Universitat Pompeu Fabra Barcelona

Mapping Patents to Technology Standards

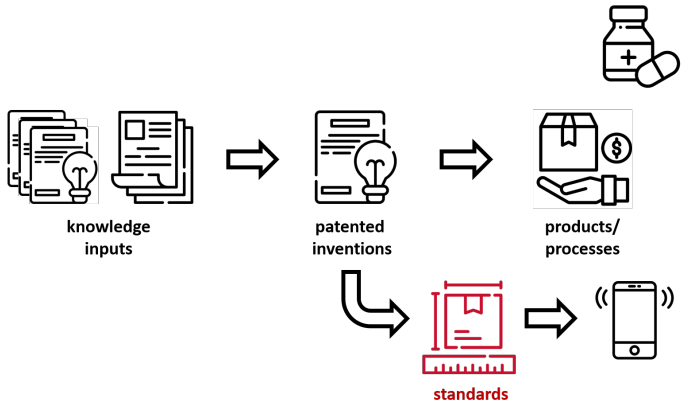
Lorenz Brachtendorf
Max Planck Institute for Innovation and Competition, Munich

Fabian Gaessler
Universitat Pompeu Fabra; Barcelona School of Economics;
Max Planck Institute for Innovation and Competition, Munich

Dietmar Harhoff
Max Planck Institute for Innovation and Competition, Munich; CEPR London

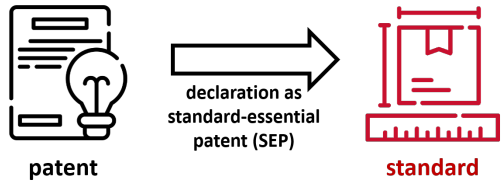
Innovation Information Initiative Technical Working Group Meeting
December 3, 2022

Motivation: from invention to innovation



- Technical specifications defining requirements for products/processes
- Published by standard-setting organizations (SSOs)

Linking patents to standards



- **Issues**
  - Overdeclaration (20-50% truly essential)
  - Underdeclaration (??%)
  - Blanket declarations (at portfolio level)
- **Alternative approaches**
  - Manual assessments → costly
  - Identifying relevant technology classes (Baron & Pohlmann, 2018) → coarse
  - Identifying relevant patents through citation network (Cho et al., 2021) → sparse

This project

We link patents to standards based on their *text-based semantic similarity*.

Outline

- **Method**
  - Standard essentiality ("relatedness") as a function of **semantic similarity** between patent publications and standard documents
  - Challenges in comparing patent and standard texts
- **Validation**
  - Comparison of similarity b/w SEP patent-standard pairs and matched controls
  - Replication of "disclosure effect" at ETSI (Bekkers et al., 2022)
  - Benchmark with manually assessed SEPs
- **Generalization**
  - Alternative algorithms to calculate semantic similarity
  - Standards from other SSOs
- **Database overview**

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

Standard document:
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 $j = 0, ..., l_{dtx} - 1, j \neq i$ within the CN averaging period (l_{dtx})."

Data sources and challenges

- **Standard documents**
  - ETSI's standards database with >40,000 standard documents
  - Documents describe multiple technologies and vary in size (>1,000 pages)
    → documents split at chapter level
- **Patent documents**
  - Around 18,000 declared SEPs (family level) ... + undeclared patents?
  - All patents with English publication from EPO, USPTO and WIPO (1980-2018)
- **Challenges**
  - Long texts
  - Many documents
  - Two distinct text corpora (structure, terminology, ...)

Approach to measure patent-standard semantic similarity

	Main	Alternatives	
Approach	octimine	tf-idf	embeddings
Open source	no	yes	yes
Libraries	n/a	tm (R), NLTK (Python)	TensorFlow, PyTorch
Reference	Natterer (2016)	Salton and Buckley (1988)	Devlin et al. (2019)
Algorithm			
Model	vector space model	vector space model	SciBERT
Pre-processing	stop-word removal, stemming, term reduction	stop-word removal, stemming, term reduction	–
Representation	latent semantic indexing	bag-of-words	document embeddings
Weighting	log-tf + entropy	tf-idf	SPECTER
Similarity metric	cosine	cosine	cosine
Patent corpus			
Sample	All	SEP subsample	SEP subsample
Documents	Most recent publication	Multiple publications	Multiple publications
Text input	Full text	Full text / no description / only claims	Full text

