NBER WORKING PAPER SERIES

MAPPING THE KNOWLEDGE SPACE:
EXPLOITING UNASSISTED MACHINE LEARNING TOOLS

Florenta Teodoridis
Jino Lu
Jeffrey L. Furman

Working Paper 30603
http://www.nber.org/papers/w30603

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2022

We are thankful for comments to Iain Cockburn, Joshua Krieger, Tim Simcoe, Martin Watzinger, and Rosemarie Ziedonis. Jeff Furman and Florenta gratefully acknowledge financial support from NSF SciSIP grants, SES-1564368 and SES-1564381, respectively. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Florenta Teodoridis, Jino Lu, and Jeffrey L. Furman. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Mapping the Knowledge Space: Exploiting Unassisted Machine Learning Tools
Florenta Teodoridis, Jino Lu, and Jeffrey L. Furman
NBER Working Paper No. 30603
October 2022
JEL No. C55,C80,O3,O31,O32

ABSTRACT

Understanding factors affecting the direction of innovation is a central aim of research in the economics of innovation. Progress on this topic has been inhibited by difficulties in measuring distance and movement in knowledge space. We describe a methodology that infers the mapping of the knowledge landscape based on text documents. The approach is based on an unassisted machine learning technique, Hierarchical Dirichlet Process (HDP), which flexibly identifies patterns in text corpora. The resulting mapping of the knowledge landscape enables calculations of distance and movement, measures that are valuable in several contexts for research in innovation. We benchmark and demonstrate the benefits of this approach in the context of 44 years of USPTO data.

Florenta Teodoridis
University of Southern California
Marshall School of Business
701 Exposition Blvd, HOH 503
Los Angeles, CA 90089
florenta.teodoridis@marshall.usc.edu

Jino Lu
University of Southern California
jinhong.lu@marshall.usc.edu

Jeffrey L. Furman
Boston University
Questrom School of Business
595 Commonwealth Ave - #653a
Boston, MA 02215
and NBER
furman@bu.edu

Code and data are available at https://github.com/jinolu/Direction_of_Innovation-HDP

I. Introduction

In 2011, Peter Thiel, the high-profile founder of PayPal and Palantir, and subsequent technology investor, famously published an investment manifesto with the subtitle, "We wanted flying cars, instead we got 140 characters." This provocative remark on contemporary technological evolution highlights a first order question regarding why innovation takes the direction that it does. However, since the NBER's seminal publication of the volume, *The Rate and Direction of Inventive Activity* in 1962 (NBER, 1962; Lerner and Stern, 2012), research has made more progress studying drivers of the inventive activity than its direction (Acemoglu, 2012; Myers, 2020).

The difficulty of developing empirical measures that characterize the relationships among ideas in knowledge space is among the key reasons for the relative paucity of such studies. In this paper, we adapt a natural language processing (NLP) technique to develop a methodology that infers the mapping of the knowledge landscape[1] – knowledge areas and the relationships between them – as captured in researchers' text documents of interest. Scholars have long used text documents such as patents, academic publications, 10k statements and news articles as a paper trail of firms' knowledge (e.g., Machlup, 1961; Griliches, 1990). A method that maps the knowledge landscape reflected in such documents enables a more comprehensive measurement of *distance* and *movement* in knowledge space.

To date, research on factors that impact the distance and movement in knowledge space primarily relies on measures derived either from observable indicators of innovation output, such

[1] We use knowledge and innovations interchangeably. However, the concept of knowledge can be extended to include any information an economic actor has. In other words, the method we describe can be applied to map not only knowledge directly related to an economic actor's innovation actions and capabilities, but also any other knowledge captured in text documents.

as taxonomies, citations, and keywords, or from earlier-stage NLP techniques of pair-wise similarity, namely vector space models such as cosine similarity and Jaccard index, and topic modeling such as Latent Dirichlet Allocation (LDA). While undoubtedly helpful, these approaches face limitations in their ability to provide a comprehensive mapping of the knowledge landscape as captured in documents of interest. For example, curated taxonomies face the tradeoff of either being stable and, thus, consistent over time or changing to accommodate the evolution of the knowledge space at the cost of classification standardization over time. Taxonomies are also socially-constructed, which affects both their accuracy and generality (Thompson and Fox-Kean, 2005; Righi and Simcoe, 2019). Citations fare better as markers of knowledge space but they, too, are subject to social processes and institutional norms (e.g., Alcacer and Gittelman, 2006; Kuhn et al., 2020).[2] NLP-based similarity measures address several of these limitations (Arts et al., 2018, 2021; Krieger, Schnitzer, and Watzinger, 2022). These are most useful in comparing pairs or groups of knowledge sets (e.g., Furman and Stern, 2011; Vakili and Zhang, 2018;). As a result, they apply best to research questions involving comparisons rather than flexible mappings of a knowledge landscape.

More broadly, mapping the knowledge space is difficult because, unlike physical space, which consists of a well-known number of dimensions, clearly-measurable distances between locations, and which is relatively stable over time, the knowledge space consists of an unknown number of dimensions, the distances between which cannot be uniquely measured, and which evolve over time in ways that cannot be anticipated until they are realized.[3] Thus, the ability to

[2] For example, Abrams, Akcigit, and Grennan (2018) suggest that a higher volume of citations in patents captures a strategic behavior of the filer rather than accurate knowledge flows.

[3] Doran (2017) conceives of ideas space as composed of an arbitrary number of dimensions and argues that there is neither a unique way to measure the distance between vectors in this multidimensional space nor the rate at which any vector is changing its position relative to its prior position or to other vectors.

measure distance and movement in the knowledge space is contingent on having access to a map of the space that is always up to date, while, at the same time, facilitates ways to trace its evolution over time. Recent developments in machine learning have the potential to address this limitation because they provide a flexible method to infer the structure of a collection of text without making any ex-ante assumptions about it.

In this paper, we extend recent efforts exploiting machine learning, and specifically NLP techniques, to present and share an algorithm that analyzes the text of any documents of interest to reveal the structure of relationships connecting the knowledge captured in the documents. Specifically, we exploit the ability of Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006),[4] a topic modeling algorithm, to uncover the structure of a collection of text for the specific purpose of constructing a map of the knowledge captured in the text. The algorithm can be applied to any corpus of text, including patents, academic publications, 10K statements, business press articles, public relations statements, or any combination. The output of the algorithm can then be used to develop measures of distance and movement in knowledge space as covariates in empirical analysis. The measurement can be done at any level of analysis, such as individual, firm, industry or region, and it can trace changes over any reference attribute, such as changes over time, geographies or demographic attributes of economic actors.

In developing these techniques, we join others' efforts to explore and democratize access to frontier developments in machine learning (e.g., Athey, 2018; Kleinberg et al., 2018; Choudhury et al., 2019) and to data that can be applied to the study of several questions of interest to research on innovation (Png, 2019; Kelly et al., 2020; Marx and Fuegi, 2020). To our knowledge, our approach is the first open access effort to systematically develop a methodology to map the

[4] For a comprehensive review of a range of topic modeling algorithms, please see Boyd-Graber, Hu, and Minmo (2017).

knowledge space in a flexible manner. Similar efforts, based on similar techniques, have been undertaken by two large bibliographic databases: The National Library of Medicine (the PubMed Related Articles algorithm) and Microsoft (the Microsoft Academic Graph's categorization schema) (Sinha et al., 2015). In each of those cases, however, the mapping schema is focused on a temporal dimension and does not provide access to the data that describes the distance between knowledge areas. Further, and most importantly, the two algorithms apply uniquely to the sets of documents in their respective databases and hence cannot be applied to other text corpora.

II. Measuring the direction of innovation

II.1. Why is it important to be able to measure the direction of innovation?

