank	Rank	Rank
Model	OLS	OLS	OLS	OLS	OLS	OLS
Patent family size	−0.0655***	−0.0739***	0.1914***	2.0483***	2.1186***	−1.9989***
	(0.0096)	(0.0095)	(0.0127)	(0.1678)	(0.1669)	(0.2177)
# Patent references	−0.0293***	−0.0289***	−0.0111***	0.5547***	0.5519***	0.2520***
	(0.0014)	(0.0014)	(0.0012)	(0.0268)	(0.0267)	(0.0221)
# NPL references	−0.0011	−0.0013	−0.0086***	−0.0551**	−0.0528**	0.0884***
	(0.0011)	(0.0011)	(0.0018)	(0.0183)	(0.0182)	(0.0249)
# Applicants	−0.0841**	−0.0843**	−0.1574***	2.8744***	2.8762***	3.0295***
	(0.0257)	(0.0257)	(0.0277)	(0.4323)	(0.4323)	(0.5177)
# Inventors	−0.2365***	−0.2363***	−0.1868***	3.9216***	3.9201***	4.0273***
	(0.0190)	(0.0190)	(0.0239)	(0.3051)	(0.3051)	(0.4267)
# US fwd. cit. (5yrs)	0.0540***	0.0492***	0.0738***	0.4124***	0.4526***	−0.0496
	(0.0028)	(0.0028)	(0.0032)	(0.0503)	(0.0504)	(0.0574)
# SEP US fwd. cit. (5yrs)		13.1382***	11.2047***		−111.0773***	−88.1917***
		(0.6311)	(0.7422)		(7.0034)	(8.4139)
# Independent claims			−0.1483***			4.9088***
			(0.0195)			(0.3477)
Length claim 1			−0.0212***			0.2553***
			(0.0005)			(0.0096)
Earliest filing year	Yes	Yes	Yes	Yes	Yes	Yes
CPC-4 FE	Yes	Yes	Yes	Yes	Yes	Yes
Standard doc. FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.49	0.49	0.53	0.08	0.08	0.09
Observations	3,689,487	3,689,487	1,755,739	3,689,487	3,689,487	1,755,739

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 7.4: Correlation of standards similarity with SEP characteristics (ITU-T)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Score	Score	Score	Score	Score	Score
Model	OLS	OLS	OLS	OLS	OLS	OLS
Patent family size	1.1320***	0.8800**	0.2349	0.6879	0.3926	0.1753
	(0.405)	(0.410)	(0.385)	(0.459)	(0.465)	(0.441)
# Patent references	−0.3427*	−0.3788**	−0.3362*	−0.5250**	−0.5385***	−0.3050
	(0.191)	(0.191)	(0.184)	(0.208)	(0.207)	(0.210)
# NPL references	0.1385	0.0413	−0.0039	0.1369	0.0293	−0.0127
	(0.115)	(0.118)	(0.137)	(0.123)	(0.126)	(0.152)
# Applicants	1.9676	1.8636	1.7843	3.5145	3.3154	2.7903
	(2.136)	(2.126)	(2.112)	(2.373)	(2.357)	(2.368)
# Inventors	−0.3275	−0.1202	−2.8127	−4.0855*	−3.6193	−5.8527**
	(1.811)	(1.804)	(1.938)	(2.310)	(2.298)	(2.624)
# US fwd. cit. (5yrs)	0.8202***	0.6319***	0.5067**	0.7259***	0.5222**	0.4385
	(0.190)	(0.197)	(0.226)	(0.234)	(0.241)	(0.312)
# SEP US fwd. cit. (5yrs)		9.6362***	2.9751		10.3144***	3.0086
		(2.879)	(2.681)		(3.267)	(3.138)
# Independent claims				5.9625***	5.7511***	0.5996
				(1.432)	(1.423)	(1.390)
Length claim 1				−0.0709	−0.0695	−0.0639
				(0.054)	(0.054)	(0.053)
Earliest filing year	Yes	Yes	Yes	Yes	Yes	Yes
Earliest declaration year	Yes	Yes	Yes	Yes	Yes	Yes
CPC-4 FE	No	No	Yes	No	No	Yes
Adjusted R^2	0.13	0.13	0.39	0.17	0.18	0.45
Observations	1,128	1,128	1,128	668	668	668

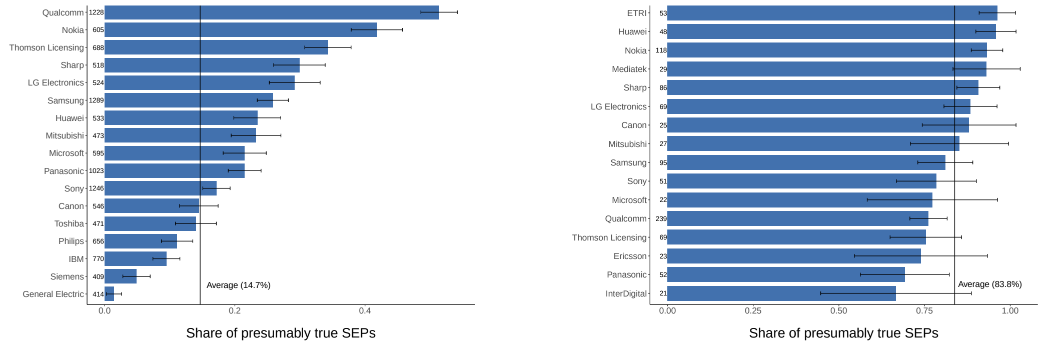
Notes: OLS regressions of similarity measures on patent family characteristics. The dependent variable *similarity score* is abbreviated as *score*. The sample consists of SEPs declared at ITU-T. Standard errors are in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

7.3 Estimating SEP portfolio shares

We estimate the share of SEPs for ITU's H.264 (MPEG-4) video compression standard. Since we have no manual assessments of SEPs declared to this standard, we are not able to compute ITU-specific estimates as input for our prediction equation. Instead, we rely on the coefficients obtained from logistic regressions using the random sample of IEEE's WiFi patents, which we already used in the prior chapter. This obviously remains a suboptimal choice, because IEEE and ITU-T standards rely on different technologies. Nonetheless, both SSOs set standards in the field of ICT and the declared SEPs indeed have several technology classes in common.[58] Furthermore, the correlations of our text similarity measure with patent characteristics for the ITU-T sample mirror those for the IEEE sample (cf. Table 6.3 with Table 7.3). Thus, we expect a similar correlational relationship between our text similarity measure and (presumed) standard essentiality. Moreover, IEEE and ITU-T share to a large extent the same SEP declarants and both allow for blanket declarations. Nonetheless, if the actual elasticity for ITU-T standards is larger (smaller) than the elasticity for IEEE standards, the estimated differences among the firms' shares of presumably true SEPs will be deflated (inflated). We therefore interpret the following out-of-sample predictions with caution.[59] We consider the full set of patents in our data as probably many SEPs are captured by blanket filings.