Approach to measure patent-standard semantic similarity

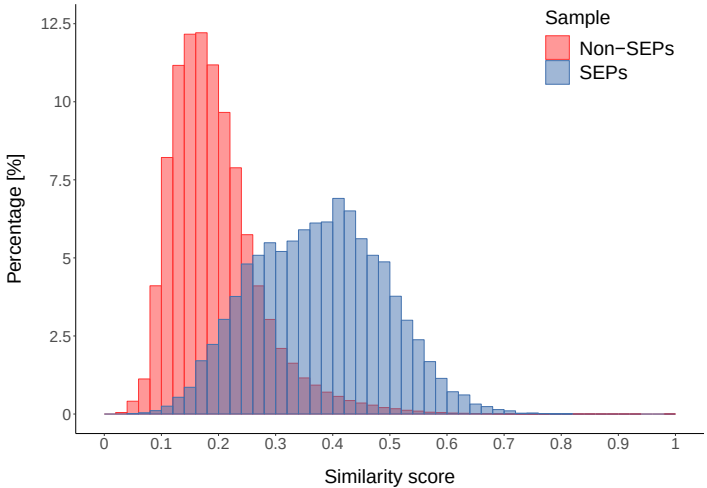
	Main	Alternatives	
Approach	octimine	tf-idf	embeddings
Open source	no	yes	yes
Libraries	n/a	tm (R), NLTK (Python)	TensorFlow, PyTorch
Reference	Natterer (2016)	Salton and Buckley (1988)	Devlin et al. (2019)
Algorithm			
Model	vector space model	vector space model	SciBERT
Pre-processing	stop-word removal, stemming, term reduction	stop-word removal, stemming, term reduction	–
Representation	latent semantic indexing	bag-of-words	document embeddings
Weighting	log-tf + entropy	tf-idf	SPECTER
Similarity metric	cosine	cosine	cosine
Patent corpus			
Sample	All	SEP subsample	SEP subsample
Documents	Most recent publication	Multiple publications	Multiple publications
Text input	Full text	Full text / no description / only claims	Full text

Summary statistics (patent level)

	Mean	Std. Dev.	Min	Median	Max
All patents (N=1,772,240)					
Similarity score	0.180	0.073	0.000	0.166	1.000
All SEPs (N=17,823)					
Similarity score	0.327	0.120	0.048	0.316	0.782
Assessed SEPs (N=2,287)					
Similarity score	0.314	0.114	0.048	0.300	0.758
Similarity score (tf-idf)	0.259	0.138	0.018	0.228	0.895
Similarity score (embeddings)	0.626	0.088	0.399	0.622	0.865

Notes: Summary statistics for *similarity score* for all patents that belong to the 3,000 most similar ones to any ETSI standard, the subset of all SEPs, and the subset of assessed SEPs. The similarity score is calculated at standard chapter level and collapsed to the max value at the patent level. Similarity score has a theoretical range between 0 and 1.

Similarity score distribution: non-SEPs vs. SEPs

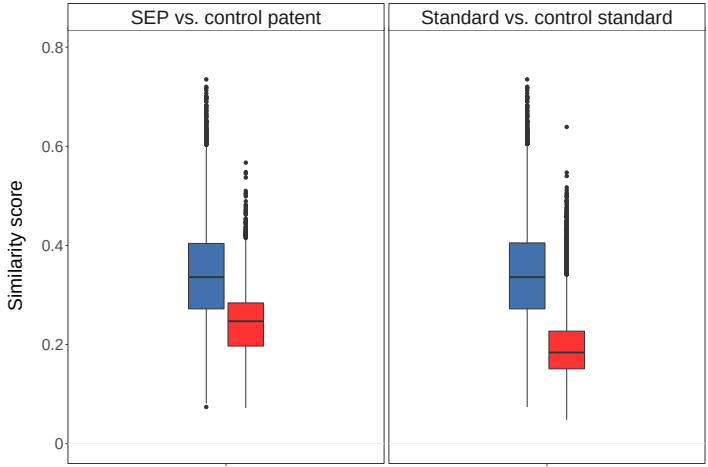


Notes: This figure shows the similarity score distribution for two different sets of patents. All non-SEP patents in the full sample (red bars) are compared to the set of SEPs declared at ETSI (blue bars).