Interest in the factors affecting the direction of innovation has been sustained in economics for more than a half century because of the central importance of knowledge for economic growth (Solow, 1956; Nelson, 1962; Romer, 1990; Lerner and Stern, 2012). For an equally long period, scholars have recognized the central role of heterogeneous knowledge in influencing competition, competitive advantage and the evolution of products, firms, and industries (Penrose, 1959; Cohen and Levinthal, 1989; Klepper and Graddy, 1990; March, 1991; Klepper, 1996). Further, the difficulties associated with obtaining a socially optimal level of diversity of innovation have been long understood to arise from difficulties with appropriability and indivisibility, which yield externalities and fixed costs that negatively affect initial inventors relative to subsequent inventors and result in underinvestment in innovation (Arrow, 1962; Scotchmer, 1991).

Acemoglu (2012) demonstrates that market incentives, as currently crafted around intellectual property protection policies, yield underinvestment in research diversity relative to the social optimum. In turn, this leads to excessive concentration of innovation in the types of technologies that offer rents in the near term at the expense of alternative approaches that may yield higher returns in the long term. Bryan and Lemus (2017) underscore this point, noting that suboptimal investment in diversity arises even in the absence of intellectual property policy interventions. Together, this line of inquiry paints a somewhat pessimistic picture regarding the relationship between research diversity and economic growth. On the more optimistic side, Acemoglu (2012) suggests that the underprovision of commercial incentives for diversity in invention may be mitigated, though not fully, through diversity in basic science research, which serves as an input into commercial innovation and for which the incentives for diversity are not dampened by commercial imperatives. Indeed, Aghion, Dewatripont and Stein (2008) argue that "the private sector's ownership of a given idea will not yield as diverse an array of useful next-generation ideas as would be generated in academia" (p. 630), since firms' incentives for commercial success impose constraints on their researchers' choices while academia's commitment to freedom in project selection enables its researchers to choose across broad sets of scientifically feasible research lines. Recent research demonstrates that increasing diversity among innovators can affect the diversity of innovations (e.g., Koning, Samila, Ferguson, 2020).

Project selection in academia is not, however, entirely free of constraints. First, the incentive system in academia gives rise to numerous concerns relative to freedom of project selection, including reputation considerations, personal satisfaction, financial remuneration, and a number of other more eclectic considerations (Merton, 1957; Stephan, 1996). For example, scholars target their research for certain journals in response to monetary awards (Franzoni et al, 2011) and reputation benefits associated with such publications, in an attempt to secure grants as urged by universities (Stephan, 1996) or to meet tenure requirements. Second, competition in ideas space and the expansion of the knowledge frontier forces researchers to specialize in increasingly narrower niches (Jones, 2009; Agrawal, Goldfarb and Teodoridis, 2016), to move away from crowded research areas (Borjas and Doran, 2015), and to race towards more rushed, lower quality projects (Hill and Stein, 2022). Third, external factors such as science and innovation policies, firm and government investments, market demand-pull factors, and features of local research environments, including geographic endowments and firm policies, can also affect researchers' project choices (Dasgupta and David, 1994; Henderson and Cockburn, 1994, 1996). For example, massive research commitments, such as the Manhattan Project, the War on Cancer, and the Large Hadron Collider at CERN, accelerate the rate of knowledge generation, although their impact on research direction is less clear. Such projects may increase research diversity by bringing together scholars from different areas and enabling broad exploration to solve specific problems, but a mission-driven focus on achieving particular outcomes and solving specific questions may direct inquiry away from alternative research trajectories.

Moreover, and coming in full circle, firms' demand for certain innovation types influences upstream incentives to innovation diversity and changes in trajectory as researchers account for the likelihood of their ideas reaching the market (e.g., Aghion and Griffith, 2008) or being available for sale in the market (Arora and Gambardella, 2010). At the same time, several studies found that academic research choices influence the innovation activity and performance of firms (e.g., Jaffe, 1989; Cohen et al., 2002; Adams et al., 2003), either directly or through spillovers (e.g., Henderson et al., 2005).

All in all, the importance of factors influencing distance and movement in knowledge space spans a wide range from implications for individual-level innovation and firm performance to social welfare and economic growth.

II.2. Difficulties in measuring the direction of innovation

Empirical measurement of the direction of innovation is challenging because it requires capturing changes in an otherwise ever-evolving multidimensional idea space. In other words, in order to estimate changes in the direction of innovation one first needs to define boundaries of research trajectories. However, this creates a paradox as boundaries of research trajectories are part of the core unknown to be estimated.

As such, most efforts to capture the direction of innovation have been a blend of two conceptual approaches: a focus on capturing indexes of similarity between innovation efforts, such as academic papers and patents, and a focus on categorizing such items into topics in order to then measure changes in the topic landscape. The two approaches are akin but distinct. The latter is needed in order to capture the direction of innovation, while the former is a proxy that is limited in its ability to reach the desired goal of measuring the direction of innovation. Specifically, similarity can be used to compare pairs or groups of innovation outcomes but does not offer any information about the set of topics that holistically describes the ideas space at any point in time. Moreover, it is not sufficient to obtain a mapping of the ideas space at one point in time. Capturing the direction of innovation requires a methodology that ensures the map updates with the changing ideas space landscape, and that allows to track those changes over time.

For example, research that categorizes direction based on curated taxonomies have the advantage of consistency but face the tradeoff of either being stable and, thus, comparable over time or being dynamic and, thus, evolving with the changing research landscape at the cost of consistency and classification standardization. Author-assigned keywords, or any other set of keywords not drawn from a defined vocabulary, fare better than stable taxonomies in capturing new knowledge directions but lack structure and may be more subject to gaming. This is a concern as it limits interpretability of hierarchical connections between keywords and changes in such relationships over time. Similarly, while the citation revolution has provided a method of tracing knowledge linkages (DeSolla Price, 1970; Griliches, 1990) and has enabled significant contributions to understanding the rate and drivers of knowledge accretion, it is subject to the concerns we note above since measuring diversity via citation maps also requires some form of categorization.

Machine learning techniques have the potential to address these limitations by helping researchers uncover the latent structure of the knowledge space based on inferred similarity of large bodies of research texts, such as academic publications, conference proceedings and patents. We exploit advancements in NLP techniques to develop an approach for mapping the knowledge landscape as captured in researchers' text documents. In doing so, we build on and extend earlier NLP-based developments.

These earlier efforts exploited a specific attribute of NLP methods, that of enabling measures of pair-wise knowledge similarity. The earliest such approach was based on vector space models (Salton et al., 1975), namely algorithms that represent text as a vector of its words. Under this representation, pair-wise knowledge similarity, as captured in the text, can be computed using standard vector measures such as cosine similarity or Jaccard index. Scholars have applied vector space models to compute the similarity between various documents such as research papers, patents, and firms' 10-K statements (Magerman et al., 2015; Hoberg and Phillips, 2016; Younge and Kuhn, 2019;). For example, Krieger, Li and Papanikolaou (2019, 2022) calculate Jaccard indexes to compare the similarity of firms' drug development projects. Arts et al. (2018) applies this method to calculate a Jaccard index of similarity between any two USPTO patents or patent portfolios. However, despite its popularity and straightforwardness, the vector space approach faces several challenges[5] that prompted scholars to explore an alternative NLP technique – topic modeling – for calculating pair-wise knowledge similarity (e.g., Kaplan and Vakili, 2015; Huang et al., 2018; Hansen et al., 2018; Iaria et al., 2018).

We extend the use of topic modeling for capturing pair-wise similarity to mapping the knowledge landscape. Specifically, our approach is targeted at exploiting the ability of topic modeling algorithms to uncover the latent topic structure of a corpus of text. A few other studies have engaged with this feature of topic modeling (e.g., Kaplan and Vakili, 2015; Hansen et al., 2018; Ghose et al., 2019; Furman and Teodoridis, 2020). Our approach extends and generalizes these efforts to a flexible implementation of topic modeling that can be employed to map the knowledge landscape as captured in any text documents. This enables developing measures of distance and movement – a richer set of attributes than pair-wise similarity.

