JOURNAL OF EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 17, Issue 3, 615–639, September 2020

Patent Similarity Data and Innovation Metrics

Ryan Whalen, Alina Lungeanu, Leslie DeChurch, and Noshir Contractor*

We introduce and describe the Patent Similarity Dataset, comprising vector space model-based similarity scores for U.S. utility patents. The dataset provides approximately 640 million pre-calculated similarity scores, as well as the code and computed vectors required to calculate further pairwise similarities. In addition to the raw data, we introduce measures that leverage patent similarity to provide insight into innovation and intellectual property law issues of interest to both scholars and policymakers. Code is provided in accompanying scripts to assist researchers in obtaining the dataset, joining it with other available patent data, and using it in their research.

I. Introduction and Overview

For decades, empirical research on patent law and innovation has benefited from access to increasingly high-quality patent datasets. Scholars have used these datasets to study innovation in a wide variety of contexts at the national level,[1] firm level,[2] team level,[3] and

*Address correspondence to Ryan Whalen, NUS Faculty of Law, 469G Bukit Timah Road 259776, Singapore; email: ryanwhalen@nus.edu.sg. Whalen is Assistant Professor of Law at the National University of Singapore, Singapore.; Lungeanu is Research Assistant Professor at Northwestern University School of Communication, Evanston.; DeChurch is Professor at Northwestern University School of Communication, Evanston.; Contractor is the Jane S. & William J. White Professor of Behavioral Sciences in the School of Engineering, School of Communication and the Kellogg School of Management, Evanston.

This research was supported by NSF award number 1856090.

[1]See, e.g., Raffaele Paci, Antonio Sassu & Stefano Usai, International Patenting and National Technological Specialization, 17 Technovation 25 (1997).

[2]See, e.g., Michele Grimaldi, Livio Cricelli, Martina Di Giovanni & Francesco Rogo, The Patent Portfolio Value Analysis: A New Framework to Leverage Patent Information for Strategic Technology Planning, 94 Technological Forecasting & Soc. Change 286 (2015).

[3]See, e.g., Margherita Balconi, Stefano Breschi & Francesco Lissoni, Networks of Inventors and the Role of Academia: An Exploration of Italian Patent Data, 33 Res. Pol'y 127 (2004).

individual level.[4] In these studies, patent data have served a wide variety of purposes.[5] For example, citations have been used as a proxy for knowledge inputs or measure of a patent's value,[6] patents themselves have been used as proxy measures for innovation more generally,[7] and the structure of the prior art citation network has been used to infer the existence of thickets of intellectual property rights.[8]

This project engages with the tradition of providing patent-related data that can enrich future research on patent law and innovation. We begin by briefly reviewing the state of available patent data, and the research that relies on it. We subsequently introduce the Patent Similarity Dataset, which uses a vector space model to compute pairwise distances between a large number of patents. After introducing vector space models generally, and explaining how the Patent Similarity Dataset was created, this article's final section goes on to describe the Patent Similarity Dataset's qualities and demonstrate how it can be used to generate a wide variety of metrics that provide new perspective on patent law and innovation.

A. The Growth in Patent Data Availability, and Patent-Data-Driven Research

One of the functions of patent law is to incentivize the disclosure of information relating to innovation.[9] As a result of this, the patent system generates a large amount of data, much of which is publicly available.[10] For decades now, researchers have been drawing on this increasingly large body of available patent data to help better understand innovation, science, and intellectual property law. Because the universe of patent data is quite large and data are available in varying formats, many of these projects require substantial data cleaning and preparation work. Researchers thus often publish their datasets both

---

[4]See, e.g., Martin G. Moehrle, Lothar Walter, Anja Geritz & Sandra Müller, Patent-Based Inventor Profiles as a Basis for Human Resource Decisions in Research and Development," 35 R&D Mgmt. 513 (2005).

[5]For a review of patent data as an economic indicator, see Zvi Griliches, Patent Statistics as Economic Indicators: A Survey," in R&D and Productivity 287 (Univ. of Chicago Press 1998); Sadao Nagaoka, Kazuyuki Motohashi & Akira Goto, Patent Statistics as an Innovation Indicator, 2 Handbook of the Economics of Innovation 1083 (B. H. Hall & N. Rosenberg, eds., North-Holland 2010).

[6]See Manuel Trajtenberg, A Penny for Your Quotes: Patent Citations and the Value of Innovations," Rand J. of Econ. 172 (1990).

[7]Daron Acemoglu, Ufuk Akcigit & William R. Kerr, Innovation Network, 113 PNAS 11483 (2016).

[8]George von Graevenitz, Stefan Wagner & Dietmar Harhoff, How to Measure Patent Thickets—A Novel Approach, 111 Econ. Letters 6 (2011).

[9]Jeanne C. Fromer, Patent Disclosure, 94 Iowa L. Rev. 539 (2008–2009).

[10]Indeed, the Patent Act requires that the PTO make patent data available. 35 U.S.C § 41(i). For an overview of IP data, see David L. Schwartz & Ted Sichelman, Data Sources on Patents, Copyrights, Trademarks, and Other Intellectual Property, in Research Handbook on the Economics of Intellectual Property Law (Edward Elgar Publishing 2019).

to ensure that their efforts are put to wide use, and to enable others to replicate or potentially improve on their analyses.

At its most basic level, "patent data" refers to the data disclosed by the patent system. Traditionally, this has included the metadata on the first page of granted patents, which details things such as the title of the invention, the patent number, the technical classifications assigned to the invention, and inventor and assignee names. In more recent years, as dataset size has become less of a limiting factor for researchers, patent datasets have offered increasing levels of detail. For instance, the NBER patent citation data are a well-known and widely used patent dataset. Since its publication in 2001, the patent citation data it provides has enabled a wide variety of innovation and IP metrics that have been used in thousands of academic articles across a wide variety of disciplines.[11]

Many other patent datasets have been created in recent decades, including those arising as a result of improved data sharing by the USPTO as well as those created by researchers interested in questions about patent law and policy. For instance, statutory changes at the turn of the century led to the publication of patent applications and the resulting access to patent prosecution data—a new type of patent data that was previously difficult to access.[12] In more recent years, the USPTO Office of the Chief Economist (OCE) has produced and disseminated a growing body of patent data.[13] Many of these initiatives build on or complement the work of researchers outside the Patent Office who have added value to patent data by cleaning and processing publicly available patent data, such as efforts to disambiguate inventors, thus allowing more nuanced analyses at the inventor or team level.[14]

In addition to datasets on patents themselves, the growth in patent-related data has also included data on patents in the legal system. For instance, Cotropia et al. have created and shared a dataset on the role patent assertion entities play in patent litigation.[15] Similarly, the OCE has used publicly available federal court data to assemble and share a patent litigation dataset,[16] and the Stanford Non-Practicing

---

[11]Bronwyn H. Hall, Adam Jaffe & Manuel Trajtenberg, The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools (National Bureau of Economic Research 2001).

[12]See 35 U.S.C § 122. See also Christopher A. Cotropia & David L. Schwartz, The Hidden Value of Abandoned Applications to the Patent System, SSRN Scholarly Paper ID 3465737 (Social Science Research Network Aug. 30, 2019).

[13]https://www.uspto.gov/learning-and-resources/ip-policy/economic-research/research-datasets

[14]Guan-Cheng Li, Ronald Lai, Alexander D'Amour, David M. Doolin, Ye Sun, Vetle Torvik, Amy Z. Yu & Lee Fleming, Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975–2010), 43 Res. Pol'y 941 (2014).

[15]Christopher A. Cotropia, Jay P. Kesan & David L. Schwartz, Unpacking Patent Assertion Entities (PAEs), 99 Minn. L. Rev. 649 (2014–2015).

[16]Alan C. Marco, Asrat Tesfayesus & Andrew A. Toole, Patent Litigation Data from US District Court Electronic Records (1963–2015), SSRN Scholarly Paper ID 2942295 (Social Science Research Network Mar. 1, 2017).