Internal and external validation

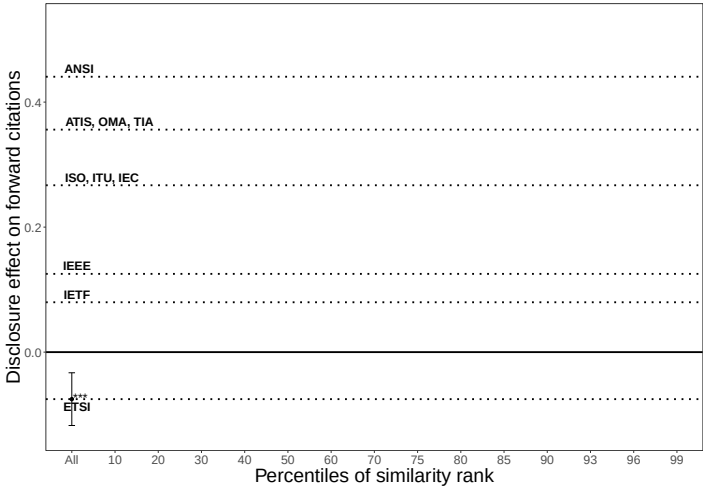
1. Similarity comparison between SEP-standard pairs and control groups
2. Replicating the ETSI "disclosure effect" (Bekkers et al., 2022)
3. Benchmark with dataset on manually assessed SEPs (true essentiality)

Comparison of SEP-standard pairs with matched control pairs



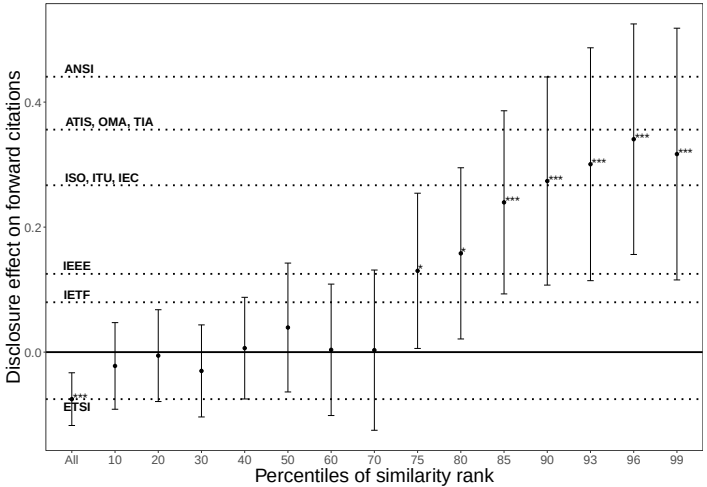
Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and matched control patents compared to the same standard (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and matched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

Effect of SEP declaration on forward cites (Bekkers et al., 2022)



Notes: Poisson estimates and 90% confidence intervals are shown. Each point corresponds to a separate regression coefficient. Standard errors are clustered on patent level. Significance levels: * p<0.1, ** p<0.05, *** p<0.01. The dotted horizontal lines show the effect sizes at other SSOs, as measured by Bekkers et al., 2022.

Effect of SEP declaration on forward cites (Bekkers et al., 2022)



Notes: Poisson estimates and 90% confidence intervals are shown. Each point corresponds to a separate regression coefficient. Standard errors are clustered on patent level. Significance levels: * p<0.1, ** p<0.05, *** p<0.01. The dotted horizontal lines show the effect sizes at other SSOs, as measured by Bekkers et al., 2022.