In Figure 7.5, we present out-of-sample predictions for firm SEP portfolios for the H.264 video compression standard. We observe substantial variation in standard essentiality across firms in the full dataset. The patents considered are not necessarily specifically declared to the MPEG standard. Hence, the shares of presumably true SEPs can only be interpreted relative to other firms. When considering the full dataset, only 14.7% of all patents are predicted to be standard-essential. Restricting the sample to patents with particularly high semantic similarity to MPEG specifications, we observe a substantially higher share of presumably true SEPs (83.8%).

Figure 7.5: SEP firm portfolio for ITU-T H.264 (MPEG) – out-of-sample predictions



(a) Similarity rank threshold: 3000

(b) Similarity rank threshold: 100

Notes: The graphs show the out-of-sample predictions on firm-level for patents related to ITU-T's H.264 (MPEG) standard. The numbers on the left-hand side of the bars indicate the patent family count. The left-hand (right-hand) side graph considers all patents (only the 100 most similar patents for any MPEG related standard text). We presume patents with predicted probabilities greater than 0.5 as standard essential.

[58] The most frequent CPC patent classes for both WiFi and MPEG are H04W, H04L, G06F, H04N, G06Q and H03M.

[59] Using a sample of declared LTE and UMTS SEPs at ETSI yields similar results for the full sample of H.264 related patents. The ranking of firms is practically identical to the one reported.

Chapter 8

Discussion and Conclusion

In this report, we introduce 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 across different technology standards administered by different standard-setting organizations.

We present the results of three distinct exercises to confirm the 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 throughout a significantly higher semantic similarity for standard-patent pairs defined by SEP declarations. We conclude that the semantic approach is suitable for measuring technological similarity between patents and standards. Second, we correlate our measure with different patent characteristics. In line with the general notion that truly standard-essential patents are of considerably high value, we find a strong and significant correlation between our measure of semantic similarity and established patent value indicators. Ultimately, we exploit information on manual essentiality assessments for a sample of SEPs declared essential to ETSI and IEEE standards. We find strong and highly significant correlations between the experts' decisions on standard essentiality and our measure of semantic similarity.

Naturally, a purely text-based determination of standard essentiality comes with some limitations. Inventors and patent attorneys may write the patent either using their own words or borrowing the terminology from standard documents. The calculated similarity scores will likely differ even if the underlying technology is the same. This introduces potential endogeneity 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. These dynamic aspects may be addressed in future versions of such semantics-based methods. Furthermore, a patent's claims solely define its scope of protection and hence, essentiality. Still, claims are typically written in a highly abstract and generic language that 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. We 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 incorporating both patent claims and description instead of the mere patent claims as input text.

For all three studied SSOs (ETSI, IEEE, ITU-T), we present descriptives on standard-patent pairs, which are either specifically declared or determined by our similarity measure. We further demonstrate our measure's usefulness in a first use case. We estimate shares of true SEPs in firm patent portfolios. In doing so, we benefit from the high accuracy of our approach when predicting standard essentiality at aggregate level. Using the results from the benchmark with manual SEP assessment data, we present out-of-sample predictions for firms' true shares of SEPs. In general, we find statistically and economically substantial differences. For instance, for LTE standards, 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. Finally, we also illustrate that our measure can be used to shed light on the number and identity of SEPs in those cases, where firms filed only blanket (i.e., unspecific) declarations.

Beyond this use case, we see several applications of our method in the academic as well as practical sphere. 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 at this point, 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 a substantial advantage of our approach lies in its scalability, and thus, time- as well as cost-efficiency. Moreover, the data generated through our method is arguably more objective and accessible than most of the proprietary datasets on SEP assessments. Against this backdrop, we hope that this report and the publicly available data invite even more scholars to empirically study the complex relationship between patents and standards.

Bibliography

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., 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. (2018). 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. forthcoming.

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.

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. 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 (2018). How Essential are Standard-Essential Patents? Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3128420.

Lerner, J., H. Tabakovic, and J. Tirole (2016). Patent Disclosures and Standard-Setting. NBER Working Paper No. 22768.

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.

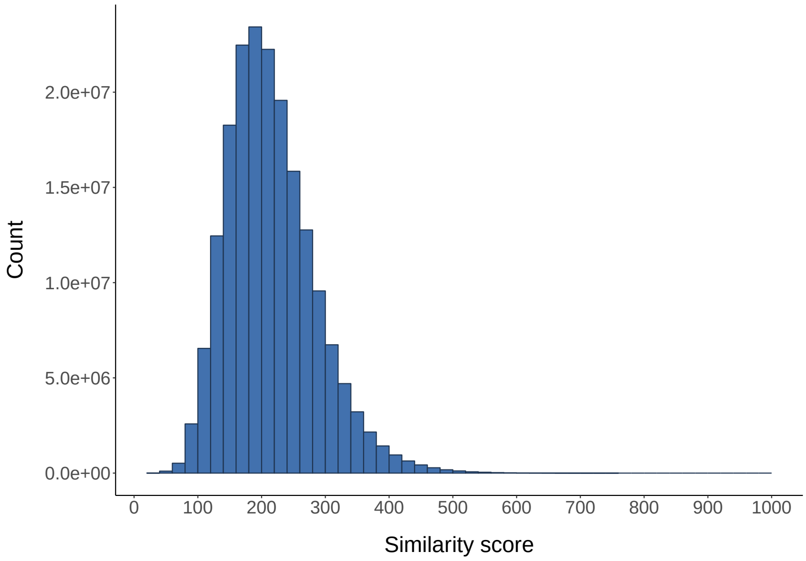
Stitzing, R., P. Sääskilahti, J. Royer, and M. V. Audenrode (2017). Over-Declaration of Standard Essential Patents and Determinants of Essentiality. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2951617.

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.