Entity (NPE) Litigation Database collates and shares data on patent litigation, with a special focus on the types of entity—for example, practicing or non-practicing—involved in the suit.[17]

The flourishing of available patent data has been accompanied by a commensurate flourishing of research utilizing those data. For instance, researchers have used patent datasets to infer patent value by proxies such as forward citation[18] or family size,[19] and to extrapolate these invention value measures to firm market value.[20] Patent data have also been used to identify patent thickets, with a variety of approaches such as using inter-firm citations to infer blocking rights,[21] or by examining citation network density.[22]

In addition to its use for assessing innovation at the patent or firm level, patent data have also helped shed light on the scientific and research processes more generally. This research—sometimes referred to as the *science of science* or, when focused on team processes, as *the science of team science*[23]—has used patent data to better understand both research inputs and outputs. For example, research has explored innovation inputs by drawing on patent classification data to explore how researchers engage in knowledge search and recombination,[24] and to better understand changes in the degree of interdisciplinarity in granted patents.[25] In addition to data about the contents of patents or the claimed invention, the information that patent data offer on inventors and where they

---

[17]Welcome to the Stanford NPE Litigation Database | NPE Litigation Database, https://npe.law.stanford.edu/.

[18]Trajtenberg, supra note 6; Dietmar Harhoff, Francis Narin, F. M. Scherer & Katrin Vopel, Citation Frequency and the Value of Patented Inventions, 81 Rev. Econ. & Stats. 511 (1999).

[19]Dietmar Harhoff, Frederic M. Schererc & KatrinVopeld, Citations, Family Size, Opposition and the Value of Patent Rights," 32 Res. Pol'y 1343 (2003); Dominique Guellec & Bruno van Pottelsberghe de la Potterie, Applications, Grants and the Value of Patent," 69 Econ. Letters 109 (2000).

[20]Bronwyn H. Hall, Adam Jaffe & Manuel Trajtenberg, Market Value and Patent Citations, RAND J. Econ. 16 (2005).

[21]von Graevenitz et al., supra note 8.

[22]Gavin Clarkson, Patent Informatics for Patent Thicket Detection: A Network Analytic Approach for Measuring the Density of Patent Space (ResearchGate 2005).

[23]Kate Börner, Noshir Contractor, Holly J. Falk-Krzesinski, Stephen M. Fiore, Kara L. Hall, Joann Keyton, Bonnie Spring, Daniel Stokols, William Trochim & Brian Uzzi, A Multi-Level Systems Perspective for the Science of Team Science," 2 Sci. Translational Med. 1 (2010).

[24]Lee Fleming & Olav Sorenson, Technology as a Complex Adaptive System: Evidence from Patent Data, 30 Res. Pol'y 1019 (2001); Lee Fleming & Olav Sorenson, Science as a Map in Technological Search," 25 Strategic Mgmt. J. 909 (2004).

[25]Xiaolin Shi, Lada A. Adamic, Belle L. Tseng & Gavin S. Clarkson, The Impact of Boundary Spanning Scholarly Publications and Patents, 4 PLoS One e6547 (2009).

work has furthered research on collaboration,[26] knowledge transfer,[27] and the effects of non-compete agreements on the labor market.[28]

Patent-data-driven research has also extended to examine the administration of patent systems. For instance, Frakes and Wasserman have used patent prosecution data to demonstrate systematic challenges facing patent examiners that contribute to the grant of low-quality patents.[29] Others have used patent prosecution data to illustrate how changes in the types of innovation might challenge examiners,[30] or to propose methods to improve the examination process.[31]

The above is by no means intended to be an exhaustive review of the research drawing on patent data. Indeed, because of the broad utility of patent data, any such review would be beyond the scope of a single article. Rather, the intent here is to highlight how useful patent data have been to researchers from a wide variety of disciplines including law, economics, sociology, management science, and more. This previous work has benefited from efforts by other researchers and by patent offices and NGOs to clean and curate increasingly detailed patent data.

Much of the past work focused on cleaning and sharing patent data is emblematic of the general rise in "metaknowledge" research.[32] As electronic publishing and indexing have increased in scope, researchers have used their increased access to metadata and improved analytic methods and power to improve our understanding of the scientific and creative processes. However, metadata studies are necessarily imprecise, making inferences about substance based on abstracted data. The metadata nature of many existing patent datasets has led researchers to rely on these coarse data when attempting to measure quite fine-grained concepts. For instance, analyses using patent citations tend to treat those citations as binary markers of influence or relationship. That is, the citation

---

[26]Stefan Wuchty, Benjamin Jones & Brian Uzzi, The Increasing Dominance of Teams in Production of Knowledge, 316 Sci. 1036 (2007).

[27]Adam B. Jaffe, Manuel Trajtenberg & Rebecca Henderson, Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations, 108 Q. J. of Econ. 577 (1993).

[28]Matt Marx, Deborah Strumsky & Lee Fleming, Mobility, Skills, and the Michigan Non-Compete Experiment, 55 Mgmt. Sci. 875 (2009); Matt Marx, Jasjit Singh & Lee Fleming, Regional Disadvantage? Employee Non-Compete Agreements and Brain Drain, 44 Res. Pol'y 394 (2015).

[29]Michael Frakes & Melissa F. Wasserman, Does the U.S. Patent & Trademark Office Grant Too Many Bad Patents? Evidence from a Quasi-Experiment, 67 Stanford Univ. L. Rev. 613 (2015); Michael D. Frakes & Melissa F. Wasserman, Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data," 99 Rev. Econ. & Statistics 550 (2016).

[30]Ryan Whalen, Boundary Spanning Innovation and the Patent System: Interdisciplinary Challenges for a Specialized Examination System," 47 Res. Pol'y 1334 (2018).

[31]Charles deGrazia, Nicholas A. Pairolero & Mike Teodorescu, Shorter Patent Pendency Without Sacrificing Quality: The Use of Examiner's Amendments at the USPTO, USPTO Econ. Working Paper No. 2019-03 (Social Science Research Network Jun. 1, 2019).

[32]James A. Evans Jacob G. Foster, Metaknowledge, 331 Sci. 721 (2011).

either exists or does not exist, leaving little room for qualitative distinction between different types of citations. Similarly, research using patent categorization to infer anything about the substance of the claimed invention necessarily treats all patents with the same classification as identical to one another. In reality, there is substantial variation both in types of prior art citations and within patent categories. Metadata elide much of that variation and in the process limit the capacity of the dataset.

Researchers have historically relied on metadata-based measures for a variety of reasons, including convenience, tractability, and limitations on access to more detailed data. However, improvements in data access, computational capacity, and natural language processing techniques mean that we can now engage more deeply with the content and substance of patent documents and need not be limited to metadata only when assembling research datasets. One of the clearest ways patent document contents can contribute to more nuanced research data is by using the text of the document to assess its content and determine how similar or dissimilar patents are from one another.

Patent similarity data have proved useful in a variety of research contexts. There is a relatively large body of work that has used patent similarity scores in the context of engineering and computer science research. For instance, recent work used patent vector space models to improve patent information retrieval[33] and recommender systems.[34] Other recent research has used patent similarity metrics to help with classification problems such as identifying standard essential patents,[35] or those patents involved in thickets.[36] Research using patent similarity metrics has also begun to appear more frequently in legal research, including work proposing similarity metrics as a supplementary measure of patent value,[37] and work that explores changes in workload at the USPTO.[38] Despite its growing popularity, the application of patent similarity scores in empirical research faces a number of challenges, including access to the computational resources to calculate the scores, operationalization of relevant metrics, and knowledge of how to

---

[33]André Rattinger, Jean-Marie Le Goff & Christian Guetl, Semantic and Topological Graphs for Patent Retrieval, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) 175 (Oct. 2019).

[34]Hyoung J. Kim, Tae S. Kim & So. Y. Sohn, Recommendation of Startups as Technology Cooperation Candidates from the Perspectives of Similarity and Potential: A Deep Learning Approach, 130 Decision Support Sys. 113229 (2020).

[35]Sven Wittfoth, Identification of Probable Standard Essential Patents (SEPs) Based on Semantic Analysis of Patent Claims, 2019 Portland International Conference on Management of Engineering and Technology (PICMET) 1 (Aug. 2019).