III. A topic modeling approach for mapping the knowledge landscape

III.1. A topic modeling primer

Topic modeling algorithms are unassisted machine learning techniques that focus on uncovering the structure underlying an otherwise unstructured collection of documents. More specifically, they are probabilistic models that employ a hierarchical Bayesian analysis of text (see e.g., Hofmann, 1999; Blei et al., 2003; Buntine and Jakulin, 2004) to analyze the words of documents in order to uncover latent topics and organize documents into categories. The algorithms are unassisted in the sense that they do not rely on *a priori* human labeling of documents,

[5] For example, topic modeling solves the curse-of-dimensionality problem of vector space models (Deerwester et al., 1990; Lei et al., 2019; Blei et al., 2003; Shahmirzadi et al., 2019) – longer text translates into difficult to operationalize high dimension vectors – by projecting the high dimensional terms into a probability distribution of topics (Blei et al., 2003; Teh et al., 2006). In other words, instead of attempting to compare bodies of text among themselves, topic modeling operates on the premise that there exists a latent structure that describes the text and attempts to uncover that structure.

however sparse, but rather construct topics based on analysis of original text. The intuition is that of a generative process, where the data are assumed to be characterized by a set of observed variables (words in the document) that developed from a set of hidden variables (the topic structure). Thus, the concept of "topic" in topic modeling is a mathematical construct. Specifically, topics capture words that appear together in the text corpora with a certain probability;[6] they are not labeled, and they do not necessarily capture categories in the same way humans would envision a categorization structure. In their comprehensive review of topic modeling techniques, Boyd-Graber, Hu, and Minmo (2017, pg. 8) explain, "For example, a [human-defined topic] might include an earthquake in Haiti, whereas a topic model might represent the same event as a combination of topics such as Haiti, natural disasters, and international aid." Thus, a benefit of topic modeling for our purposes is that it captures all patterns in the data, even those that might not be *ex ante* on researchers' mind. Stated differently, topic modeling generates a comprehensive mapping of relationships between ideas in a text corpus.

Because the set of words comprising a topic represent a latent characterization of the words in the text, the topics are influenced by the words in the text corpus. For example, if the text corpus is not cleaned of words, such as "and", "or", "in", then topics including these words would be generated. This is not wrong in a technical sense, since indeed documents would have these words in common, but it is likely uninformative for management research. There are several such vocabulary pre-processing steps that users of topic modeling algorithms typically consider. The most common include eliminating conjunctions and prepositions, stemming (reducing words to their stem root), lemmatization (reducing words to their dictionary form), n-grams (deciding if the

[6] The algorithms exploit patterns of word co-occurrence to infer the set of "hidden" topics, under the assumption that words that frequently occur together are more likely to be semantically correlated (Blei et al., 2003; Hoffmann, 1999; Barde and Bainwad, 2017). This helps address, to a certain extent, issues of synonymity and polysemy.

meaning of n-group of words should be considered together or as individual words), and embeddings (learning the meaning of words and groups of words from a similar context).

In all cases, the main goal of vocabulary pre-processing steps is to ensure the generated topics are informative. This is most important if users are generating topics with the aim of interpreting the topics afterwards. This is a secondary concern for this paper as we are principally interested in capturing distance and movement in the knowledge landscape. In other words, in our case, topics are useful in that they provide a mapping of the knowledge space that can inform on distance and movement. The more detailed the map i.e., the more patterns (topics) revealed in the data, the more accurate the measurement of distance and movement, even if some topics might be deemed less informative than others purely from a categorization perspective. A parallel to geographical maps is informative: (1) the more detailed a map, the better suited to accurately calculate distance and movement in geographical space, and (2) the choice of map type (i.e., "topic" type), such as topological, physical, or a combination, does not affect the accuracy of the measures even if some "topics" might be viewed as less informative than others from a categorization perspective. For example, distance is measured equally accurately in a topological map, a road map or a physical map, even if some would consider a certain map type more intuitive for measuring distance.

A final important characteristic is the number of topics in topic modeling. The number of topics generated by a topic modeling algorithm is not equivalent to the number of categories describing the text corpus, i.e., the number of distinct bins of document types. If a topic modeling algorithm classifies a corpus of text in twenty topics, this does not mean that each document will belong uniquely to one of the twenty topics. Instead, documents can be classified by the topic modeling algorithm as belonging to multiple topics with each such membership being characterized by a certain probability. The probability can be viewed as the weight of belonging to a topic. Thus, the number of distinct bins that characterizes the text corpus is equivalent to all realized combinations of weighted topic modeling topics. The theoretical lower bound for this number is N combinations of i, where i ranges from 1 to N, and N is the number of topics identified by the topic modeling algorithm ($\sum_{i=1}^{N}\frac{N!}{i!(N-i)!}$). For example, a topic modeling analysis that generates four topics (T1, T2, T3, T4), can have documents categorized not only in T1, T2, T3 and T4, but also in {T1, T2}, {T1, T3}, {T1, T4}, {T2, T3}, {T1, T2, T3}, {T1, T2, T4},{T1, T3, T4}, {T2, T3, T4} and {T1, T2, T3, T4}, to a total of 13 categories. Moreover, this is a lower theoretical bound because the assignment to these categories is weighted and the weights range from 1 to 100 (i.e., probabilities of 0.01 to 1). The realized number of categories is the subset of bins that have at least one document assigned to the bin. As a result, the total number of categories describing a corpus of text is not the number of topics identified by the algorithm, but rather a larger number, since topic membership is probabilistic and documents can belong to any combination of topics.

III.2. An approach to mapping the knowledge space

A number of topic modeling algorithms are available, of which LDA is the most popular among social scientists (Blei et al., 2003). We depart from this popular choice and select a different topic modeling algorithm as the basis of our approach – HDP – which constitutes an extension of the LDA algorithm. We select HDP because of its built-in mechanism to identify the optimal number of topics (in a mathematical sense) that describe a corpus of text. LDA groups documents into the number of topics chosen by the researcher. Placing the choice of the number of topics into the hands of the researcher leads to some significant challenges for our set goal. Choosing a number of topics that is too large could result in category definitions that are overly narrow, while choosing a number that is too small could result in category definitions that are too broad.[7] This is a concern even if topic assignment is probabilistic. For example, if the optimal number of topics is too small, the weights that would otherwise be split between the optimal number of topics would be summed up thus obscuring a more granular categorization. Conversely, if the number of topics is too large, weights that would indicate assignment to distinct patterns would be further split creating the illusion of a more granular categorization. HDP, which can be viewed as a non-parametric extension of LDA, relaxes this constraint, as the algorithm identifies the optimal number of topics that characterize a corpus of text and then assigns documents probabilistically to these topics.

We modify the off-the-shelf HDP algorithm to output (1) the top 30 words describing each topic[8], i.e., those assigned the highest probability of describing each corpus of text, and (2) the set of topics describing each input text with a probability greater than 0.01. Because the main goal of topic modeling algorithms is prediction, the off-the-shelf version of HDP only considers these attributes internally and does not present them to the user. Our modification does not change the internal processing of the topic modeling – we rely on computer scientists' expertise – but rather adjusts the output of the algorithm to display information of interest to innovation researchers. The final output of our algorithm includes (1) a list of input text IDs, e.g., patent numbers, and the probabilities with which each text belongs to the HDP-generated topics (naturally, the probabilities

[7] In the original paper introducing the HDP algorithm, Teh et al. (2016) benchmark the HDP against the LDA. Nevertheless, before settling on our proposed approach, we executed several LDA analyses and compared the output with the HDP analysis. The generated categories were qualitatively similar. Furthermore, we compared the number of topics generated by the HDP with the output of algorithms that scientists use to guide their choice of optimal number of topics in an LDA analysis and found no discrepancies. Last, we recorded the run time between these different approaches and found the HDP to run in a fraction of the time when compared to the LDA.

[8] The algorithm can be easily adjusted to output more than the top 30 words. However, in our testing, we found that beyond 30 words the probability of the words describing the text takes very low values, i.e., the information that such words carry is marginal at best.

add up to 1, thus fully describing the text), and (2) the list of HDP-generated topics and the words that describe each topic, along with their respective assigned weights.