Appendix

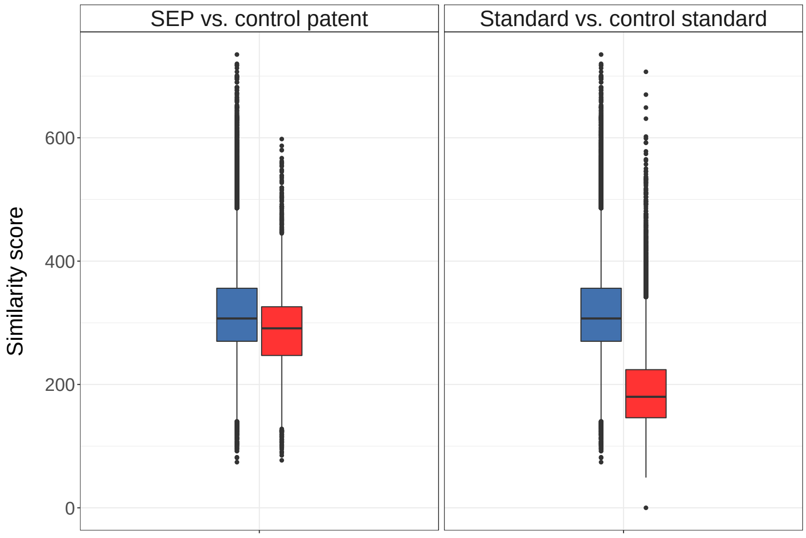
Appendix A: Figures

Figure A-1: Similarity score distribution over ETSI standard documents



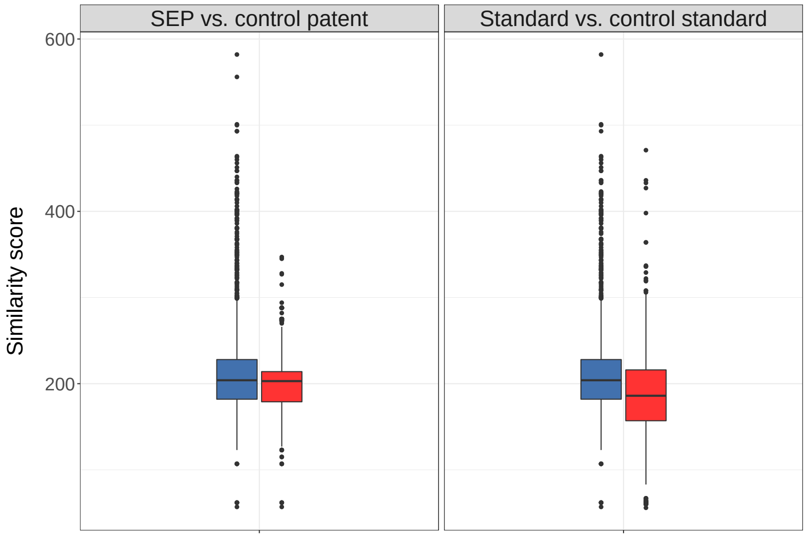
Notes: This figure shows the similarity score distribution for the full sample of ETSI patent-standard pairs. Sample size $N = 187,397,890$.

Figure A-2: Comparison of ETSI SEP - standard pairs with control groups (censored data)



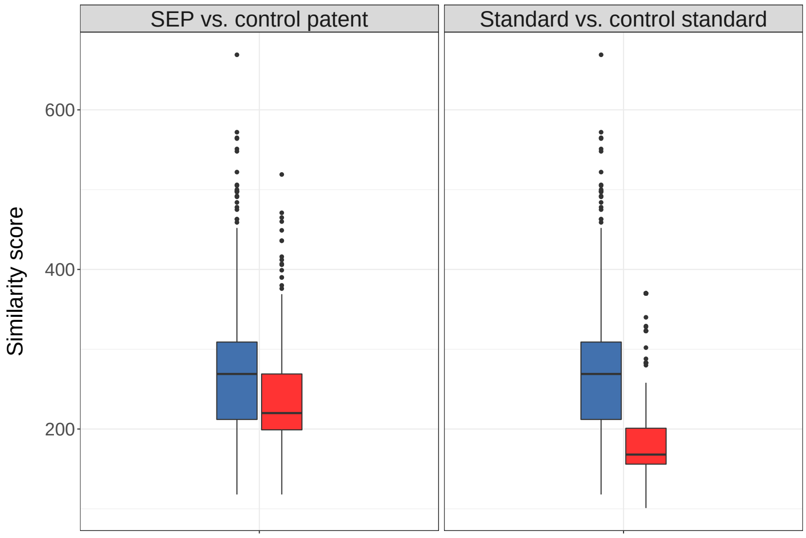
Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and similar control patents compared to the same standard (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and similar control standards (red). The censored dataset is used in this representation. Differences are significant, but are considerably less pronounced relative to the results with truncated data. Statistics are shown in Table B-2.

Figure A-3: Comparison of IEEE 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 similar 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 similar control standards (red). The censored dataset is used in this representation. Differences are significant, but are considerably less pronounced relative to the results with truncated data.

Figure A-4: Comparison of ITU-T SEP - standard pairs with control groups (censored data)

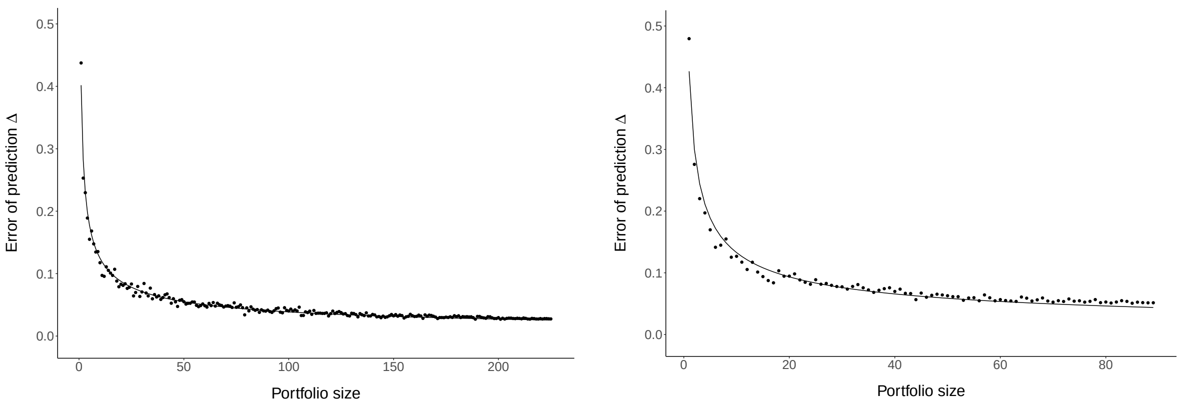


Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and similar control patents compared to the same standard (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and similar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