[36]Mateusz Gątkowski, Marek Dietlb, Łukasz Skrokb, Ryan Whalenc & Katharine Rocketta, Semantically-Based Patent Thicket Identification, 49 Res. Pol'y 103925 (2020).

[37]Jonathan H. Ashtor, Investigating Cohort Similarity as an Ex Ante Alternative to Patent Forward Citations, 16 J. Empirical Legal Stud. 848 (2019).

[38]Whalen, supra note 30; Ryan Whalen, Complex Innovation and the Patent Office, 17 Chi.- Kent J. Intell. Prop. 226 (2018).

apply them. The Patent Similarity Dataset shared here seeks to reduce these barriers to entry by sharing similarity data, the code necessary to reproduce them, and demonstrations of innovation measures that can be derived from the dataset.

II. The Dataset

The Patent Similarity Dataset provides document-level similarity measures for granted patents. These similarity scores provide insight into the degree of similarity in the linguistic content of patent pairs. Because patent text is largely comprised of a description of the claimed technology and specific claim language, patent similarity scores can be thought of as providing insight into the similarity of the inventions that are claimed in the documents. Once computed, these scores can be used to assess innovation and patent policy in a variety of novel ways. For instance, weighting prior art citations based on the similarity between the citing and cited documents allows for more nuanced measures of innovation input and patent impact. Similarly, one can use these scores to identify different types of innovation, such as that emerging from a single disciplinary field or those that span diverse areas of expertise.[39] These and other applications of patent similarity data will be demonstrated below after introducing vector-based models and describing how the dataset was assembled.

A. Vector Space Models

Vector space models are called that because they represent documents with a numerical vector. By representing documents as n-dimensional vectors, one can use vector and matrix analyses to gain insight into their relationships with one another. Corpora vector spaces are often represented as a matrix, with a row for each document and columns representing the relevant dimensions. There are a wide variety of methods to identify and measure model dimensions, and as natural language processing (NLP) methods develop, new methods are introduced with some regularity.

Perhaps the simplest method to situate documents in a vector space is to rely on vocabulary terms, representing each unique word with a column in the matrix. Doing so generates a matrix with n rows, and m columns, where n is the number of documents in the corpus and m is the number of unique terms in the corpus. Cell values can represent the number of times a term appears in each document. With this matrix in hand, one can quite simply compute vocabulary similarity measures by taking the cosine of the rows (i.e., vectors) for document pairs. A somewhat more nuanced, yet still relatively simple, vocabulary-based approach reweights terms based on the degree to which the term helps distinguish the document from other documents in the corpus. This term frequency-inverse document frequency (TF-IDF) measure takes into account the number of times each term appears in a document and the number of documents it appears in across the

[39]Whalen, supra note 30.

corpus. By reweighting term scores, TF-IDF helps strengthen the vocabulary signal and leads to somewhat better similarity scores.

However, these simple vocabulary-based measures have a number of weaknesses. Because each unique term in the corpus is represented as a matrix column, the matrix is very sparse. Perhaps more importantly, these methods do not account for varied relationships between words. Each unique term is treated as its own dimension, when in reality some terms are closely related to one another (e.g., "car" and "automobile") while others are not (e.g., "finance" and "calcium"). More nuanced models have been developed to address these weaknesses. They include methods such as latent semantic indexing (LSI), which applies singular value decomposition on the original document-term matrix, generating a lower-rank *n*-dimensional document-concept matrix that partially addresses term relatedness;[40] and Latent Dirichlet Allocation (LDA), a probabilistic model that assigns weighted probabilities that terms relate to topics or dimensions.[41]

Recent developments in machine learning techniques have led to further vector space model developments, including "deep learning" approaches to vector space modeling. To construct the patent distance dataset we use Doc2Vec—one of these more recent neural-network-based models—because of its wide adoption and performance advantages,[42] and because recent research suggests it performs well in identifying similar inventions.[43] Doc2Vec is an extension of the popular Word2Vec model, which represents words as embeddings (i.e., vectors) that enable sophisticated natural language processing tasks.[44] Using Doc2Vec, we can represent documents as vectors, enabling comparisons between documents.[45]

---

[40]Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer & Richard Harshman, Indexing by Latent Semantic Analysis, 41 JASIS 391 (1990).

[41]David M. Blei, Andrew Y. Ng & Michael I. Jordan, Latent Dirichlet Allocation," 3 J. Machine Learning Res. 993 (2003).

[42]Michal Campr & Karel Ježek, Comparing Semantic Models for Evaluating Automatic Document Summarization, Text, Speech, and Dialogue, in P. Král & V. Matoušek, eds., Lecture Notes in Computer Science 252 (Springer International Publishing 2015).

[43]lEA Helmers, Franziska Horn, Franziska Biegler, Tim Oppermann & Klaus-Robert Müller, Automating the Search for a Patent's Prior Art with a Full Text Similarity Search, 14 PLoS One (2019).

[44]Quoc V. Le & Tomas Mikolov, Distributed Representations of Sentences and Documents, arXiv:1405.4053 [cs] (2014).

[45]Word2Vec produces word vectors by using a three-layer neural network featuring an input layer, a hidden layer, and an output layer. There are different algorithmic applications of Word2Vec, but in each the essential function of the hidden layer is to predict words based on their context. The training process tunes the hidden layer to produce the most accurate predictions, based on the input layer. Doc2Vec extends this approach by adding additional input nodes that represent the documents. Thus, when the Doc2Vec training is complete, one has both the word embeddings as well as document embeddings. The similarity scores featured in the Patent Similarity Dataset rely on these document embeddings.

B. Creating the Dataset

To create a Doc2Vec model from a corpus of documents, one first needs the input corpus. In the case of patent documents, there are numerous sources of the full text of granted patents. For the purpose of generating the Patent Similarity Dataset, we drew on the data provided by the USPTO's Office of the Chief Economist (OCE). The OCE provides regular database dumps containing, among other patent data, the full text of patent description and claims sections.[46] This dataset covers all patents granted between 1976 and the end of 2019.

After downloading these data, we use the text of the utility patent documents[47]—comprising the description and independent claims text—as the input data for the Doc2Vec model.[48] The model estimates a 300-dimension vector representation for each input document. These vectors or "embeddings" can be thought of as points in multi-dimensional semantic space that represent the contents of each patent document.

Once the model is computed, we use the patent embeddings to calculate a variety of cosine similarity scores: the similarities between all citing/cited patent pairs, and the similarities of the 100 most similar patents to each patent in the dataset. These pre-calculated similarity scores are available for download, along with the Doc2Vec embedding vectors and pre-calculated model object that will enable further similarity score calculations for texts not used in the model generation (e.g., patent applications, or new patents not in the dataset). Interpreting vector space model similarity scores can be challenging. Research suggests that they generally agree with human readers' similarity assessments,[49] and we offer some validation techniques below and in the accompanying code notebooks to show that the similarity scores in the dataset track expectations,

---

[46]In addition to the similarity scores, this project also shares a Python script that automates the downloading and database assembly process for those who wish to work with the OCE patent data.

[47]We exclude non-utility patents such as plant and design patents because they differ in many ways from utility patents, and including them in the model training may reduce the model's accuracy. For the purpose of enabling longitudinal analysis, we also exclude reissued patents from the analyses below; however, their vectors can be calculated and included in results if researchers so desire.

[48]The model was computed using the Gensim Python library, using the distributed bag of words (DBOW) algorithm with 10 epochs. Radim Řehůřek & Petr Sojka . Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks 45 (ELRA Valletta, Malta May 2010). As in most vector space models, words are treated as "tokens" and thus separating text into words is a vital step of the process. For most patents doing so is straightforward. However, some inventions, such as those detailing chemical compounds, raise some challenges. For simplicity and consistency, we treated any whitespace-separated strings as tokens or words. Thus, chemical formulae with no space between components will be treated as a single word, while ones with whitespace between components will be treated as multiple words. We use both the description and independent claims to help ensure that the important text discussing each invention is included in the model calculation. All patents require at least one independent claim and, indeed, these claims are legally the most important part of the patent in describing its scope, so we include these. Meanwhile, the description field describes the invention in more general terms and contextualizes it, providing useful input for the model.