Benchmark with 2,300 manually assessed SEPs

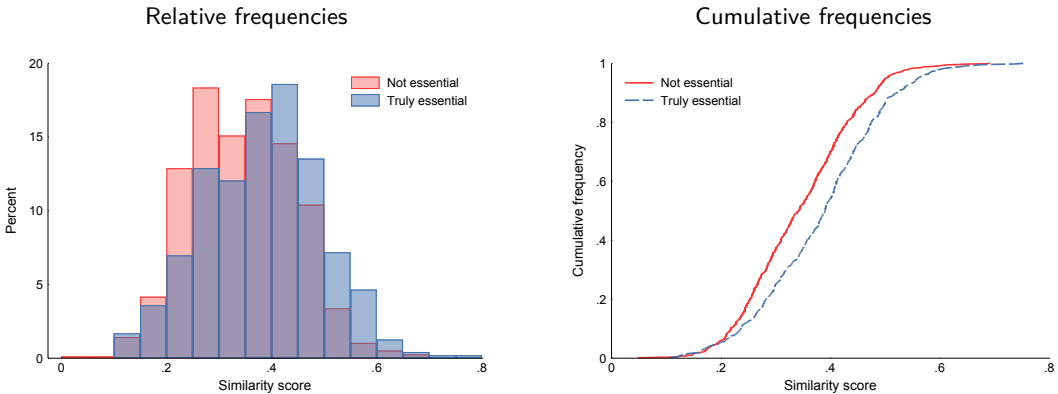
UNITED STATES DISTRICT COURT
CENTRAL DISTRICT OF CALIFORNIA

TCL COMMUNICATION TECHNOLOGY HOLDINGS, LTD., *et al.*, Plaintiffs/Counterclaim-Defendants, v. TELEFONAKTIEBOLAGET LM ERICSSON, *et al.*, Defendants/Counterclaim-Plaintiffs,	CASE NO: SACV 14-341 JVS(DFMx) Consolidated with CASE NO: CV 15-2370 JVS(DFMx)

FILED
CLERK, U.S. DISTRICT COURT
December 21, 2017
CENTRAL DISTRICT OF CALIFORNIA
BY: Karla Tunis DEPUTY

Public Redacted Document

Distribution of similarity scores of declared SEPs by assessment outcome



Notes: The left-hand graph shows the similarity score distributions of the two subsets of assessed SEPs declared to ETSI LTE standards: not essential SEPs (red bars) and truly essential SEPs (blue bars). The right-hand graph shows cumulative frequencies for both subsets.

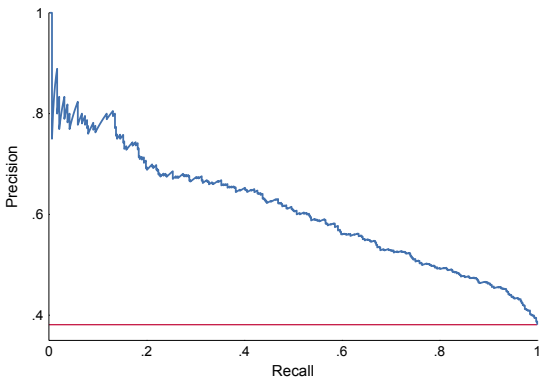
Logistic regressions: true standard essentiality

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score	0.8258***	0.6993***	0.5441***	0.3454**	0.5115***
	(0.1319)	(0.1382)	(0.1649)	(0.1746)	(0.1527)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest Decl. Year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo R^2	0.024	0.053	0.124	0.156	0.120
AUC	0.606	0.667	0.730	0.755	0.726
Observations	1,241	1,241	1,241	1,241	1,241

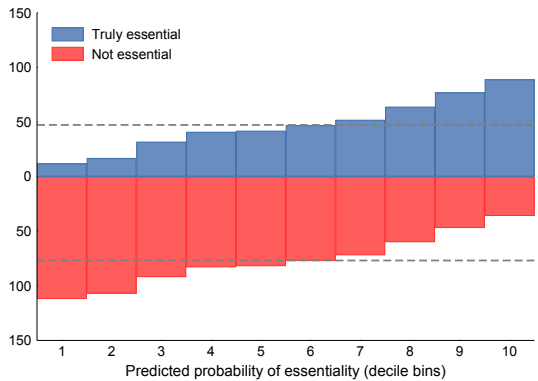
Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. Mean value of dependent variable: 0.381. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Predictive performance

Precision-recall curve



Essentiality status by predicted essentiality



Notes: The left-hand graph illustrates the precision-recall curve. The right-hand graph visualizes the composition of SEPs by essentiality status within bins of the predicted probability of true essentiality. Visualization adopted from Baron & Pohlmann (2021).