The algorithm can be used to analyze the entire body of text comprising the dataset of interest in one go, or separately for subsets of the dataset, split based on certain attributes of interest. More specifically, scholars might be interested in measuring movement and distance in knowledge space relative to certain reference attributes, such as changes over time, geographies or demographic attributes. However, all topic modeling algorithms treat the input text as a one-time group for which the latent categorization needs to be revealed. In other words, current algorithms cannot automatically track the evolution of topics over a certain attribute by updating the set of keywords in each category over time. Scholars have two options to make this possible.

One approach is to have the entire dataset evaluated by the topic modeling algorithm in a single run and then to have the measures of interest constructed relative to the reference attribute of interest. The advantage rests with the consistency of the resulting mapping of the knowledge space. The disadvantage is the tall computational requirement. While there is no conceptual limit to the amount of text that can be analyzed, the computational requirements, including the processing time, increase exponentially with the amount of data.[9]

A second, alternative approach for obtaining a mapping over a desired reference attribute is to split the data by the attribute, such as by year, or by state, and execute the algorithm for each data subset. This approach requires an additional step to connect the resulting sets of topics into a knowledge map. To make this possible, we add to our algorithm an optional step that calculates

[9] For example, the application to 44 years of patent abstracts that we include in this paper took an average of three weeks for each run of the algorithm using Cloud services. Considering that one might need to execute several instances of the algorithm to tune certain parameters, such as vocabulary settings, the run time to obtain the knowledge map can easily extend to months of data processing time. Moreover, we find it difficult to even begin to estimate the computational requirements for repeating the same process using the full text of the patents, given that the approach would increase the size of our dataset several times fold.

the cosine vector similarity between the topics generated by the HDP algorithm over each attribute of interest (e.g., yearly HDP topics). Specifically, we employ a weighted cosine similarity algorithm where the weights are the HDP-generated probabilities that capture the relevance of each word for each topic. This simple vector-space similarity technique works well because the words describing each topic do not suffer, by construction, from any of the issues that limit the feasibility of vector-space algorithms to produce an informative similarity measure (e.g., Deerwester et al., 1990; Lei et al., 2019; Blei et al., 2003; Shahmirzadi et al., 2019).

III.3. Benefits and limitations

There are four main benefits to our approach when compared with existing methods for capturing distance and movement in knowledge space. First, our technique captures changes in the knowledge landscape earlier than most other approaches of mapping the space. The benefit is a direct consequence of the logic underlying topic modeling algorithms, that of revealing all patterns characterizing a text corpus, including emergent ones. By contrast, many taxonomies are relatively stable over time and hence might not change frequently enough to reflect emerging ideas or innovation trajectories that are exhausted.

Second, our approach facilitates a more accurate measurement of distance in knowledge space. Fundamentally, measuring distance requires (1) understanding the degree of similarity between different knowledge areas and (2) the knowledge space positioning of economic actors. Existing techniques for developing similarity measures help with the former but are not directly suited to capture the latter. Taxonomies are relatively weak in inferring the former, but fare better in capturing the latter. The mapping of the knowledge space generated by our algorithm provides a method that is conducive to reaping the benefits of both worlds.

Third, mapping the knowledge landscape allows scholars to capture movement in the space. To characterize movement, researchers need access to a (1) detailed mapping of the knowledge space that is (2) most current and that (3) retains the evolution of the knowledge space to the most current version. First, the level of detail is important because it allows researchers to observe even small changes in movement. The coarser the classification of knowledge areas, the lower the ability to observe movement. Second, a map that is current would facilitate access to more precise changes. Classifications that are updated infrequently would not capture contemporaneous information about movement in knowledge space. Third, a classification system that does not retain its history of changes would lose at least some information about movement in the knowledge space. Taxonomies, citations, keywords and other similar approaches are less likely to meet all these criteria than the technique we describe.

In addition, a key advantage of the approach we describe is its flexibility. Using this approach, a researcher can choose the type and amount of data to analyze, the reference attributes and the unit of analysis. The algorithm can be applied to any text dataset the researcher deems informative for their topic of interest, including, for example, the abstracts or full text of patents or academic publications, firms' 10K statements, standard setting documents, public relations statements, court opinions, business press articles, firm product catalogs, industry conference brochures, etc. Moreover, there is no restriction in combining these datasets. For example, one might want to analyze the innovation portfolio of firms, comprised of academic papers and patents, or the strategy of firms captured in public relation statements and business press articles, or firms' attempts to secure intellectual property as captured in standard setting documents and patents. Other approaches for capturing firms' knowledge are usually restricted to one type of data. For example, taxonomies do not classify patents and academic papers together and citations between academic papers and patents have only recently been systematically documented and made public (e.g., Marx and Fuegi, 2020). Moreover, the technique we describe allows to trace changes over any reference attribute, such as changes over time, geographies or demographic attributes of economic actors, and at any level of analysis, such as individual, firm, industry or region.

In considering how to engage with our technique, it is also important to note that it is not superior to all other approaches of capturing knowledge attributes in all instances. In cases where fields are relatively well-defined, taxonomies can form a valuable basis for inquiry about distance and movement in knowledge space. For example, the internationally created and maintain US patent classification schema, the Cooperative Patent Classification (CPC), provides a detailed mapping of the knowledge space as captured in the text data of patents. Moreover, the advantages of our technique are contingent on the type of available data. The text can reveal information about the knowledge of economic actors in so far it reflects that knowledge. For example, scholars have long noted the limitations of patents (e.g., Pakes and Griliches, 1980; Griliches, 1990) and academic publications (Nelson, 2009) as paper trails of innovation. The amount of available data is also relevant. First, like many other NLP techniques including vector space model, topic modeling does not perform well with short and sparse text data (Popescul et al., 2001; Cheng et al., 2014; Puniyani et al., 2010) and it is difficult to infer robust patterns when the available information represents the topic of interests sparingly. Increasing the amount of data is not sufficient to address the problem – the data need to be relevant, as well. Even substantial amounts of text will not enable algorithms to uncover underlying idea patterns if that text reflects redundant information about actors' knowledge bases.

III.4. A step-by-step guide to implementing our algorithm

In this section, we introduce a step-by-step guide on how to implement our technique.[10] In the next section, we follow the steps to benchmark and demonstrate our algorithm using patent data.

Step 1: Select the text corpora in line with your research goals

Before employing our technique, it is important to ask if the benefits the approach brings, as listed above, are relevant in the context of your topic of interest. The consideration is akin to selecting the most appropriate regression model. Just because several approaches are available of which some are newer than others, does not mean the most recent approach is the most apt. Moreover, the choice of technique does not need to be exclusive. The HDP-based approach can be used in conjunction with other techniques to develop several measures of concepts of interest, each of which possesses particular advantages and disadvantages.

If our technique fits the research topic, the next step is to select the data type. As we discussed in the previous section, our technique could be applied to any text data and any combination of text data. However, even if our algorithm accommodates such an approach, the choice of data rests with the researcher who has the expertise to judge the fit between various dataset options and the research topic of interest. Choices of bin split, such as by year or by state, and unit of analysis, such as firm or region, are part of this step.

For example, in Furman and Teodoridis (2020), we applied our algorithm to 14 years of academic publications and conference proceedings in computer science, electrical engineering and electronics with the goal of developing measures that capture changes in the diversity and trajectory of researchers in response to automation of certain research tasks. Because we were interested in researchers' choices, we decided that the text data of academic publication abstracts

[10] The algorithm is available at https://github.com/jinolu/Direction_of_Innovation-HDP.

was appropriate to trace topics of research. We employed the HDP-based technique because other approaches available to us, such as taxonomies and citations, each suffered from limitations in their ability to capture the diversity and the trajectory of research. Nevertheless, we included measures based on these other data in addition to the HDP-based ones and discussed advantages and disadvantages of each. Next, we chose to apply the HDP-based algorithm on data split in yearly bins because, at that time, taken together, the text corpus was too large to be analyzed all at once and because our focus was on tracing changes in diversity and trajectory over time.