[49]See Peter D. Turney & Patrick Pantel, From Frequency to Meaning: Vector Space Models of Semantics, 37 J. Artificial Intell. Res. 141 (2010).

suggesting the metrics proposed below will be useful, especially when applied in the aggregate. Nonetheless, individual scores should be interpreted carefully.

III. Patent Similarity Data and Measures

Because patents are of interest to scholars from a wide variety of disciplines with a wide variety of interests, the Patent Similarity Dataset has a commensurately wide variety of potential applications. It can be used to provide insight at the individual patent level—for example, measures of impact or interdisciplinarity within a specific invention—at the inventor level—for example, the degree to which an inventor produces inventions that are similar to one another—and even at the firm or location level—for example, the degree of variation over time in a firm's patented invention output.

A. Comparing Patents by Classification

As an initial point of inquiry, one might wish to determine the degree to which semantic similarity tracks with existing measures used to infer a patent's contents or topic. Absent access to data that leverage patent text, metadata provide one of the best ways to infer what type of technology a patent claims. The PTO assigns classifications to each application as it is submitted. These classifications both determine which art unit will assess the application and help guide the prior art search.[50] Previous research has leveraged these classifications to infer the topical substance of the claimed technology.[51]

Figure 1 demonstrates how the semantic similarity data we share here track with these classification data. Here we see that patents paired at random are the least similar to one another, while those that are matched based on their cooperative patent classification (CPC) section are somewhat more similar to one another, those matched on class yet more similar, and those matched on subclass most similar.[52] This corresponds with what one would expect, as moving down the CPC classification tree successively narrows the topics covered and we would expect patents in narrower classes to have more in common.[53] At the same time, we also see substantial variation within levels, which is also to be expected when comparing patents at random, even when they are drawn from the same CPC subclass. This accentuates one of the advantages of using the text over a metadata-based approach such as using classification to infer content. Classification data

---

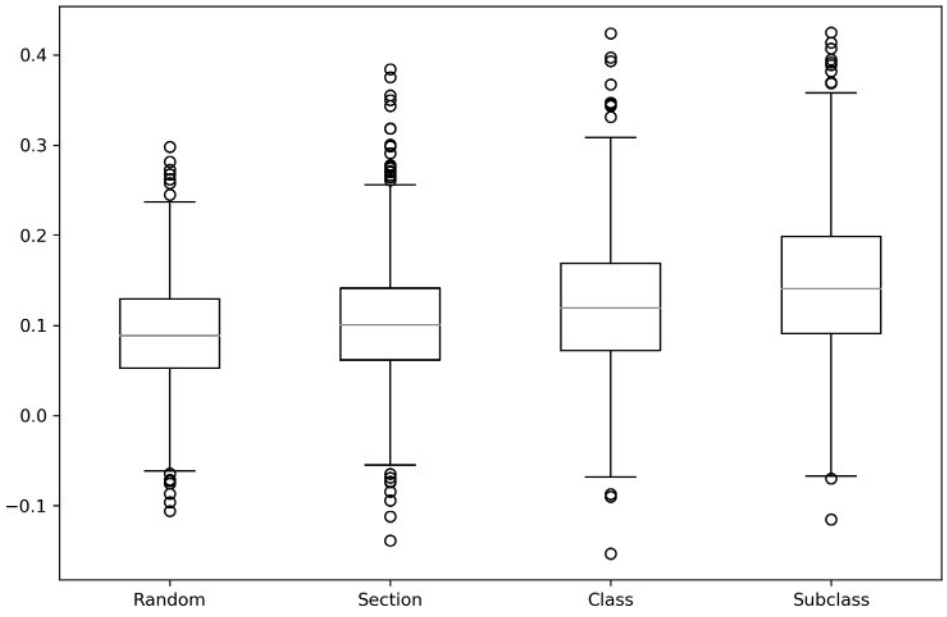
[50]See Cooperative Patent Classification, http://www.cooperativepatentclassification.org.

[51]See, e.g., Lee Fleming, Recombinant Uncertainty in Technological Search, 47 Mgmt. Sci. 117 (2001); Shi et al., supra note 25.

[52]T tests for mean equivalence show that each of these differences is statistically significant at the $p < 0.0001$ level.

[53]For example, consider a patent classified into subclass A45B. It fits within section "A," which broadly covers "Human Necessities," class "A45," which covers "Hand or Travelling Articles," and subclass "A45B," which covers "Walking sticks; umbrellas; ladies' or like fans."

Figure 1: Patent similarity by classification level.NOTE: Showing similarity distributions for 1,000 patent pairs matched at random, or by CPC section, class, or subclass.



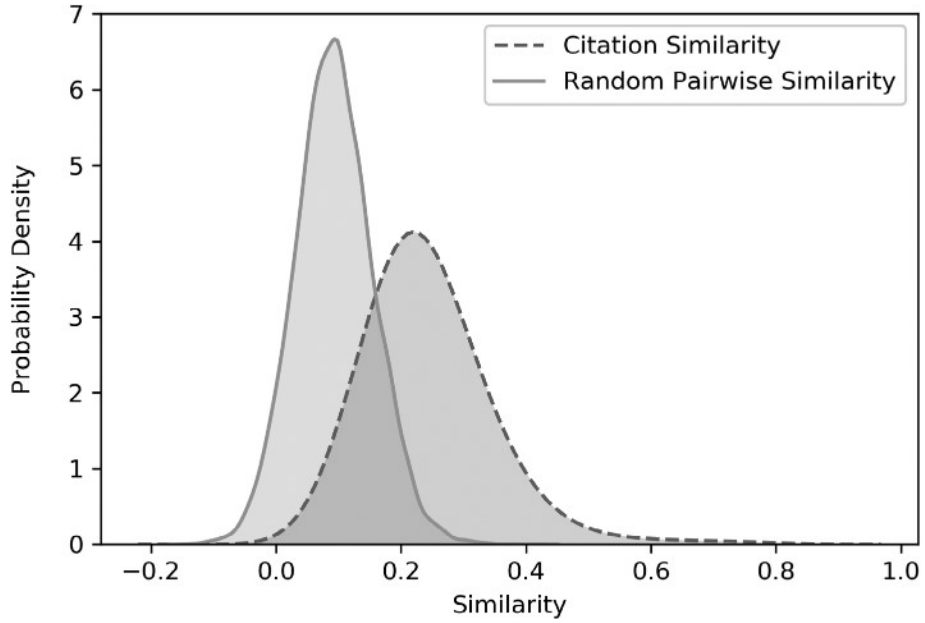
elide all the intra-category differences that exist, whereas a full text approach can capture more nuanced relationships.

B. Citation-Based Metrics

There is a large body of research using patent citations as components for a variety of measurements. These include measures of invention impact or value, innovation inputs, and knowledge flows. However, as discussed above, prior art citations are traditionally treated as binary constructs—they either exist or they do not. In reality, citations vary along a wide range of dimensions—one of which is the degree to which the cited and citing document are similar to one another. For instance, when considering citations as evidence of an invention's impact, the similarity of the citing patent provides insight into the type of impact the invention had. Citations from proximate inventions suggest that the invention had influence on the development of its own technical area, whereas citations from semantically dissimilar inventions suggest it had more wide-ranging impact. The Patent Similarity Dataset allows one to use these similarity variations to develop new and useful metrics.

Simply comparing the distribution of similarities between citing/cited patents and random patent pairs demonstrates that, as one would expect, prior art citations tend to

Figure 2: Similarity between citing patent pairs and random patent pairs.



be to other patents that are relatively similar to the citing patent. The average similarity between non-citing patent pairs is a relatively low 0.09 with a relatively tight distribution. Citing pairs on the other hand tend to be more similar to one another at 0.26 (see Figure 2).

We can also see that over time, the average citation similarity has decreased. Figure 3 plots the average backward citation similarity—that is, the similarity between a citing patent granted in year X and its cited prior art—showing a trend toward citing more and more dissimilar prior art over time.