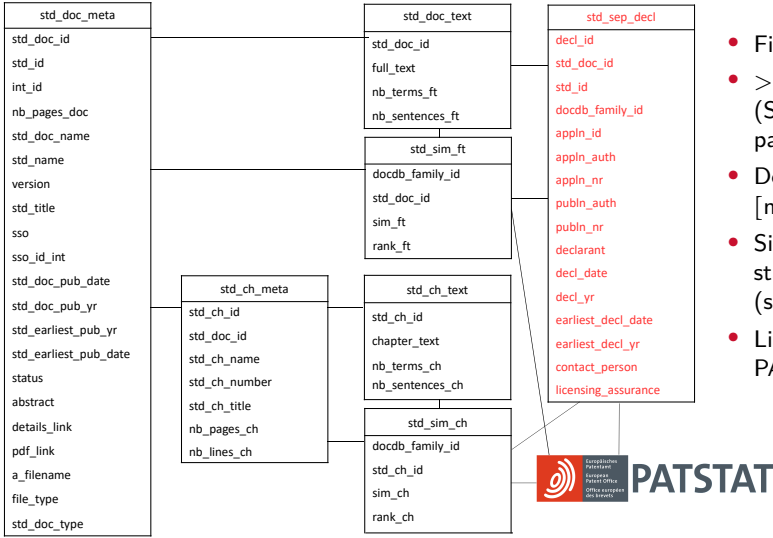
Generalizability of approach

- **Alternative patent text inputs**
  - Full text vs. without description vs. only claims ✓
  - Earliest vs. latest publication in patent family ✓
- **Alternative algorithms**
  - tf-idf ✓
  - embeddings (✓)
- **Extension to other standards**
  - IEEE standards ✓
  - ITU-T standards ✓

Brachtendorf et al. (2020). Approximating the standard essentiality of patents. *EPO Academic Research Programme*, June 12, 2020.

Brachtendorf et al. (2022). Truly standard-essential patents? A semantics-based analysis. *Journal of Economics & Management Strategy*, 20(2), 589-624.

Database: Semantic similarity of patent-standard pairs (ETSI, IEEE, ITUT)



- Files in tab-delimited *.csv format
- >60,000 standard documents (SSO, title, version, publication date, pages, url, ...) + chapter info
- Document/chapter texts [n/a in Harvard Dataverse]
- Similarity info for >200 million standard document-patent family pairs (similarity score, similarity rank)
- Link to SEP declarations and PATSTAT [external]

Harvard Dataverse:
doi.org/10.7910/DVN/B2RJSX

Conclusion

Summary

- New method to link patents to standards based on semantic similarity
- Method robust to different algorithms and standard technologies
- Similarity information of >200 million standard-patent pairs made available

Use cases

- SEP litigation (Helmers & Love, 2022) and portfolio licensing (Baron & Pohlmann, 2021)
- Strategic patenting (Righi & Simcoe, 2022)
- Contributions to technological progress (at firm/sector/country level)

Next steps

- Refinement of similarity measure
- Enlargement to other standards
- ...

Universitat Pompeu Fabra Barcelona

Thank you for your attention!

Fabian Gaessler
Department of Economics and Business
Universitat Pompeu Fabra
fabian.gaessler@upf.edu

Standard-essential patents (SEPs)

- **Definition**
  - SEPs protect inventions that are part of technical standards
  - Standard implementation requires SEP licenses, otherwise infringement
- **Licensing SEPs**
  - Multiple SEPs for one component/technology
  - Multiple SEP owners

  → Problem of *royalty stacking*
- **How licensing fees are calculated**
  - Bottom-up approach:
    sum of fees for individual components determine aggregate royalty rate
  - Top-down approach:
    overall licensing fees distributed **in proportion to relevant SEPs**