Figure A-5: The error of prediction as a function of portfolio size

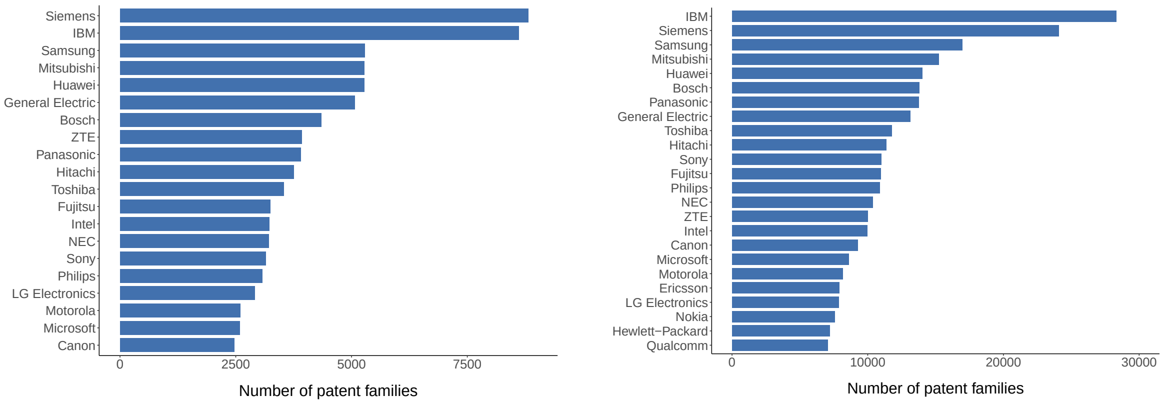
(a) UMTS

(b) GSM



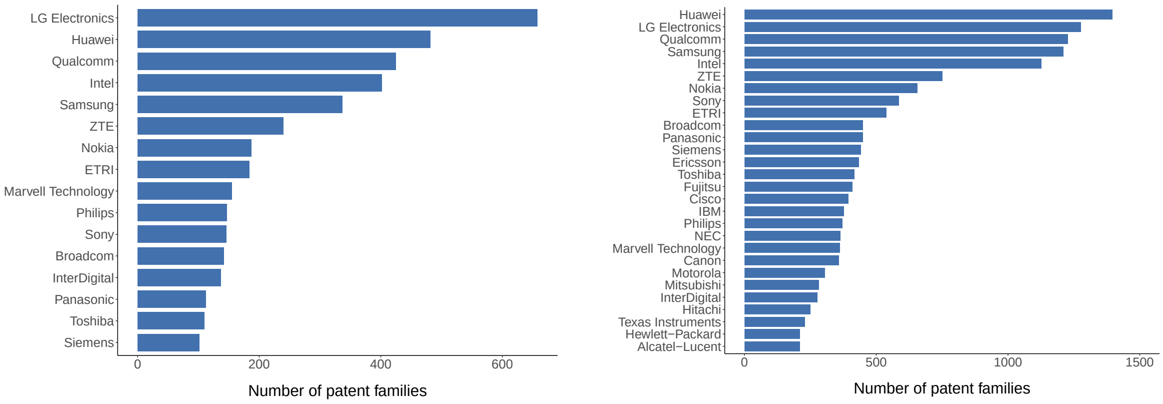
Notes: The error of prediction Δ is plotted as a function of portfolio size where portfolios are randomly drawn from the test sample of UMTS and GSM patents. Non-linear least squares fits are shown. The fitted functions are power law functions.

Figure A-6: Patents by firm (IEEE standards)



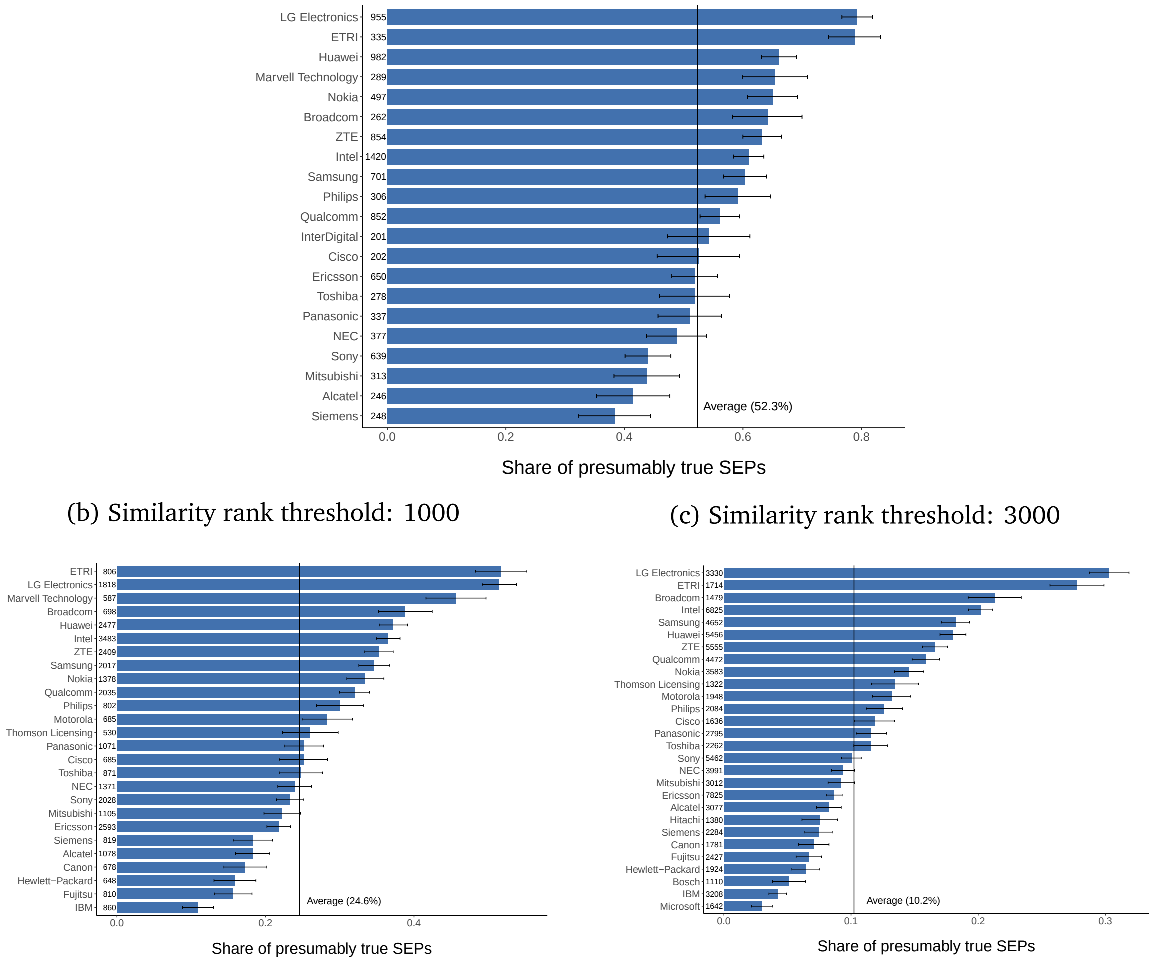
Notes: The graphs show the number of patent families by patent applicant. We exclude individual patent applicants. On the left-hand side we consider patents that are among the 100 most similar patents for a given IEEE standard document. On the right-hand side, the 500 most similar patents are shown.

Figure A-7: Patents by firm (IEEE, WiFi standards)



Notes: The graphs show the number of patent families by patent applicant. We exclude individual patent applicants. On the left-hand side we only consider patents that are among the 100 most similar patents for a given standard document that relates to the 802.11 (WiFi) standards family. On the right-hand side, the 500 most similar patents are shown.

Figure A-8: SEP firm portfolios for IEEE 802.11 (WiFi) – out-of-sample predictions