There are many ways citation similarity data can be leveraged to provide insight into knowledge inputs and invention impact. Figure 4 describes four such measures in visual terms, with the central patent document representing the focal patent, the one-directional arrows showing backward and forward citation relationships, and the bidirectional arrows representing each of four proposed measures: (1) prior art proximity; (2) prior art homogeneity; (3) impact proximity; and (4) impact homogeneity. Each of these is described below and operationalized in the accompanying data analysis scripts. To demonstrate the measures, we sample 10,000 random patents from each year. The sample for backward-citation-based metrics begins in 1986 and extends to 2019, while the forward-citation-based sample extends from 1976 to 2009. To preserve comparability across patents invented at different times, we limit forward citation analyses to citations occurring within 10 years of grant.

Figure 3: Mean citation similarity over time.NOTE: Showing the average pairwise similarity across all prior art citations, averaged by year.

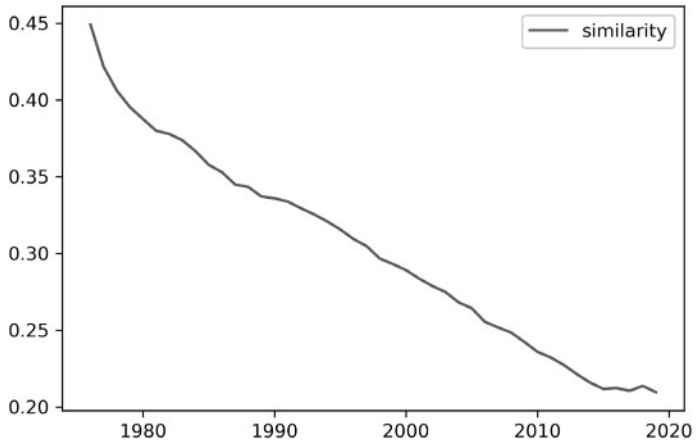
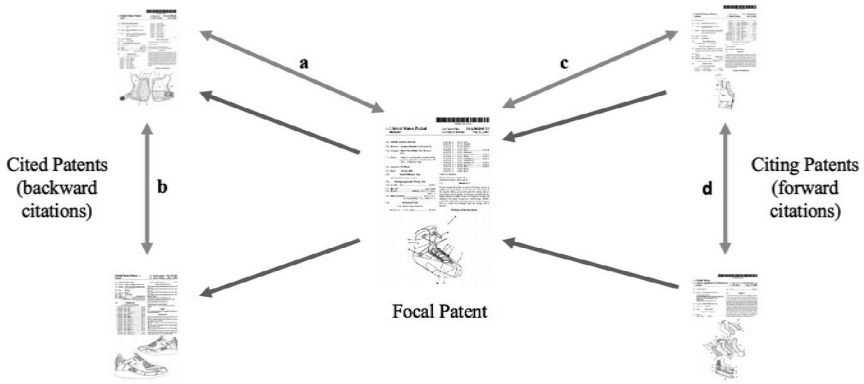


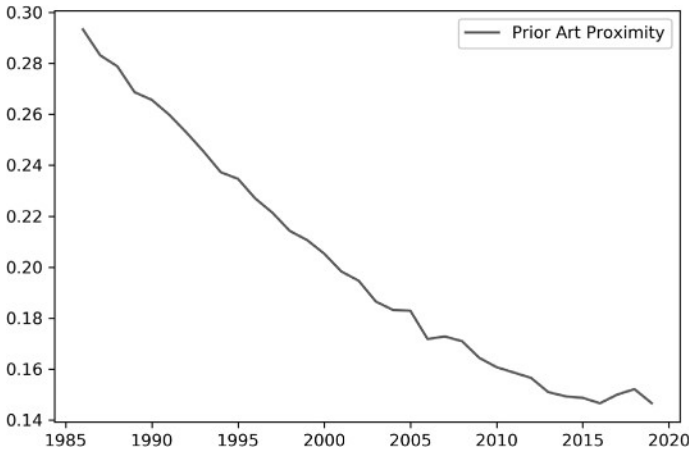
Figure 4: Citation based similarity measures.NOTE: a = prior art proximity; b = prior art homogeneity; c = impact proximity; d = impact homogeneity. Unidirectional arrows represent citations, bidirectional arrows represent the relationships upon which each metric is based.



Prior Art Proximity

Prior art proximity measures the degree to which a patent cites prior art that is similar or dissimilar to itself. To do so it measures the degree of similarity between the citing patent and its backward cited references. To calculate a patent's prior art proximity score, we

Figure 5: Prior art proximity.NOTE: Showing the yearly average minimum similarity between patents and their cited prior art (backward citations). The downward sloping curve shows that patents tend to cite to inventions that are increasingly distant from themselves.



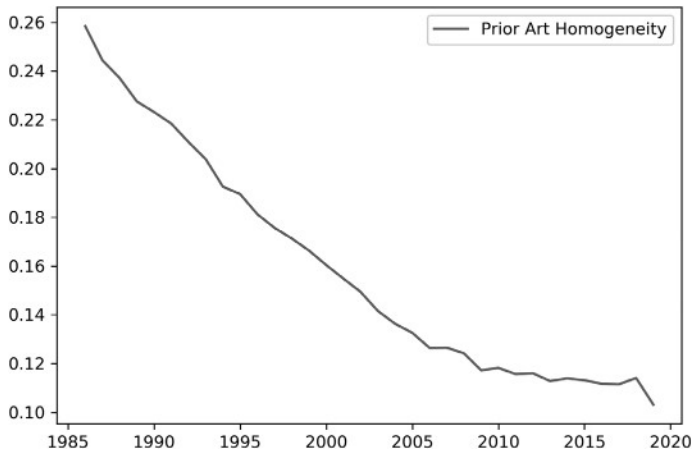
first measure the cosine similarity between it and each of its cited references. We then take the minimum of this set of scores.[54] Defined in this manner, a high knowledge proximity score means that a patent cites predominantly proximate and similar knowledge. On the other hand, a low knowledge translation score occurs when a patent draws on at least one piece of very dissimilar of knowledge. Figure 5 demonstrates this measure on a random sample of patents, showing a general decrease in knowledge proximity over time, but a leveling out in recent years.

Prior Art Homogeneity

Prior art homogeneity measures the degree to which a patent cites to areas of knowledge space that are distant from one another. Here, the comparison is between the co-cited prior art rather than between the focal invention and its cited prior art. To calculate prior art homogeneity, we first measure the pairwise cosine similarity between each co-cited pair of patents. The minimum of these scores represents the greatest degree of dissimilarity between knowledge areas that are cited by the focal patent. Inventions that bring together highly dissimilar areas of knowledge can be thought of as "boundary spanning" inventions that are qualitatively different than other types of inventions. Research

[54]Here we operationalize these citation-based metrics by using the minimum similarity. Depending on the substantive questions of interest, one might prefer to use different statistics (e.g., mean, median). We therefore enable this in the accompanying code.

Figure 6: Prior art homogeneity.NOTE: Showing the yearly average minimum similarity between co-cited prior art. The downward sloping curve shows that patents tend to cite to multiple different technological areas that are increasingly diverse from one another.



suggests that these types of inventions are more challenging for the PTO at the examination stage.[55] Figure 6 shows a general trend toward a decrease in the similarity between the various prior art cited by individual patents.

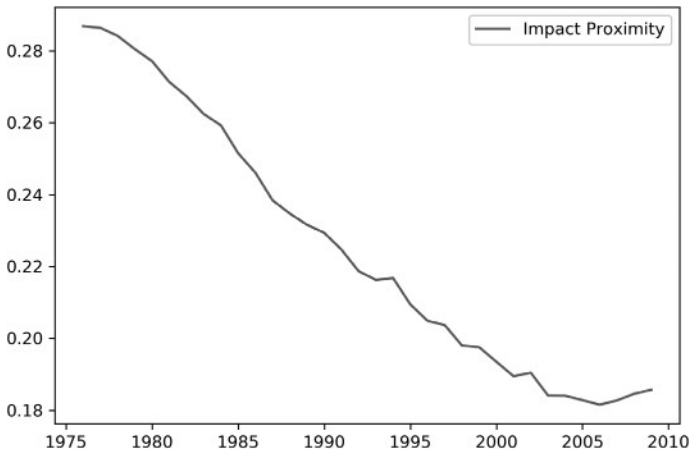
Impact Proximity

Impact proximity measures the degree to which a patent is cited by future patents that are similar or dissimilar to itself. A patent with a high impact proximity score has been cited as prior art by later patents that are similar to it—that is, its impact has been on proximate areas of the knowledge space—whereas a low impact proximity score demonstrates that the patent was cited by later patents that feature very different content than the original. To calculate impact proximity, we first calculate the pairwise similarity between a cited patent and all the citing references that are granted within 10 years of the cited patent. We then take the minimum of this set of scores, which reflects the single most dissimilar technical area by which the invention is cited. Figure 7 shows a general decrease in impact proximity over time, suggesting that patents have been cited by increasingly dissimilar prior art.