Step 2: Process the vocabulary

We suggest following the standard approach in topic modeling for formatting the input text. Doing this requires that we introduce definitions for two terms, “stemming” and “lemmatization”: (1) Stemming algorithms reduce words to their stems. For example, stemming could reduce each of the words ‘signals,’ ‘signaling,’ and ‘signaled,’ to the same stem, ‘signal.’ (2) Lemmatization resolves words to their dictionary form. For example, lemmatization could reduce ‘am,’ ‘are,’ and ‘is,’ to ‘to be.’ Stemming is simpler because it typically involves simply cutting off suffixes and conjugations, whereas lemmatization requires additional computational linguistics to recognize and resolve parts of speech as well as other linguistic nuances. To format the input text, we suggest eliminating all stop words and words with only one character, lowercasing all text, and then, lemmatizing the words from their inflected forms. We suggest lemmatization instead of stemming because, unlike stemming, lemmatization performs morphological analysis of the words and thus brings context to the lemma (Singh and Gupta, 2017).

More advanced vocabulary processing steps could be taken. These are not, however, crucial if the research goal is to obtain a mapping of the knowledge space reflected in your text document rather than to interpret generated topics. The more advanced steps depart from a bag-of-words assumption, where all words are considered equidistant from one another in terms of meaning, to account for the linguistic context. There are two such techniques: n-grams and embeddings. N-grams allow the algorithm to distinguish between words that have a certain meaning when taken together and another when considered separately. It is common to focus on bigrams, namely on groups of two words that have a meaning distinct from that of each of the words separately (e.g., Krieger, Schnitzer, and Watzinger, 2022; Arts et al, 2021). Embeddings take the goal of capturing nuanced meaning one step further by considering the meaning of words in context. The topic modeling is trained to learn the context from a training dataset of text that uses words in a similar manner as the text to be analyzed. For example, a popular approach is to use a Wikipedia embedding if the text to be processed is considered to use the English language in a similar manner with the Wikipedia articles. Technically, an embedding is thus a special type of word representation where the words that appear frequently together are considered to be related or similar in meaning. With rapid advancements in computer science, embedding such as the Wikipedia one, are available off the shelf. More advanced users can also train their own embeddings.

Step 3: Generate the topics and the document membership data for each HDP bin

Once the vocabulary pre-process decisions are made and implemented, the modified HDP algorithm we describe can be executed. We follow best practices and suggest the 1,000 iterations built-in stopping rule (Boyd-Graber et al., 2017; Mimno and Blei, 2011). More advanced users can adjust the stopping rule based on their needs. Regardless of the chosen stopping rule, we suggest checking the algorithm's performance by plotting the model likelihood and number of identified topics per number of iterations. These are the classic performance metrics used by HDP specialists

(Boyd-Graber et al., 2017).[11] The plots indicate how the likelihood metric and the algorithm-identified number of topics change with each iteration. A good point to stop iterating is when the plots show a relatively flat line, namely no more changes in the likelihood metric and in the algorithm-identified number of topics. This convergence of the likelihood metric and that of the number of topics indicates that the algorithm performance is optimized, and the optimal number of topics is identified (Boyd-Graber et al., 2017). Once completed, the algorithm outputs the list of topics and the probabilistic topic membership data for each input text document.

Step 4: Generate measures

The ultimate goal of the topic modeling method we describe is to construct measures that can be used as covariates in causal or correlational empirical analysis. These measures can be constructed at any level of analysis, as needed. For example, the HDP output data can be used to calculate measures of industry concentration or firm knowledge diversity, such as the Herfindahl-Hirschman index, the Euclidean distance and the Gini index. Researchers can devise any other measures as it fits the topic of their inquiry.

IV. Algorithm benchmarking and benefits

IV.1. HDP on patent data

To benchmark and demonstrate the algorithm's benefits, we apply it to 44 years of USPTO utility patents, from 1976 to 2019, and share the resulting data. This yields 5,871,621 patent records analyzed through our algorithm. We explain the patent application and our rationale for focusing on patent data to benchmark and demonstrate the benefits of our algorithm by following the step-by-step guide above.

[11] The likelihood estimates how likely the model can describe the data, given the current model parameters.

Step 1: Select the text corpora in line with your research goals

Our main goal is to examine whether our algorithm performs as we describe. We focus on patents because the data provide two advantages towards this goal. First, patent data are a widely used paper trail of knowledge production in studies of innovation research, whose value and limitations have been extensively documented. Second, patents are categorized using a carefully curated taxonomy, the Cooperative Patent Classification (CPC) schema, that is frequently updated to reflect changes in the patent knowledge space;[12] following each update, the patents are re-categorized as needed. This limits some typical disadvantages of taxonomies, namely classification standardization over time at the cost of delayed integration of new knowledge trajectories, but not all. Specifically, the limitation of equidistance assumption between classes persists in the CPC taxonomy (e.g., Ghosh et al., 2016). We exploit this variation in limitations to benchmark our algorithm and to demonstrate its benefits.

In the previous section, we listed four benefits of the HDP approach. The patent application helps demonstrate the first three; the fourth one is intuitive and follows consequently. Specifically, if our approach indeed captures emergent trends, then it should yield a knowledge map that is strongly correlated with the CPC-based one, given that the CPC team painstakingly tracks and documents emergent trends. At the same time, our approach should capture more accurate information about distance and movement in the knowledge space because the CPC considers patent classes to be equidistant, whereas our approach captures degrees of similarity. The effect is likely not large given the extensive classification effort. Rather, we anticipate observing it in situations that mechanically lead to higher measurement noise driven by the equidistance assumption.

---

[12] https://www.cooperativepatentclassification.org/faq "How often is the CPC scheme expected to be revised? It is expected that there will be multiple revisions in a year. A multiyear plan will be established beforehand."

We run our algorithm on patent abstracts. Although there is no conceptual limit for extending to the full patent text, the limitation is computational. Computational requirements increase exponentially with the amount of text data. We follow prior research, which indicates that patent abstracts carry important information about the knowledge contained in patents, albeit not without limitations (e.g., Stuart, 2000). We apply the algorithm to the entire corpus of patent abstracts, from 1976 to 2019. We follow this approach because we want to obtain a topic-based mapping of the knowledge space that is conceptually comparable with the CPC schema, which also maps the entire corpus of patents.

Step 2: Process the vocabulary

First, we follow the standard suggested approach described above for formatting the text of the patent abstracts and eliminate all stop words and words with only one character. We also lowercase all abstract text, and we lemmatize the words. Second, we move from a bag-of-words assumption and implement a bigram approach. We follow the guidance of a large body of work that sets to improve on the patent classification techniques that highlight the benefit of bigrams in this particular context (D'hondt et al., 2013).

Step 3: Generate the topics and the document membership data for each HDP bin

We apply the HDP algorithm to the resulting collection of processed yearly patent abstracts and obtain as output the list of 77 HDP-generated topics and the words that describe each topic, along with their respective assigned probabilities. We use the built-in stopping rule of 1,000 iterations. To ensure the stopping rule is optimal, we plot the likelihood metric and the number of algorithm-identified topics per number of iterations (Figures 1a and 1b). The shape of the plotted curves suggests convergence at 1,000 iterations.

While the resulting collections of words comprising topics differ from what we might expect from a categorization system, it is good practice to check that the topics are, by and large, relatable. We manually check our 77 topics for this reason. To demonstrate, we include two examples of patent abstracts and their 3-digit CPC categorization alongside their assigned HDP topics in Figures 2a and 2b. It is important to note that, in both cases, the probabilities add up to one, and thus fully describe the patents across the different topics that characterize the latent structure of the patent corpus. It is also important to note that, in both cases, the HDP-assigned topics intuitively align with the 3-digit CPC classification of the patents.