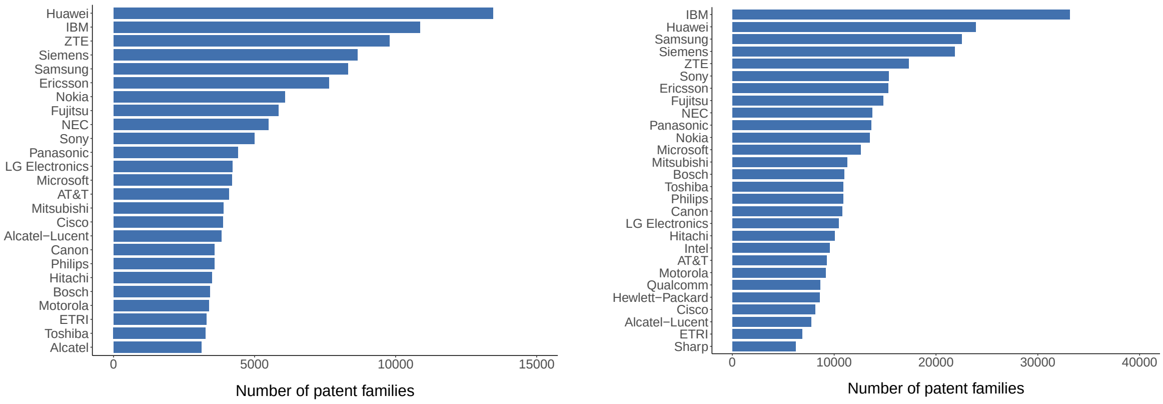

(a) Similarity rank threshold: 250

(b) Similarity rank threshold: 1000

(c) Similarity rank threshold: 3000

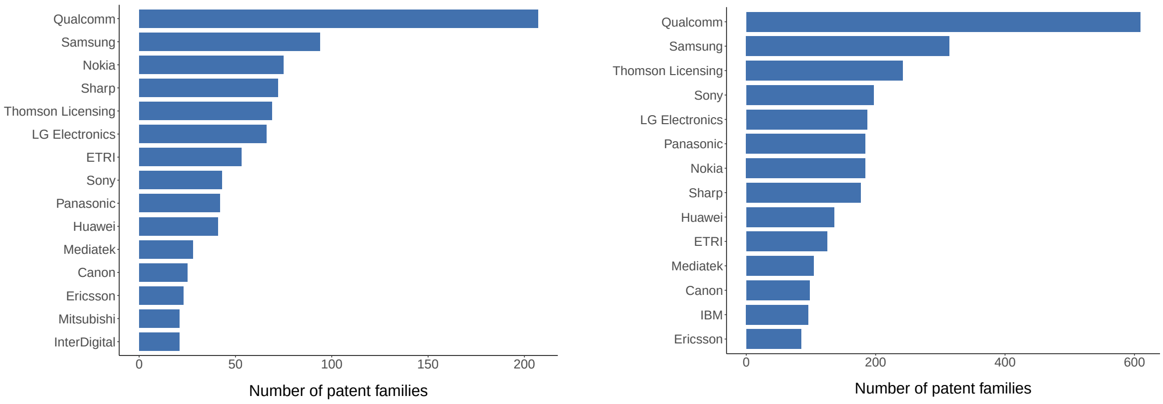
Notes: The graphs show the out-of-sample predictions on firm-level for patents that relate to the IEEE 802.11 (WiFi) standard. The numbers on the left-hand side of the bars indicate the patent family count. The top figures shows predictions for patents that are among the 250 most similar patents for any standard text that relates to the standard. In the bottom left-hand graph the 1000 most similar ones and in the bottom right-hand graph all patents in the data (i.e. the 3000 most similar patent families) are considered. We presume patents with predicted probabilities greater than 0.5 as standard essential. 95% confidence intervals are shown.

Figure A-9: Patents by firm (ITU-T)



Notes: The graphs show the number of patent families by patent applicant. We exclude individual patent applicants. On the left-hand side we consider patents that are among the 100 most similar patents for a given ITU-T standard document. On the right-hand side, the 500 most similar patents are shown.

Figure A-10: Patents by firm (ITU-T, H.264 standards)



Notes: The graphs show the number of patent families by patent applicant. We exclude individual patent applicants. On the left-hand side we only consider patents that are among the 100 most similar patents for a given standard document that relates to the H.264 (MPEG) standards family. On the right-hand side, the 500 most similar patents are shown.

Appendix B: Tables

Table B-1: Comparison of different manual SEP assessment studies

Study	Year	Standards	Patent count	Metric wording
Article One	2012	LTE	3,116	Highly essential
Concur IP	2017	GSM, UMTS, LTE	>5,200	
Cyber Creative Institute	2011	LTE	1,147	Really essential
Cyber Creative Institute	2012	LTE	1,601	Truly essential
Cyber Creative Institute	2013	LTE	2,129	Truly essential
Fairfield	2005	UMTS		
Fairfield	2008	UMTS	380	
Fairfield	2010	LTE, SAE	1,115	Probably essential
Jefferies	2011	LTE	1,400	Essential
iRunway	2012	LTE	4,599	Seminal
PA Consulting	2016	LTE	4,628	Probably essential

Notes: Table adapted and amended from Mallinson (2017).

Table B-2: T-statistics for the comparison of ETSI 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-3: Summary statistics (manual ETSI 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-4: Predicting standard essentiality (ETSI)

	(1)	(2)	(3)
	LTE	UMTS	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 was deemed essential by the evaluators. 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-5: Confusion matrix (ETSI)

		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-6: Predicting standard essentiality with most relevant characteristics (ETSI)

	(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: These specifications are used for out-of-sample predictions presented in Section 5.3. The dependent variable is a dummy equal to one if the patent family was deemed essential by the evaluators. 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-7: Summary statistics (ETSI SEP sample)

	Mean	SD	Median	Min	Max	N
Similarity score	326.7000	119.1000	315	48	782	14713
Similarity rank	285.3000	481.1000	80	1	3000	14713
Granted US patent	0.6610	0.4740	1	0	1	17240
# US fwd. cit. (5yrs)	30.9600	49.1800	15	0	1392	15756
# SEP US fwd. cit. (5yrs)	3.8280	7.2790	1	0	122	17247
# Independent claims	3.9290	2.4930	3	1	19	12308
Length claim 1	133.7000	62.9500	123	1	398	12189
Patent family size	8.3140	10.1000	6	1	472	17240
# Patent references	18.1600	28.5900	11.50	0	962	17240
# NPL references	15.1600	105.1000	4	0	12854	17240
# Applicants	1.8410	1.6540	1	1	20	16918
# Inventors	3.0010	1.7440	3	1	19	16902
Section-specific declaration	0.2470	0.4310	0	0	1	17247
Earliest declaration year	2010.6000	4.6130	2011	1900	2017	15680
Earliest filing year	2005.5000	6.6910	2007	1950	2016	17038

Notes: Summary statistics for patent characteristics of SEPs declared at ETSI. Patent characteristics are at patent family level.

Table B-8: Summary statistics (IEEE SEP sample)

	Mean	SD	Median	Min	Max	N
Similarity score	263.2510	81.7290	252	79	582	650
Similarity rank	867.2950	897.1150	513	1	2970	650
# US fwd. cit. (5yrs)	22.9080	30.1080	14	0	281	650
# SEP US fwd. cit. (5yrs)	1.0170	2.2030	0	0	23	650
# Independent claims	3.9540	2.5280	3	1	19	388
Length claim 1	126.8110	72.4450	111	1	393	391
Patent family size	6.1120	5.8830	5	1	66	650
# Patent references	21.1170	30.6110	13	0	300	650
# NPL references	20.0490	72.9040	4	0	727	650
# Applicants	1.6810	1.4740	1	1	12	649
# Inventors	2.6720	1.7300	2	1	11	650
Earliest declaration year	2006.4910	6.0760	2007	1988	2019	650
Earliest filing year	1999.4580	7.1870	1999	1977	2017	650

Notes: Summary statistics for patent characteristics of SEPs declared at IEEE. Patent characteristics are on patent family level.