⇒ **SEPs: Everybody wants some!!**

Interpretation of semantic similarity

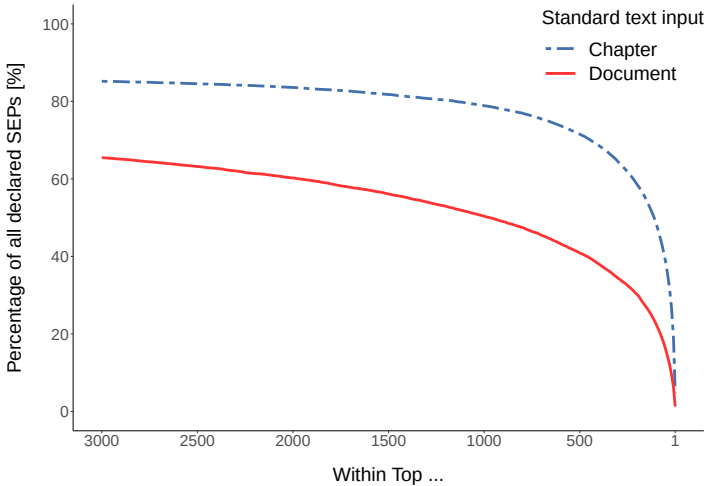
We interpret the semantic similarity between patents and standards as a measure of their **technological relatedness**.

- First, both patent and standard documents are highly technical texts and can be reasonably compared to each other.
- Second, standard documents are utilized by patent examiners, patent attorneys and inventors alike, which underlines their role as informative descriptions of technological solutions.

Technological relatedness is a necessary but not a sufficient condition for standard essentiality.

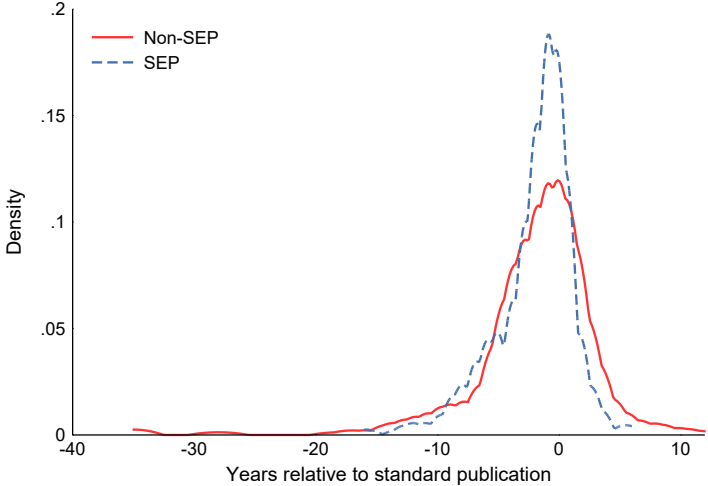
By and large, technological relatedness should be a fair proxy of standard essentiality.

Aggregate share of SEPs by rank



Notes: This graph shows the aggregate share of SEPs by their similarity rank at document level (red line) and chapter (blue line) level.

Time difference between patent filing and standard publication



Notes: This graph shows the time difference between patent (priority) filing and standard publication in years relative to standard publication. We distinguish between SEPs (blue) and non-SEPs (red) with high similarity (i.e., a similarity score of at least 0.6 and a similarity rank of 5 or better.

Patent characteristics of SEPs and non-SEPs with high similarity score

High similarity patents	SEPs (N = 186)			Non-SEPs (N = 239)				
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Diff.	p-value
Priority year	2006.70	2007.00	3.73	2006.26	2008.00	5.44	-0.44	0.342
Patent family size	9.62	7.00	7.88	5.04	4.00	3.47	-4.59	0.000***
# Applicants	1.83	1.00	1.39	1.78	1.00	1.44	-0.05	0.717
# Inventors	2.46	2.00	1.35	2.46	2.00	1.35	0.00	0.996
Corporate applicant	0.94	1.00	0.25	0.86	1.00	0.35	-0.07	0.014**
SEP-holding applicant	1.00	1.00	0.00	0.84	1.00	0.37	-0.16	0.000***
# Patent references	19.23	13.00	28.45	15.03	11.00	21.81	-4.20	0.086*
# NPL references	23.73	10.00	44.40	8.47	4.00	20.32	-15.25	0.000***
# Claims	18.87	18.00	10.97	20.58	20.00	11.18	1.71	0.138
Length claim 1	110.53	98.00	62.00	116.06	97.00	62.79	5.53	0.394
# US fwd. cit. (5yrs)	34.38	18.00	47.34	25.08	14.00	30.78	-9.31	0.017**
# SEP US fwd. cit. (5yrs)	6.95	3.00	10.20	3.24	1.00	6.07	-3.71	0.000***
Patent transferred	0.04	0.00	0.20	0.02	0.00	0.13	-0.03	0.105
Years betw priority filing and std publ.	2.56	2.00	2.80	3.48	2.00	4.61	0.91	0.018**