Step 4: Generate measures

The goal for the HDP-generated data is to develop covariates that enable empirical analysis of the direction of innovation. For this example, we focus on a measure of firm-level diversification and we calculate an annual measure of knowledge diversity at the assignee level.[13] We define diversity based on the breadth an assignee's knowledge portfolio at each point in time, i.e., the set of knowledge topics tackled in any particular year. We calculate the diversification measure as a yearly spread of patent innovations across topics, as identified by the HDP algorithm. Specifically, we calculate the Herfindahl-Hirschman index (HHI) across the HDP identified topics and subtract it from one. For each assignee, we first calculate the weighted number of patents in each HDP identified topic in each year, where the weights equal the HDP-generated probability of each patent belonging to the topic. Then, we divide the resulting number by the assignee's total number of patents in each year to obtain the weighted yearly share of patents across the HDP identified topics. We use these shares to calculate the HHI. Last, we subtract the HHI from one to obtain the index of diversification. The approach to subtract from one is a mathematical convenience that allows

[13] We only focus on assignee type "US Company or Corporation" and "Foreign Company or Corporation", excluding individual assignees and governments.

us to interpret the resulting index as one that increases with the level of diversification, since higher values of the HHI indicate an increase in concentration of knowledge topics. (There are, of course, other measures of concentration or dispersion that we could using, including the Ellison-Glaeser dartboard index (1997) which could also address issues associated with smaller firms in any particular industry, whose smaller patent portfolios may lead to measurement challenges in an annual measure of knowledge diversification.)

IV.2. Algorithm performance and benefits

We compare the HDP-based diversification measure with an equivalent one calculated using 3-digit CPC patent classes. To obtain the CPC-based diversification index, we use a similar approach as above to calculate the HHI across CPC patent classes, at the assignee-year level. However, because the CPC assumes classes are equidistant, we also assume equally spread shares of patents across classes for patents assigned to more than one CPC 3-digit class.

We take two main steps to benchmark and demonstrate the benefits of our algorithm. First, we show that the HDP- and CPC-based diversification measures are correlated. Table 1, column 1 shows the correlation. We estimate a linear model with assignee fixed effects and robust standard errors clustered at the assignee level, since we are interested in within-assignee correlations between the two diversification measures.

Second, we show that the strength of the correlation varies were expected. Specifically, the equidistance assumption of the CPC taxonomy means that the lower the number of patents of an assignee, the higher the likelihood that the diversification index is over or underestimated. This is a mathematical consequence of the HHI formula. Thus, if our approach is better equipped to capture the distance in knowledge space between patents, then we should see a higher discrepancy between the CPC-based diversification index and the HDP-based one for assignees with a lower number of patents, when compared with assignees with a relatively higher number of patents. We show evidence of these effects in Table 1, columns 2 and 3. Specifically, we focus on assignees with a count of patents in the 90th percentile of the distribution (column 2) and in the 99th percentile (column 3). We find that the CPC-based diversification measure overestimates the level of diversification of assignees with few patents, relative to assignees with a larger number of patents, as evidenced by the coefficients of the interaction terms. Specifically, the interaction term is higher in column 2 (p-value=0.000) then in column 3 (p-value=0.000), namely when we employ a stricter cut-off for assignees with many patents. In other words, the larger the number of patents an assignee has, the lower the likelihood that the equidistance assumption of the CPC taxonomy mechanically miss-estimates diversification, and hence the closer to 1 the estimated correlation between the two diversification measures; a correlation of 1 would denote that the CPC- and HDP-based indexes are perfect substitutes. The result is robust to any other cut-off points for distinguishing between assignees with fewer versus more patents.

To further demonstrate the benefits of our algorithm, we exploit another source of variation where the equidistance assumption might be playing a role in mechanically over- or under-estimating diversity. In some industries, patents that are classified in two or more 3-digit CPC classes are, in fact, closer or further away in knowledge space than the 50-50 split implied by the equidistance assumption of the CPC-taxonomy. It is then possible that the magnitude of the correlation between the CPC- and the HDP-based diversification measures might vary across industries. To show evidence of these effects, we take advantage of the dataset provided by Kogan et al. (2017), which matches the USPTO patent database to the CRSP/Compustat database. We merge our dataset with the Kogan et al. (2017) data using the unique patent identifiers to link our assignee-level diversity measures with the CRSP/Compustat's permanent company identifiers. We then retrieve the Standard Industry Classification (SIC) code for each assignee from the CRSP/Compustat database. We focus on the SIC because the 2-digit SIC codes provide a more granular categorization than the 2-digit North America Industry Classification (NAICS) schema, which is also available in the CRSP/Compustat data. Our goal is to evaluate the possibility that there is heterogeneity in the correlation between the CPC- and HDP-based diversity measures. The higher aggregation level of the 2-digit NAICS codes means less of an opportunity to observe the presence of such heterogeneity. Nevertheless, because our data includes NAICS codes, we are able to confirm similar patterns across industry classification systems, as expected. Certainly, the patterns we observe need to be interpreted with care given that the statistical power for across-industry estimates decreases with increases in the level of granularity of industry classification codes (e.g., from 2-digit to 3-digit NAICS codes).

In Table 2, we show evidence of variation in the correlation strength between the two diversification measures using 2-digit SIC codes. As before, we estimate a linear model with assignee fixed effects and robust standard errors clustered at the assignee level. We restrict the estimation to SIC codes with a number of observations above the 25th percentile; as mentioned previously, within-industry estimations with few observations have a high sampling variance and hence lack statistical precision. The restriction reduces the number of observations from 53,193 to 43,360, and the number of 2-digit SIC codes from 67 to 12; clearly our data are not sufficient to provide robust estimates for those industry codes with observations below the 25th percentile. Table 2, column 1 shows results restricted to industry codes with a number of observations above the 25th percentile and column 2 shows results restricted to industry codes with a number of observations above the 50th percentile. As expected, there are industries where the CPC-based diversification index is overestimated, such as SIC 28 "Chemical and Allied Products" and others where the index is underestimated, relative to the HDP-based index, such as SIC 35 "Industrial and Commercial Machinery and Computer Equipment." There are also industries where the two indexes capture similar levels of diversification, such as SIC 13 "Oil and Gas Extraction", SIC 20 "Food and Kindred Products" and SIC 33 "Primary Metal Industries." Among all industries, and within the bounds of our data sample, SIC 28 seems to exhibit the highest and most persistent discrepancy. Figure 3 shows plots of the two diversification indexes for several firms categorized under SIC 28. In all cases, the HDP-based index exhibits similar yearly patterns as the CPC-based index, but at a lower value. Intuition suggests that the effects might be driven by drug and other chemical developments that have practical applications to a wider set of areas. For example, Figure 4 shows the CPC and HDP classification of a Pfizer Inc patent that introduces chemical substances that are "antibacterial agents and agents for promoting growth in farm animals." The patent is classified in two 3-digit CPC classes - C07D "Heterocyclic Compounds" and A23K "Fodder." Mechanically, the HHI index formula gives equal weights to the two CPC classes. However, the patent is clearly more about the chemical substance, i.e., CPC class C07D, then about growth in farm animals i.e., A23K. Our method captures this difference as demonstrated by the weights assigned to the two topics describing the patent. The first topic, the one describing the chemical substance, and hence the equivalent to the C07D classification, is given a weight of 0.89. The second topic, the one describing the farm animal connection, and hence the equivalent to the A23K classification, is given a weight of only 0.11.

It is important to note that, through this exercise, our goal is not to sharply identify all areas where the HDP-based diversification index might be different in magnitude from the CPC-based one. Nor is our main goal to improve upon the CPC classification system. Rather, we conduct comparisons between the HDP- and CPC-based approaches to benchmark and demonstrate the benefits of the HDP method. The CPC classification is well documented, and the patent system as a paper trail of knowledge is well understood, thus allowing us to observe if our HDP-based approach behaves as expected, where expected. The highest return to using the HDP-based method comes from applying the algorithm to text corpora that lack a robust classification system, thus uncovering knowledge patterns that were otherwise hidden to the researcher.

V. Conclusion

In this paper, we describe a topic modeling-based algorithm that helps map the knowledge space as captured in text documents. This enables researchers to measure distance and movement in the knowledge landscape in a more flexible and comprehensive manner. We apply our method to 44 years of patent data and compare the HDP approach with the established and carefully curated CPC taxonomy, to benchmark our method and to demonstrate its benefits.