Table B-9: Logistic regressions: Standard essentiality (IEEE) – Patent level

	(1)	(2)	(3)	(4)	(5)
Similarity score (cens.)		0.0015***	0.0018**	0.0013	0.0033
		(0.0004)	(0.0007)	(0.0008)	(0.0029)
# Independent claims	0.0260**	0.0294**	0.0325*	0.0456*	0.0924
	(0.0126)	(0.0138)	(0.0197)	(0.0263)	(0.0848)
Length claim 1	0.0000	0.0000	−0.0001	−0.0004	−0.0005
	(0.0003)	(0.0003)	(0.0004)	(0.0005)	(0.0010)
# Inventors	0.0094	0.0073	0.0010	0.0085	−0.0610
	(0.0240)	(0.0239)	(0.0343)	(0.0463)	(0.1077)
# Applicants	0.0253	0.0178	0.0270	0.0806	0.2182
	(0.0235)	(0.0235)	(0.0343)	(0.0570)	(0.1485)
Patent family size	0.0023	0.0039	0.0032	0.0083	−0.0756
	(0.0027)	(0.0028)	(0.0036)	(0.0058)	(0.0464)
# Patent references	0.0008	−0.0006	−0.0012	−0.0041	−0.0207
	(0.0016)	(0.0019)	(0.0027)	(0.0034)	(0.0133)
# NPL references	−0.0011	−0.0009	−0.0011	0.0005	0.0059
	(0.0014)	(0.0012)	(0.0014)	(0.0015)	(0.0047)
# SEP US fwd. cit. (5yrs)	0.0110	0.0115	0.0115	−0.0163	0.1065
	(0.0227)	(0.0231)	(0.0313)	(0.0356)	(0.1278)
Earliest filing year	No	No	Yes	Yes	Yes
Earliest declaration year	No	No	No	Yes	Yes
CPC-3 FE	No	No	No	No	Yes
Pseudo R^2	0.07	0.15	0.21	0.34	0.50
AUC	0.69	0.77	0.79	0.86	0.92
Observations	144	144	114	92	58

Notes: The dependent variable is a dummy equal to one if the patent family was deemed essential by the evaluators for IEEE standards. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard text for the standard specified in the SEP declaration are selected for the regressions. The data are on patent family level. Marginal effects of one unit change are reported. The sample size varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table B-10: Logistic regressions: Standard essentiality (WiFi sample)

	(1)	(2)	(3)	(4)	(5)
Similarity score (cens.)		0.0016***	0.0000	0.0006	0.0065
		(0.0005)	(0.0005)	(0.0025)	(0.0289)
# Independent claims	0.0129	0.0115	0.0002	−0.0002	0.0483
	(0.0103)	(0.0107)	(0.2398)	(0.0315)	(0.5117)
Length claim 1	−0.0005	−0.0003	0.0000	−0.0005	−0.0071
	(0.0003)	(0.0003)	(0.0018)	(0.0036)	(0.0483)
# Inventors	0.0389	0.0434	−0.0000	0.0343	−0.0161
	(0.0357)	(0.0337)	(0.0177)	(0.1180)	(0.0619)
# Applicants	0.1258***	0.0650	0.0001	−0.0114	−0.0491
	(0.0487)	(0.0488)	(0.1342)	(0.0392)	(0.1891)
Patent family size	−0.0013	−0.0057	−0.0001	−0.0270	−0.0695
	(0.0090)	(0.0089)	(0.1728)	(0.1003)	(0.3526)
# Patent references	−0.0070*	−0.0058	0.0000	−0.0085	−0.0413
	(0.0037)	(0.0039)	(0.0586)	(0.0192)	(0.0529)
# NPL references	0.0056	0.0027	−0.0000	0.0089	0.0112
	(0.0043)	(0.0042)	(0.0170)	(0.0268)	(0.0278)
# SEP US fwd. cit. (5yrs)	0.0207	0.0267	0.0001	0.1693	1.4849
	(0.0241)	(0.0227)	(0.2037)	(0.5817)	(5.7201)
Earliest filing year	No	No	Yes	Yes	Yes
Earliest declaration year	No	No	No	Yes	Yes
CPC-3 FE	No	No	No	No	Yes
Pseudo R^2	0.17	0.27	0.78	0.73	0.69
AUC	0.76	0.87	0.98	0.98	0.97
Observations	122	122	62	50	43

Notes: The dependent variable is a dummy equal to one if the patent family was deemed essential by the evaluators for IEEE 802.11 (WiFi) standard specifications. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard text for the standard specified in the SEP declaration are selected for the regressions. Marginal effects of one unit change are reported. The sample size varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table B-11: Summary statistics (ITU-T SEP sample)

	Mean	SD	Median	Min	Max	N
Similarity score	287.4080	98.3990	283	32	669	1137
Similarity rank	859.3350	910.8540	455	1	2989	1137
# US fwd. cit. (5yrs)	10.5990	16.0180	6	0	206	1137
# SEP US fwd. cit. (5yrs)	0.5440	1.1680	0	0	12	1137
# Independent claims	3.9280	2.7060	3	1	17	679
Length claim 1	123.7560	70.7380	107	1	393	673
Patent family size	6.9420	7.8600	5	1	123	1137
# Patent references	15.8290	18.2200	11	0	154	1137
# NPL references	11.1730	29.3810	3	0	596	1137
# Applicants	1.7220	1.5490	1	1	14	1135
# Inventors	2.6850	1.7750	2	1	17	1137
Earliest declaration year	2008.6160	7.1630	2011	1983	2019	1130
Earliest filing year	2002.6720	8.7460	2005	1976	2017	1137

Notes: Summary statistics for patent characteristics of SEPs declared at ITU-T. Patent characteristics are on patent family level.

Appendix C: Robustness Checks

The semantic algorithm we rely on in this paper 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 as often used in the literature (see Chapter 3.1).

For a small subset of our data, we show that measuring standard essentiality using the common text-based approaches is relatively simple. We use the text mining package 'tm' in R 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 is then converted into a (sparse) document-term-matrix. Words are weighted by term frequency-inverse document frequency (tf-idf). We additionally remove very 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 with the corresponding standard documents identified by its engineers yielding 117,282 text-based comparisons for only 657 patent families. In Table C-1, we report logistic regression results for the full text comparison using the alternative similarity measures as described before. Table C-2 reports results when only patent claim texts are chosen for the semantic similarity calculation. Comparing the effect sizes of the similarity score measures in both tables, the effects are larger when also patent title, abstract and descriptions are taken into account supporting the importance of considering all patent text information.

We compute micro-average precision and recall scores. Including patent characteristics, we obtain 63.4% precision and recall using patent full text data. 62.7% are obtained when only claim texts are used. These values are comparable, yet slightly inferior to the similarity measure calculated based on the proprietary octimine algorithm.

Table C-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 one if the patent family was deemed essential by the evaluators for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard in the sample of manual SEP assessments are selected for the regressions. For patents the full text is taken. Similarity scores range from 0 to 1. 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 varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table C-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 one if the patent family was deemed essential by the evaluators for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and the most similar standard in the sample of manual SEP assessments are selected for the regressions For patent only claim texts are chosen. Similarity scores range from 0 to 1. 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 varies as observations are dropped when fixed effects are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Appendix D: Database Documentation

In this Appendix, we detail all variables that are part of the database on patents and standards. The order of the variables follows the entity-relationship diagram as presented in Figure 4.1.