Notes: This table compares patent characteristics between SEPs and non-SEPs with high similarity (i.e., a similarity score of at least 0.6 and a similarity rank of 5 or better. The unit of observation is at the patent level. Reported **p**-values based on an unpaired **t**-test. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Predictive performance overview

	Similarity score			Similarity score (tf-idf)		
Model	Precision	Recall	F1	Precision	Recall	F1
Logit	0.65	0.66	0.65	0.64	0.66	0.64
NB	0.62	0.63	0.57	0.60	0.63	0.57
SVM (Linear)	0.64	0.65	0.59	0.63	0.64	0.57
SVM (Poly)	0.64	0.65	0.60	0.63	0.65	0.59
SVM (Radial)	0.63	0.64	0.60	0.64	0.65	0.61
XGB	0.64	0.65	0.63	0.64	0.65	0.63
RF	0.63	0.64	0.59	0.63	0.64	0.60

Notes: We use the Lasso specification (column 5) to compare five different machine learning models for the prediction of true standard essentiality. The logistic classifier *Logit* is our main model and yields the highest scores for all three performance metrics. *NB* is the Naive Bayes classifier and *SVM* is the support vector machine. To account for potential non-linear separability, we use *radial* and *polynomial* kernel functions for the *SVM* classifier. We further report the performance of two ensemble algorithms: *RF* is a random forest and *XGB* is the Extreme Gradient Boosting algorithm. Scores are weighted according to the frequency of the corresponding classes.

Logistic regressions: true standard essentiality (patent text input)

	DV: SEP truly essential (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
Patent publication:	Latest			Earliest		
Patent text:	Full	No desc	Claims	Full	No desc	Claims
Similarity score (tf-idf)	0.6266***	0.4832***	0.4328***	0.5719***	0.3706***	0.3710**
	(0.1158)	(0.1328)	(0.1316)	(0.1195)	(0.1438)	(0.1452)
Controls	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso
Pseudo R^2	0.120	0.122	0.118	0.120	0.112	0.112
AUC	0.727	0.729	0.726	0.726	0.720	0.720
Observations	1,241	1,241	1,241	1,241	1,241	1,241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. The similarity (*Similarity score (tf-idf)* is based on different text input: different patent publications and input text. Concerning the patent publication, either the latest publication (*Latest*) or the earliest publication (*Earliest*) in the DOCDB patent family is chosen. Concerning the patent text either full text (*Full*), full text without description (*No desc*), or claims text only (*Claims*) is chosen. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Control variable categories are collapsed to one (*Controls*). Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Logistic regressions: true standard essentiality

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score (tf-idf)	0.7046***	0.6624***	0.5838***	0.4492***	0.6266***
	(0.1043)	(0.1091)	(0.1245)	(0.1308)	(0.1158)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest Decl. Year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo R^2	0.028	0.060	0.131	0.161	0.120
AUC	0.612	0.665	0.735	0.760	0.727
Observations	1,241	1,241	1,241	1,241	1,241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. *Similarity score (tf-idf)* and *Similarity score (embeddings)* are based on two alternative open source algorithms: tf-idf and embeddings. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Logistic regressions: true standard essentiality

	DV: SEP truly essential (d)				
	(6)	(7)	(8)	(9)	(10)
Similarity score (embeddings)	0.7983***	0.6651***	0.5347*	0.4365	0.5502**
	(0.2509)	(0.2555)	(0.2782)	(0.2961)	(0.2670)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority Year FE	No	No	Yes	Yes	Lasso
Earliest Decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo R^2	0.006	0.042	0.120	0.155	0.112
AUC	0.556	0.641	0.726	0.754	0.720
Observations	1,241	1,241	1,241	1,241	1,241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. *Similarity score (tf-idf)* and *Similarity score (embeddings)* are based on two alternative open source algorithms: tf-idf and embeddings. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.