The approach we describe is not the first to leverage text data to map the knowledge space. Indeed, curated taxonomies have been useful in this regard for quite some time. Scholars have also made use of keywords and citation linkages. Because all these approaches suffer from limitations, more recently, scholars have been exploiting NLP methods such as vector space models (e.g., cosine similarity, Jaccard indexes) to reflect the similarity between ideas or sets of ideas. These are particularly good at capturing similarity between two bodies of text i.e., the distance between two bodies of text, but face greater challenges in mapping the entirety of a relevant knowledge space. The HDP-based algorithm that we introduce and describe in this paper builds on and extends these efforts to a flexible approach that can be applied to any corpus of text. Along with LDA, HDP belongs to a family of probabilistic topic models that can uncover the structure underlying an otherwise unstructured collection of documents. Whereas LDA requires that scholars calibrate its operation by specifying the number of topics into which a corpus of text should be categorized, HDP determines the optimal number of categories from the corpus itself and probabilistically assigns documents to those categories. Unlike human-generated models, it adjusts category boundaries as the corpus of text updates.

While our work in this paper has focused on a relatively classic level of aggregation – firm-year –, the approach is sufficiently flexible to be applied to different other levels of analysis. For example, the algorithm can be used for both broad sources of text (e.g., the entire patent corpus) or narrower sources (e.g., specific time periods, fields, or subfields), and it can be applied to trace changes over time (e.g., month, year, seasons), by geographies (e.g., cities, counties, states) or by other demographic attributes (e.g., gender, ethnicity, age).

Once applied to a body of text, our approach can be used to describe distance and movement of ideas in knowledge space, measures that benefit several research questions of interest to management scholars. For example, in Furman and Teodoridis (2020), we successfully apply the same techniques to a large body of academic papers to evaluate the role of automating research technologies in influencing the rate and direction of academic research, an important topic of study that has been lagging due to difficulties of measuring changes in the direction of innovation. Also, Lu (2022) applies these techniques to evaluate the impact of an innovation policy targeted at increasing attention to a certain area of research on alternative research lines, with implications for the socially optimal diversity of research (Arrow, 1962; Scotchmer, 1991; Acemoglu, 2012). His effort extends frontier papers on this topic that are limited in their ability to capture implications for changes in attention to research trajectories.

More broadly, our paper adds to efforts to democratize access to tools and techniques that enable more nuanced measures of innovation activity. The approach we describe can be applied to any information gleaned from texts generated by corporate R&D, published statements, or corporate documents, product announcements, trade journals, or other sources of text that characterize knowledge bases. Available techniques based on taxonomies, keywords, or citations do not immediately, or at all, extend to these other data sources. Our technique can generate comparable measures across such varied data sources and even based on combinations of such datasets.

References

Abrams, David S., Ufuk Akcigit, and Jillian Brennan (2019) "Patent Value and Citations: Creative Destruction or Strategic Disruption?" working paper, also *U of Penn, Inst for Law & Econ Research Paper No. 13-23*.

Acemoglu, Daron (2012) “Diversity and Technological Progress,” in *The Rate and Direction of Inventive Activity Revisited*, Stern S, Lerner J (eds). University of Chicago Press: Chicago, IL, 319– 356.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein (2008) “Academic freedom, private-sector focus, and the process of innovation,” *RAND Journal of Economics*, 39(3), 617-635.

Adams, James D., Eric P. Chiang, and Jeffrey L. Jensen (2003) “The influence of federal laboratory R&D on industrial research,” *Review of Economics and Statistics* 85(4), 1003-1020.

Agrawal, Ajay, Avi Goldfarb, and Florenta Teodoridis (2016) “Understanding the Changing Structure of Scientific Inquiry,” *American Economic Journal: Applied Economics*, 8 (1): 100–128.

Alcacer, Juan and Michelle Gittelman (2006) “Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations,” *Review of Economics and Statistics*, 88(4), 774-779.

Arora, Ashish, and Alfonso Gambardella (2010) "The market for technology." *Handbook of the Economics of Innovation*, 1, 641-678.

Arrow, Kenneth (1962) “Economic Welfare and The Allocation of Resources for Invention,” in *The Rate and Direction of Inventive Activity*, Princeton University Press and NBER.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez (2018) “Text Matching to Measure Patent Similarity,” *Strategic Management Journal*, 39 (1): 62–84.

Arts, Sam, Hou, Jianan, & Gomez, Juan Carlos (2021) “Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures,” *Research Policy*, 50(2), 104144.

Athey, Susan (2018) “The impact of machine learning on economics,” in Agrawal, Ajay, Avi Goldfarb, and Joshua Gans, Eds.,. *The Economics of Artificial Intelligence: An agenda*, University of Chicago Press, 507-547.

Azoulay, Pierre, Waverly Ding, and Toby Stuart (2009) “The impact of academic patenting on the rate, quality and direction of (public) research output,” The Journal of Industrial *Economics*, 57(4), 637-676.

Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona E. Murray (2015) “Retractions,” *Review of Economics and Statistics*, 97(5), 1118-1136.

Barde, Bhagyashree Vyankatrao and Anant Madhavrao Bainwad (2017) “An Overview of Topic Modeling Methods and Tools,” In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS),* 745–50. IEEE.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, 3 (Jan): 993–1022.

Borjas, George J. and Kirk B. Doran (2012) “The collapse of the Soviet Union and the productivity of American mathematicians,” *Quarterly Journal of Economics,* 127(3), 1143-1203.

Borjas, George J. and Kirk B. Doran (2015) “Prizes and productivity how winning the fields medal affects scientific output." *Journal of Human Resources,* 50(3), 728-758.

Boyd-Graber, Jordan, Yeuning Hu, and David Mimno (2017) "Applications of topic Models," *Foundations and Trends in Information Retrieval,* 20(20), 1-154.

Bryan, Kevin and Jorge Lemus (2017) “The direction of innovation,” *Journal of Economic Theory* 172, 247-272.

Buntine, Wray and Aleks Jakulin (2004) “Applying Discrete PCA in Data Analysis,” In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 59–66. AUAI Press.

Cheng, X., Lan, Y., Guo, J., Yan, X. (2014) “Btm: Topic modeling over short texts,” *IEEE Transactions Knowledge Data Engineering*, 26(12), 2928–2941.

Choudhury, Prithwiraj, Dan Wang, Natalie A. Carlson, and Tarun Khanna (2019) “Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles,” *Strategic Management Journal*, 40 (11): 1705–32.

Cohen, Wesley M. and Daniel A. Levinthal (1989) “Innovation and learning: the two faces of R&D,” *The Economic Journal*, 99(397), 569-596.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh (2002) “Links and impacts: the influence of public research on industrial R&D.” *Management Science*, 48(1), 1-23.

de Solla Price, Derek (1976) “A General Theory of Bibliometric and Other Cumulative Advantage Processes,” *Journal of the American Society for Information Science*, 27(5): 292–306.

Dasgupta, Partha and Paul A. David (1994) “Toward a new economics of science,” *Research Policy,* 23(5) 487-521.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990) “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, 41 (6): 391–407.

D'hondt, Eva, Suzan Verberne, Cornelis Koster, and Lou Boves (2013) “Text representations for patent classification,” *Computational Linguistics* 39(3), 755-775.

Doran, Kirk (2017) “The Direction of Science: Natural Experiments,” presentation at the June 2017 DRUID Conference, New York, NY.

Ellison, Glenn, and Edward L. Glaeser (1997) “Geographic concentration in US manufacturing industries: a dartboard approach,” *Journal of Political Economy*, 105(5), 889-927.

Franzoni, Chiara Franzoni, Chiara, Giuseppe Scellato, and Paula Stephan (2011) “Changing incentives to publish,” *Science* 333(6043), 702-703.

Furman, Jeffrey L. and Scott Stern (2011) “Climbing atop the shoulders of giants: The impact of institutions on cumulative research,” *American Economic Review*, 101(5), 1933-63.

Furman, Jeffrey L. and Florenta Teodoridis (2020) “Automation, Research Technology, and Researchers’ Trajectories: Evidence from Computer Science and Electrical Engineering,” *Organization Science*, 31 (2): 330–54.