Table: STD_DOC_META

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document ID
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each standard document.
Domain: Bigint
Source database: Generated within the process

STD_ID

Name: Unique ID of the standard
Also known as: Standard ID
Description: This is the unique ID for the standard, e.g. when there are mutliple versions or releases of the same standard.
Domain: Bigint
Source database: Generated within the process

INT_ID

Name: Internal ID of the document of the standard
Also known as: internal ID
Description: This is the ID of a document of the standard provided by the database source.
Domain: Characters
Source database: IEEE, ITU-T, ETSI

NB_PAGES_DOC

Name: Number of pages in the document
Also known as: n/a
Description: Gives the number of pages for the standard document.
Domain: Int
Source database: Generated within the process

STD_DOC_NAME

Name: Name of the document of the standard
Also known as: Standard document name
Description: This is the name of the standard document.
Domain: Characters
Source database: IEEE, ITU-T, ETSI

STD_NAME

Name: Name of the standard
Also known as: Standard name
Description: This is the name of the standard.
Domain: Characters
Source database:

VERSION

Name: Version of the standard document
Also known as: Standard version
Description: This is the version of the standard, e.g. the same standard can have multiple versions which are logically ordered.
Domain: Characters
Source database: ETSI; generated within the process

STD_TITLE

Name: Title of the standard
Also known as:
Description: Thuis is the title of the standard found in the actual document.
Domain: Characters
Source database: IEEE, ITU-T, ETSI

SSO

Name: Standard setting organization
Also known as: n/a
Description: Gives the corresponding standard-setting organization.
Domain: Characters
Source database: Generated within the process

SSO_ID_INT

Name: ID of a standard-setting organization.
Also known as: n/a
Description: The ID defines the first figure of the std_doc_id, std_id and std_ch_id.
Domain: Int
Source database: Generated within the process

STD_DOC_PUB_DATE

Name: Date of publication of the standards document
Also known as: Standard document publication date
Description: This refers to the official publication date specified on the document of the standard.
Domain: %Y-%m-%d
Source database: Generated within the process

STD_DOC_PUB_YR

Name: Year of publication of the standard document
Also known as: Standard document publication year
Description: This gives the publication year as specified in the standard document.
Domain: Int
Source database: Generated within the process

STD_EARLIEST_PUB_DATE

Name: The earliest date of publication of the standards
Also known as: Earliest standard publication date
Description: This is the publication date of the first version of the standard.
Domain: %Y-%m-%d
Source database: Generated within the process

STD_EARLIEST_PUB_YR

Name: The earliest year of publication of the standards
Also known as: Earliest standard publication year
Description: This is the publication year of the first version of the standard.
Domain: Int
Source database: Generated within the process

STATUS

Name: Status
Also known as: n/a
Description: This is the status of the standard document as specified in the provided datasources.
Domain: Characters
Source database: IEEE, ITU-T, ETSI

ABSTRACT

Name: Abstract
Also known as: n/a
Description: This is a summary of the document of the standard.
Domain: Longtext
Source database: IEEE, ITU-T, ETSI

DETAILS_LINK

Name: Link for details
Also known as: n/a
Description: This URL refers to official website of the standard providing further details on the standard.
Domain: Characters
Source database: IEEE, ITU-T, ETSI

PDF_LINK

Name: Link for PDF
Also known as: n/a
Description: Gives the URL to the download of the actual PDF document of the standard.
Domain: Characters
Source database: ITU-T, ETSI

A_FILENAME

Name: Name of the file
Also known as: n/a
Description: This is the filename of the actual document (usually of type PDF).
Domain: Characters
Source database: IEEE; generated within the process

FILE_TYPE

Name: The type of the file
Also known as: Document type
Description: This is the type of the document (e.g. .pdf, .doc, .docx, ...)
Domain: Characters
Source database: Generated within the process

STD_DOC_TYPE

Name: Type of the document of the standard
Also known as: Standard document type
Description: This refers to the type of the technical standard. E.g., TS, EN, ...
Domain: Characters
Source database: ETSI

Table: STD_CH_META

STD_CH_ID

Name: Unique ID of the chapter of the standard
Also known as: Standard chapter ID
Description: This is the primary key of the table STD_CH_META.
Domain: Bigint
Source database: Generated within the process

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document id
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each standard document.
Domain: Bigint
Source database: Generated within the process

STD_CH_NAME

Name: Name of the chapter of the standard
Also known as: Standard chapter name
Description: This is the unique name of the chapter of the standard.
Domain: Characters
Source database: Generated within the process

STD_CH_NUMBER

Name: Number of the chapter of the standard
Also known as: Standard chapter name
Description: This is the number of the chapter of the standard. It can be of type character (e.g. for appendices, annexes, ...).
Domain: Characters
Source database: Generated within the process

STD_CH_TITLE

Name: Title of the chapter of the standard
Also known as: Standard chapter title
Description: This is the title of the chapter of the standard as specified in the table of contents of the document.
Domain: Characters
Source database: Generated within the process

NB_PAGES_CH

Name: Number of pages of the chapter
Also known as: n/a
Description: Gives the number of pages of the chapter as shown in the table of contents of the standard document.
Domain: Int
Source database: Generated within the process

NB_LINES_CH

Name: The number of lines in the chapter
Also known as: n/a
Description: Ths gives the number of lines of the chapter as identified in the original PDF document.
Domain: Int
Source database: Generated within the process

Table: STD_DOC_TEXT

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document id
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each

standard document.
Domain: Bigint
Source database: Generated within the process

FULL_TEXT

Name: The full text of the standard document
Also known as: n/a
Description: This is the full text of the standard document. The text is not cleaned and may contain control characters (e.g. \n, \r, ...)
Domain: Longtext
Source database: IEEE, ITU-T, ETSI

NB_TERMS_FT

Name: Number of terms in the full text
Also known as: Number of words
Description: This gives the number of terms in the full text.
Domain: Int
Source database: Generated within the process

NB_SENTENCES_FT

Name: Number of sentences in the full text
Also known as: n/a
Description: Gives the number of sentences in the full text of the standard document.
Domain: Int
Source database: Generated within the process

Table: STD_CH_TEXT

STD_CH_ID

Name: Unique ID of the chapter of the standard
Also known as: Standard chapter ID
Description: This is a primary key of the tableSTD_CH_META.
Domain: Bigint
Source database: Generated within the process

CHAPTER_TEXT

Name: Text of the chapter
Also known as:
Description: This is the text of the chapter of the standard document. It may contain line breaks and other control characters.
Domain: Longtext
Source database: IEEE, ITU-T, ETSI

NB_TERMS_CH

Name: Number of terms in the chapter
Also known as: Number of words
Description: This gives the number of terms in the chapter text.
Domain: Int
Source database: Generated within the process

NB_SENTENCES_CH

Name: Number of sentences in the chapter
Also known as: n/a
Description: Gives the number of sentences in the chapter of the standard document.
Domain: Int
Source database: Generated within the process

Table: STD_SEC_META

STD_SEC_ID

Name: Unique ID of the section of the standard
Also known as: Standard section ID
Description: This is the primary key in the table and represents the unique ID for the section within the document of a standard as shown in the table of contents.
Domain: Bigint
Source database: Generated within the process

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document id
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each standard document.
Domain: Bigint
Source database: Generated within the process

STD_SEC_NAME

Name: Name of the section of the standard
Also known as: Standard section name
Description: This is the unique name of the section of the standard document.
Domain: Characters
Source database: Generated within the process

STD_SEC_NUMBER

Name: Number of the section of the standard
Also known as: Standard section number
Description: This is the number of the section as specified in the table of contents of the standard document. It may contain characters (e.g. section numbers of appendices, annexes, ...).