---

[55]Whalen, supra note 30.

Figure 7: Impact proximity.NOTE: Showing the yearly average minimum similarity between patents and their citing prior art (forward citations). The downward sloping curve shows that patents tend to be cited by increasingly distant technology fields.



Impact Homogeneity

Impact homogeneity measures the degree to which a patent is related to a diverse set of future patents through its forward citations. In the context of patents, a published patent with a high impact homogeneity score has been cited as prior art by inventions in a relatively narrowly localized area of the knowledge space, whereas a patent with a low impact homogeneity score has been cited by patents in diverse areas of the knowledge space. To calculate impact homogeneity, we first calculate the pairwise similarity scores for all co-citing references. The minimum among these scores measures the greatest dissimilarity between citing references and is taken as the homogeneity score, although the mean can be informative as well. Figure 8 shows a decrease in impact homogeneity, suggesting that, over time, patents have been cited by pairs of subsequent art that are increasingly dissimilar from one another.

The four similarity-weighted citation metrics proposed above can provide insight into patents and innovation that traditional binary citation measures are unable to capture. They show that patents are increasingly citing to less and less similar prior art (prior art proximity), that the prior art cited is itself increasingly diverse (prior art homogeneity), that inventions are progressively influencing fields that are more different from their own (impact proximity), and that they are also influencing fields that are more diverse (impact homogeneity). Some of this may be influenced by the overall increasing diversity of technologies. As more inventions are claimed, the "technology space" increases in size. All else equal, this leaves any two patents granted today less similar to one another than two patents granted 20 years ago (see Figure 9). Indeed, this can be measured and

Figure 8: Impact homogeneity.NOTE: Showing the patent-wise average minimum similarity between co-citing prior art. The downward sloping curve shows that patents tend to be cited by multiple different technological areas that are increasingly diverse from one another.

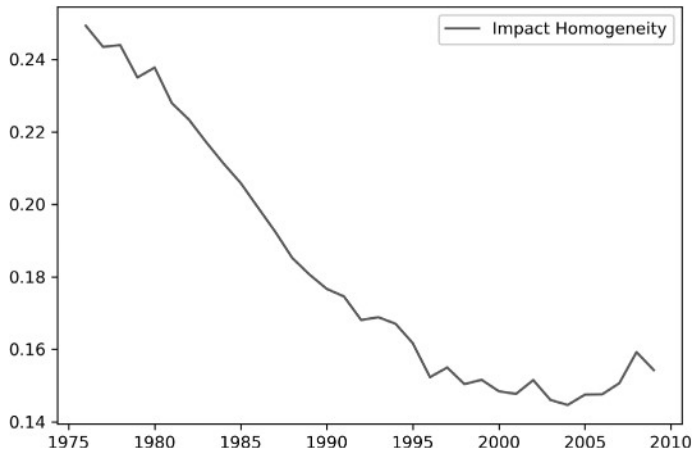


Figure 9: Increasing invention diversity.NOTE: Showing the average pairwise similarity for 10,000 random pairs of patents granted in the same year. The decrease in similarity reflects the increasing diversity of patented inventions over time.

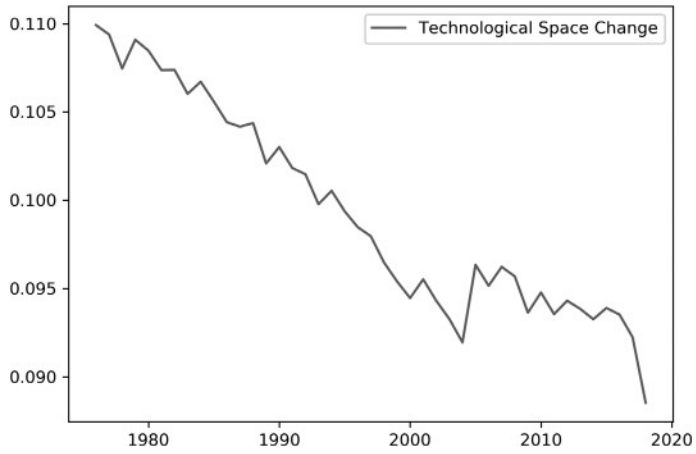
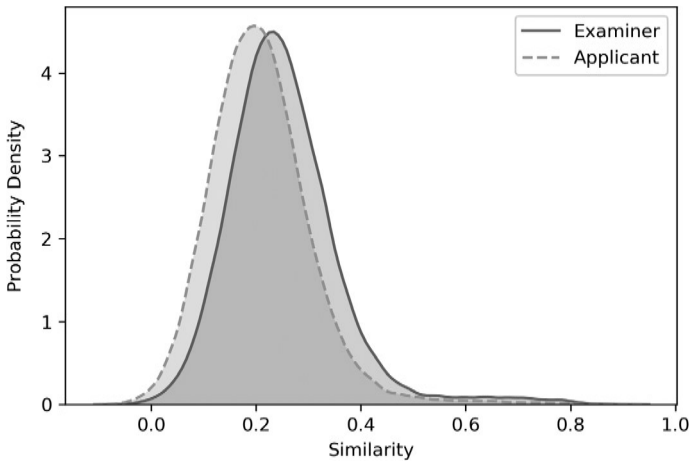


Figure 10: Examiner versus application citations.NOTE: Compares the similarity distribution for 50,000 randomly chosen examiner citations versus 50,000 randomly chosen applicant citations. The difference in means is statistically significant ($T = 56.87$, $p < 0.0001$).



controlled for.[56] Depending on one's substantive question of interest, controlling for changes in the size of the "technology space" may make sense. That said, the change over time is relatively small in scale compared to the changes we see in the above citation-based measures.

In addition to this, one might also be interested in using citation similarity data to provide insight into the patent application and examination process. Previous research has suggested that examiners and applicants may focus on different types of prior art when adding citations to patent applications,[57] and that examiners may find the prior art that they identify more useful in determining patentability.[58] Patent similarity data can provide new perspective on these issues. One simple way to do so is simply to compare the similarity distributions for citations added by applicants and examiners. Doing so reveals that examiners tend to cite to more-similar prior art (see Figure 10). This could perhaps be because examiners are better at finding prior art or, alternately, because applicants strategically exclude citations to more similar inventions. Regardless of why

[56]See the accompanying code notebook for a demonstration of how to control for this.

[57]Juan Alcácer, Michelle Gittelman & Bhaven Sampat, Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis, 38 Res. Pol'y 415 (2009).

[58]Christopher A. Cotropia, Mark A. Lemley & Bhaven Sampat, Do Applicant Patent Citations Matter? 44 Res. Pol'y 844 (2013).

these citation tendencies may differ, the Patent Similarity Dataset makes identifying and measuring them a straightforward process.

C. Patent Neighbors

The Patent Similarity Dataset also includes data on each patent's 100 nearest neighbors—the 100 patents from the dataset that are most similar to the focal patent—and their accompanying similarity scores. These data can be used for a wide variety of analyses, including those that provide perspective on how crowded an invention's "neighborhood" is.

As an example, consider the neighborhoods of both litigated and non-litigated patents. To examine whether they differ from one another, we begin with the litigated patent data,[59] and identify the similarity between each litigated patent and its nearest neighbor. We then compare these similarity scores with the similarity between non-litigated patents and their nearest neighbors. Having a very similar nearest neighbor suggests that the patent in question is in a more "crowded" intellectual property space, with perhaps many other competing, blocking, or related patents, whereas having only more distant neighbors suggests that an invention is relatively unique. By comparing the distributions of the nearest neighbor similarities for both litigated and non-litigated patents, we can see that, on average, litigated patents tend to have much more similar nearest neighbors than their non-litigated counterparts, and a somewhat bimodal distribution of these scores (see Figure 11).