Domain: Characters
Source database: Generated within the process

NB_PAGES_SEC

Name: Number of pages of the section
Also known as: n/a
Description: Gives the number of pages of the section as shown in the table of contents of the standard document.
Domain: Int
Source database: Generated within the process

NB_LINES_SEC

Name: The number of lines in the section
Also known as: n/a
Description: Ths gives the number of lines of the section as identified in the original PDF document.
Domain: Int
Source database: Generated within the process

VOID

Name: Void section
Also known as:
Description: Indicates whether a section has become void over the process of standards development. Currently only available for ETSI.
Domain: Int 0..1
Source database: ETSI; generated within the process

Table: STD_SEC_TEXT

STD_SEC_ID

Name: Unique ID of the section of the standard
Also known as: Standard section ID
Description: This is the primary key in the table and represents the unique ID for the section within the document of a standard as shown in the table of contents.
Domain: Bigint
Source database: Generated within the process

SEC_TEXT

Name: Text of the section
Also known as:
Description: This is the text of the section of the standard document. It may contain line breaks and other control characters.
Domain: Longtext
Source database: IEEE, ITU-T, ETSI

NB_TERMS_SEC

Name: Number of terms in the section
Also known as: Number of words
Description: This gives the number of terms in the section text.
Domain: Int
Source database: Generated within the process

NB_SENTENCES_SEC

Name: Number of sentences in the section
Also known as: n/a
Description: Gives the number of sentences in the section of the standard document.
Domain: Int
Source database: Generated within the process

Table: STD_SEP_DECL

DECL_ID

Name: ID of the declaration
Also known as: Declaration ID
Description: The ID is not unique in the table, but unique on declaration level. The table is on patent-standard document level (if available).
Domain: Int
Source database: Generated within the process

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document id
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each standard document.
Domain: Bigint
Source database: Generated within the process

STD_ID

Name: Unique ID of the standard
Also known as: Standard ID
Description: This is the unique ID for the standard, e.g. when there are mutliple versions or releases of the same standard.
Domain: Bigint
Source database: Generated within the process

DOCDB_FAMILY_ID

Name: ID of the docdb patent family
Also known as: n/a
Description: Patent family of the underlying patent declared to the standard-setting process.

Domain: Int
Source database: DOCDB

APPLN_ID

Name: ID of the application
Also known as: n/a
Description: Unique ID of the patent application.
Domain: Int
Source database: DOCDB, PATSTAT

APPLN_AUTH

Name: The authority of the application
Also known as: Application authority, country
Description: Patent authority where the application was filed.
Domain: Characters 1...2
Source database: DOCDB

APPLN_NR

Name: The number of the application
Also known as: Application number
Description: Number issued by patent authority
Domain: Characters
Source database: DOCDB

PUBLN_AUTH

Name: The authority of publication
Also known as: Publishing office
Description: Patent authority that issued the publication of the application
Domain: Characters 1...2
Source database: DOCDB

PUBLN_NR

Name: The number of publication
Also known as: Publication number
Description: Number given by patent authority
Domain: Characters
Source database: DOCDB

DECLARANT

Name: The name of the declarant
Also known as: n/a
Description: This is the name of the organization filing the SEP declaration.
Domain: Characters 1...1000
Source database: Declaration databases (IEEE, ITU-T, ETSI)

DECL_DATE

Name: Date of the declaration
Also known as: Filing date
Description: This is the date the declaration was filed.
Domain: %Y-%m-%d
Source database: Declaration databases (IEEE, ITU-T, ETSI)

DECL_YR

Name: Year of the declaration
Also known as: Filing year
Description: The year the declaration was filed.
Domain: Int
Source database: Declaration databases (IEEE, ITU-T, ETSI)

EARLIEST_DECL_DATE

Name: The earliest date of the declaration of the SEP
Also known as: Earliest filing date
Description: This is the earliest date the SEP was declared to the standard-setting process.
Domain:%Y-%m-%d
Source database: Generated within the process

EARLIEST_DECL_YR

Name: The earliest year of the declaration
Also known as: Earliest filing year
Description: This is the earliest year the SEP was declared to the standard-setting process.
Domain: Int
Source database: Generated within the process

CONTACT_PERSON

Name: The contact person of the declaration
Also known as: n/a
Description: This is the contact person as indicated in the declaration letter.
Domain: Characters
Source database: Declaration databases (IEEE, ITU-T, ETSI)

LICENSING_ASSURANCE

Name: Licensing assurance
Also known as: n/a
Description: Provides further details on the licensing assurance.
Domain: Characters
Source database: Declaration databases (IEEE)

Table: STD_SIM_FT

DOCDB_FAMILY_ID

Name: ID of the docdb patent family
Also known as: n/a
Description: Unique ID of the patent family.
Domain: Int
Source database: DOCDB

STD_CH_ID

Name: Unique ID of the chapter of the standard
Also known as: Standard chapter ID
Description: This is the primary key of the table STD_CH_META.
Domain: Bigint
Source database: Generated within the process

SIM_FT

Name: Semantic similarity between patent and standard full text
Also known as: score, similarity score
Description: This is the semantic similarity score calculated for the given patent-standard pair.
Domain: 0–1000
Source database: Dennemeyer Octimine

RANK_FT

Name: Relative similarity of patent for standard full text
Also known as: rank, similarity rank
Description: Gives the relative similarity of the patent for the standard chapter.
Domain: 1–3000
Source database: Dennemeyer Octimine

Table: STD_SIM_CH

DOCDB_FAMILY_ID

Name: ID of the docdb patent family
Also known as: n/a
Description: Unique ID of the patent family.
Domain: Int
Source database: DOCDB

STD_DOC_ID

Name: Unique ID of the document of the standard
Also known as: Standard document id
Description: This is the primary key of the table STD_DOC_META and assigns a unique ID to each standard document.
Domain: Bigint
Source database: Generated within the process

SIM_CH

Name: Semantic similarity between patent and standard chapter
Also known as: score, similarity score
Description: This is the semantic similarity score calculated for the given patent-standard pair.
Domain: 0–1000
Source database: Dennemeyer Octimine

RANK_CH

Name: Relative similarity of patent for standard chapter
Also known as: rank, similarity rank
Description: Gives the relative similarity of the patent for the standard chapter.
Domain: 1–3000
Source database: Dennemeyer Octimine