The patent similarity data demonstrations provided thus far have focused on patent citations and nearest neighbors, and largely on individual patents. However, patent similarity data can also provide new forms of insight when one shifts focus from individual inventions to things like inventors, teams, and firms.

D. Inventor-, Team-, and Location-Level Analyses

At the inventor level, patent similarity data can be used to better understand a given inventor's area of expertise. This can be done by first situating each of the inventor's inventions within semantic space, and then calculating their pairwise similarity scores. Those scores can be used to create an "expertise network" that graphs the inventor's inventions and the similarities between them. These expertise networks provide insight into the type of innovator an inventor is. An inventor with a tightly grouped body of patents has historically invented in a relatively focused area of the knowledge space, whereas an inventor with significant distance between his or her inventions has worked in a more diverse set of areas.

To demonstrate, compare the invention networks of four well-known technology company CEOs (see Figure 12). We can see that Bill Gates's inventions are on average

---

[59]David L. Schwartz, Ted M. Sichelman & Richard Miller, USPTO Patent Number and Case Code File Dataset Documentation, SSRN Scholarly Paper ID 3507607 (Social Science Research Network Dec. 1, 2019).

Figure 11: Nearest similarity—litigated versus non-litigated patents.NOTE: Showing the similarity distribution for the nearest neighbor similarity of patents that go on to be litigated and those for random patents.

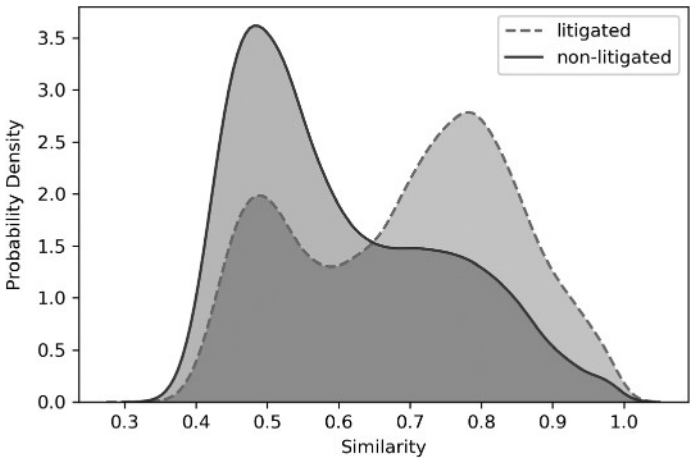
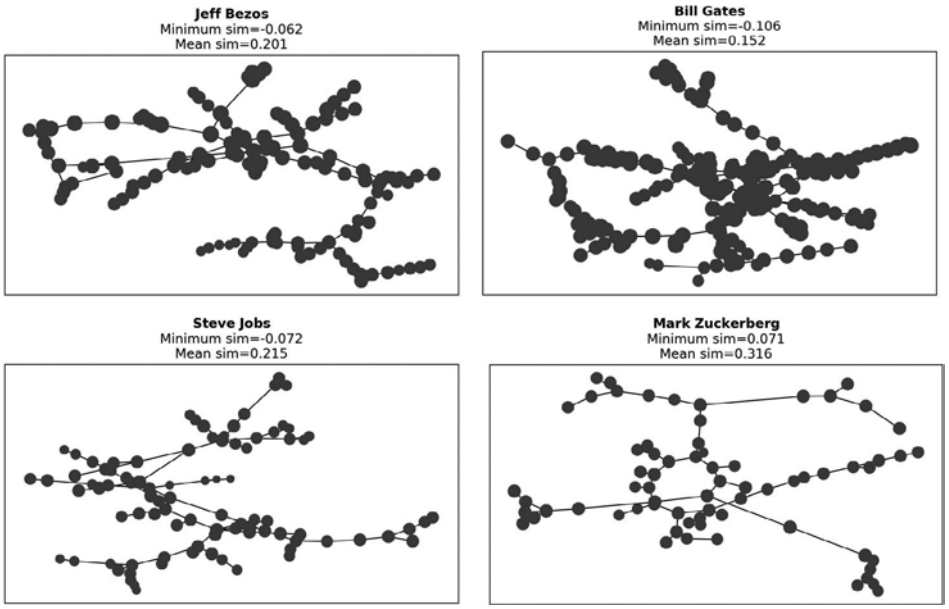


Figure 12: Intra-inventor-similarity networks.NOTE: Nodes are patents on which the individual is listed as an inventor, they are linked together based on their pairwise semantic similarity, and then the minimum spanning tree of the network is shown.



less similar to one another than Mark Zuckerberg's. Furthermore, he has a lower minimum similarity—suggesting that his two least similar inventions are less closely related than Zuckerberg's. All this suggests that, at least according to the patent record, Bill Gates has invented in a wider variety of areas than Mark Zuckerberg. Jobs and Bezos are somewhere between the two, with average similarities higher than Gates but lower than Zuckerberg, and similarly low minimum similarities.

Focusing on the inventor level can also provide new perspective on teams of inventors. An inventor's core area of expertise can be estimated by identifying the centroid of his or her invention network. To do so, we take the inventor's mean patent vector and locate it within the patent vector space. This location can be thought of as an inventor's "average" invention, estimating the core of his or her expertise. Inventors can then be compared to one another, to reveal whether or not they have historically tended to work in similar or dissimilar technical areas. Teams consisting of members with similar expertise backgrounds will have high similarity scores between their member centroids, whereas teams with more diverse inventing backgrounds will have low similarity scores between their member centroids. We can visualize these results in a number of ways. For instance, again using the tech CEOs assessed above, we can compare pairwise similarity between inventors and visualize the resulting team network (see Figure 13).

Figure 13: Inter-inventor similarity network.

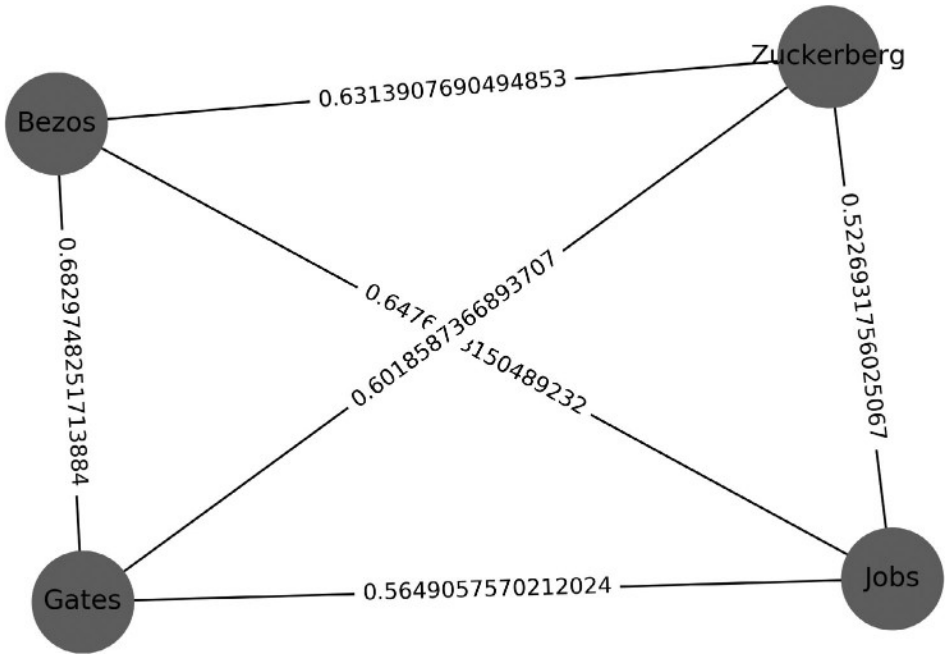
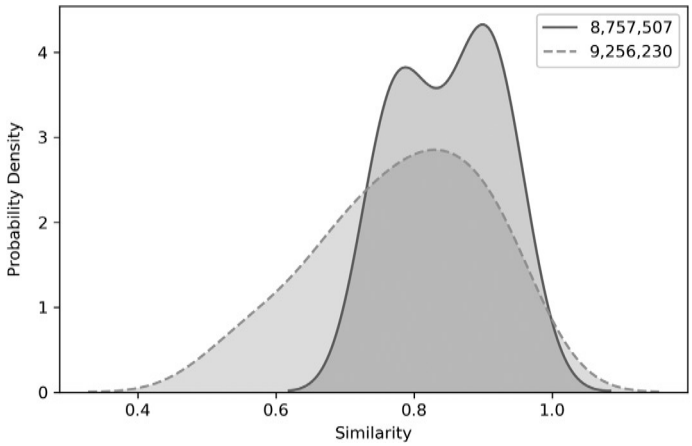


Figure 14: Patent team similarity distributions.NOTE: Showing the intra-team similarity scores for inventors on two patents.



Alternately, one might be interested in the average or distribution of pairwise similarity scores among team members. This, too, can be calculated with relative ease from the Patent Similarity Dataset. To demonstrate, Figure 14 compares the team member similarity scores for the teams of inventors listed on two of Google's Nest thermostat-related patents. We can see that the 8,757,507 patent was invented by individuals who had on average more similar inventing histories than those who invented the 9,256,230 patent.[60]

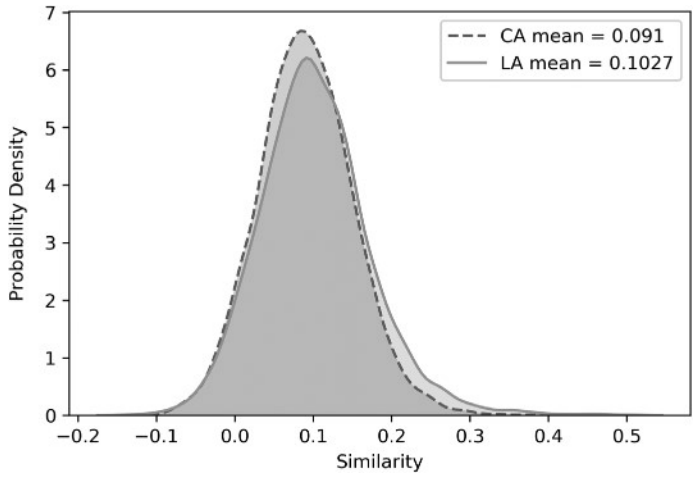
Many team-level metrics that the Patent Similarity Dataset enables could be similarly used on other groupings of inventors or inventions. For instance, one might be interested in the patent portfolios of firms or the similarities of inventors who work at those firms. Alternately, one might be curious about particular geographic locations, such as cities or states, and their inventing histories. Because it is easy to integrate with existing patent datasets, these sorts of firm- or location-level analyses are also relatively straightforward to implement using the Patent Similarity Dataset. For instance, one can quite easily measure the average pairwise similarity of inventions granted in one state (say, California) and compare it against that of another (say, Louisiana) (see Figure 15). We can see that Californian patenting is somewhat more diverse, with lower intra-state similarity scores than Louisiana.

E. Other Applications

The above is meant to illustrate a few of the wide variety of ways that patent semantic similarity data can be applied to the study of innovation and intellectual property law.

[60]The average pairwise team member similarity on the 507 patent is 0.85, while that on the 230 patent is 0.78.

Figure 15: California patent similarity versus Louisiana patent similarity.NOTE: Comparing the similarity distribution for 10,000 randomly selected patent pairs with inventors' addresses listing California with 10,000 randomly selected Louisianan inventions. The difference in means is statistically significant ($T = 14.61$, $p < 0.0001$).



In addition to these demonstrated measures, there are many more ways researchers can apply these similarity scores. For instance, patent similarity data have the potential to provide insight into legal disputes by giving additional perspective on litigated patents. Alternately, the PTO's patent examination process is replete with questions that may benefit from the potential insight that patent similarity data provide. One of this article's coauthors has argued elsewhere that semantic similarity data may provide valuable insight into developing empirical patentability metrics.[61]

Beyond providing insight into the administration of the patent system and enforcement of patent rights, patent similarity data can also provide additional perspectives on technological development more generally. Indeed, patent similarity is already used as a component of some patent landscaping techniques.[62] Other research suggests that semantic similarity data can have utility in a wide range of applications, including the identification of patent thickets[63] or estimation of a patent's value at the time of grant.[64]

---

[61]Laura G. Pedraza-Fariña & Ryan Whalen, A Network Theory of Patentability, 87 Univ. Chi. L. Rev. (2020).

[62]Aaron Abood & Dave Feltenberger, Automated Patent Landscaping," 26 Artificial Intell. L. 103 (2018).

[63]Gątkowski et al., supra note 36.

[64]Ashtor, supra note 37.

A publicly available patent similarity dataset and accompanying code makes implementing or improving on these existing techniques easier for researchers.

F. Obtaining the Patent Similarity Dataset

In publishing the Patent Similarity Dataset, we hope to facilitate its use by not only sharing and describing it, but by providing sample code that can be repurposed by researchers. By developing and operationalizing similarity-based metrics and providing the required code we hope to reduce the barrier to entry that many researchers face in applying natural language processing techniques to their own research. Thus, in addition to the publicly available data and description, we have provided an accompanying Python Jupyter notebook appendix that demonstrates how to use patent similarity data, and how to join them with other existing patent databases.[65]

- The patent similarity data includes the following files:[66]
- Patent vectors—this contains the 300-dimension vectors for each patent in JSON format.
- Citation similarity—this contains the cosine distance between all citing/cited pairs in the patent dataset. It is provided as a weighted edge list.
- Most similar—this contains the patent numbers and similarity scores for the 100 most similar patents to each granted utility patent in JSON format.

In addition to these files, we have also shared Python scripts that will download the public patent data provided by the COE and convert them into a SQLite database, as well as scripts that will add the patent similarity data as tables to that database.[67] Finally, we also share the saved Doc2Vec model, which can be used to calculate similarities for other patent pairs or arbitrary input texts, as well as scripts that can be used to re-compute the Doc2Vec model locally should users wish to alter the model parameters.

IV. Conclusion

This article has introduced the Patent Similarity Dataset, described its creation and structure, and demonstrated a variety of ways it can be used to produce novel insight of use to intellectual property and innovation researchers. It is our hope that by providing these data and related code we will make it more feasible for scholars to leverage advances in natural language processing in their own research. Combining these patent similarity

[65]The notebook can be retrieved from https://github.com/ryanwhalen/patent_similarity_data.

[66]The data files are available on the Zenodo data repository: https://zenodo.org/record/3552078 (DOI: 10.5281/zenodo.3552078).

[67]The patent download script can be found at https://github.com/ryanwhalen/patentsview_data_download.

data with other sources of patent data creates a powerful integrated database that enables researchers to move beyond metadata-based patent research and engage more deeply with patent content and the complex relationships between inventions (Table 1).

APPENDIX A: FILE DESCRIPTIONS

Table 1: Dataset files and descriptions

Filename	*Structure*	*Description*
vectors.json	Two columns: 1. Patent number 2. JSON list containing 300-dimension document embedding	This file contains one row for each patent. Each row contains the patent_id and that patent's doc2vec vector. The approximate uncompressed size is 39 GB.
most_sim.json	Two columns: 1. Patent number 2. JSON lists containing 100 tuples, each with (patent_id, similarity_score) structure	This file contains one row for each patent. Each row contains the patent_id and a list of tuples. Each tuple represents one of that patent's 100 nearest neighbors and the similarity score between that neighbor and the focal patent. The approximate uncompressed file size is 21 GB.
cite_sims.csv	Three columns: 1. Citing patent number 2. Cited patent number 3. Similarity score	This file contains one row for each prior art reference. Each row shows the citing patent, the cited patent, and the pairwise similarity between the two. The approximate uncompressed file size is 3 GB.
patent_doc2v_model.model	Binary	This file contains the genism model object, which can be used to embed documents not used in the training set, for instance, patents granted after the end of 2019. The uncompressed size is approximately 500 MB.

Note: Data available at https://zenodo.org/record/3552078.