Furman, Jeffrey L. and Florenta Teodoridis (2020) “Machine learning could improve innovation policy,” *Nature Machine Intelligence* 2(2), 84.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2019) "Modeling consumer footprints on search engines: An interplay with social media." *Management Science*, 65(3), 1363-1385.

Ghosh, Anindya, Ram Ranganathan, and Lori Rosenkopf (2016) "The impact of context and model choice on the determinants of strategic alliance formation: Evidence from a staged replication study," *Strategic Management Journal*, 37(11), 2204-2221.

Griliches, Zvi (1990) "Patent Statistics as Economic Indicators: A Survey," *Journal of Economic Literature*, 28(4): 1661–707.
Grupp, Hariolf (1990) "The concept of entropy in scientometrics and innovation research: An indicator for institutional involvement in scientific and technological developments," *Scientometrics*, 18(3-4), 219-239.
Hansen, Stephen, Michael McMahon, and Andrea Prat (2018) "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 133 (2): 801–70.
Henderson, Rebecca, and Iain Cockburn (1994) "Measuring competence? Exploring firm effects in pharmaceutical research," *Strategic Management Journal*, 15(S1), 63-84.
Henderson, Rebecca and Iain Cockburn (1996) "Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery," *RAND Journal of Economics*, 27(1), 32-59.
Henderson, Rebecca, Adam Jaffe, and Manuel Trajtenberg (2005) "Patent citations and the geography of knowledge spillovers: A reassessment: Comment," *American Economic Review*, 95(1), 461-464.
Hill, Ryan and Carolyn Stein (2022) "Race to the bottom: Competition and quality in science," Working Paper.
Hoberg, Gerard and Gordon Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 124 (5): 1423–65.
Hoffmann, Thomas (1999) "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289-296.
Huang, Allen H., Reuven Lehavy, Amy Y. Zang, and Rong Zheng (2018) "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science*, 64 (6): 2833–55.
Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger (2018) "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science," *The Quarterly Journal of Economics*, 133 (2): 927–91.
Jaffe, Adam B. (1989) "Real effects of academic research," *American Economic Review*, 957-970.
Jones, Benjamin (2009) "The burden of knowledge and the 'Death of the Renaissance Man': Is innovation getting harder?" *The Review of Economic Studies*, 76(1), 283-317.
Kaplan, Sarah and Keyvan Vakili (2015) "The Double - edged Sword of Recombination in Breakthrough Innovation," *Strategic Management Journal*, 36 (10): 1435–57.
Kauffman, Stuart A. (1993) *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA.
Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2020). Measuring technological innovation over the long run. *American Economic Review: Insights*.
Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018) "Human decisions and machine predictions," *Quarterly Journal of Economics*, 133(1), 237-293.
Klepper, Steven (1996) "Entry, Exit, Growth, and Innovation over the Product Life Cycle," *The American Economic Review*, 562–83.
Klepper, Steven (1997) "Industry Life Cycles," *Industrial and Corporate Change*, 6 (1): 145–82.
Klepper, Steven and Elizabeth Graddy (1990) 'The evolution of new industries and the determinants of market structure," *Rand Journal of Economics*, 21(1), 27–44.
Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, Noah Stoffman (2017) "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics*, 132(2), 665–712.
Koning, Rembrand, Samila, Samson and Ferguson, John Paul (2020) "Inventor Gender and the Direction of Invention," *AEA Papers and Proceedings*, 110, 250-254.
Krieger, Joshua L., Danielle Li, Dimitris Papanikolaou (2022), "Missing Novelty in Drug Development," *The Review of Financial Studies*, 35(2), 636–679, also NBER Working Paper #24595 (2019).
Krieger, Joshua Lev, Monika Schnitzer, and Martin Watzinger (2022) "Standing on the Shoulders of Science," Harvard Business School Working Paper, No. 21-128.

Kuhn, Jeffrey., Kenneth Younge, and Alan Marco. (2020) “Patent citations reexamined,” *RAND Journal of Economics*, 51(1), 109-132.

Lei, Lei, Jiaju Qi, and Kan Zheng (2019) “Patent Analytics Based on Feature Vector Space Model: A Case of IoT,” *IEEE Access*, 7: 45705–15.

Lerner, Josh and Scott Stern (2015) *The Rate and Direction of Inventive Activity Revisited*. Chicago: IL: University of Chicago Press.

Lu, Jino (2022) “Demand for innovation and Direction of Research: The Roles of Intellectual Distance and Research Productivity,” working paper.

Magerman, Tom, Bart Van Looy, and Koenraad Debackere (2015) “Does Involvement in Patenting Jeopardize One’s Academic Footprint? An Analysis of Patent-Paper Pairs in Biotechnology,” *Research Policy*, 44 (9): 1702–13.

Machlup, Fritz (1961) “Patents and Inventive Effort: The evidence is insufficient to prove or disprove the claim that patent protection promotes inventive effort,” *Science* 133(3463), 1463-1466.

March James G (1991) “Exploration and Exploitation in Organizational Learning,” *Organization Science* 2, 71–87.

Marx, Matt and Aaron Fuegi (2019) “Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles,” *Boston University Questrom School of Business Research Paper*, no. 3331686.

Merton, Robert K. (1957) “Priorities in scientific discovery: A chapter in the sociology of science,” *American Sociological Review*, 22(6) 635-659.

Mimno, David and David Blei (2011) "Bayesian Checking for Topic Model," *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, Jul 27-31, 227–237

Myers, Kyle (2020) “The Elasticity of Science,” *American Economic Journal: Applied Economics*, 12(4), 103–134.

National Bureau of Economic Research (1962) *The Rate and Direction of Inventive Activity Economic and Social Factors*. Princeton, NJ: Princeton University Press.

Pakes, Ariel and Zvi Griliches (1980), "Patents and R&D at the Firm Level: A First Report," *Economics Letters*, 5 (4), 377-81

Penrose, Edith T. (1959) *The Theory of the Growth of the Firm*. Oxford, UK: Basil Blackwell.

Png, Ivan P L (2019) “US R&D, 1975–1998: A New Dataset,” *Strategic Management Journal*, 40 (5): 715–35.

Popescul, A., L. Ungar, D. Pennock, and S. Lawrence. (2001) “Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments,” *in Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference.*

Puniyani, K., Eisenstein, J., Cohen, S. B., & Xing, E. (2010) "Social links from latent topics in microblogs," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, 19-20.

Righi, Cesare, and Timothy Simcoe (2019) "Patent examiner specialization." *Research Policy*

Rivkin, Jan W. and Nicolaj Siggelkow (2003) "Balancing Search and Stability: Interdependencies Among Elements of Organizational Design, *Management Science*, 49(3), 255-350.

Rivkin, Jan W. and Nicolaj Siggelkow (2007) "Patterned Interactions in Complex Systems: Implications for Exploration," *Management Science*, 53(7), 1068-1085.

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975) “A Vector Space Model for Automatic Indexing,” *Communications of the ACM*, 18 (11): 613–20.

Scotchmer, Suzanne (1991) “Standing on the shoulders of giants: Cumulative research and the patent law,” *Journal of Economic Perspectives*, 5(1), 29-41.

Shahmirzadi, Omid, Adam Lugowski, and Kenneth Younge (2019) “Text Similarity in Vector Space Models: A Comparative Study,” In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 659–66. IEEE.

Singh, Jasmeet and Vishal Gupta (2017) “A Systematic Review of Text Stemming Techniques,” *Artificial Intelligence Review*, 48 (2): 157–217.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang (2015) "An Overview of Microsoft Academic Service (Mas) and Applications," In *Proceedings of the 24th International Conference on World Wide Web*, 243–46.

Solow, Robert (1956) "A contribution to the theory of economic growth," *Quarterly Journal of Economics* 70(1), 65-94.

Stephan, Paula (1996) "The economics of science," *Journal of Economic Literature*, 34(3), 1199-1235.

Stephan, Paula (2012) "The economics of science," *Journal of Economic Literature*, 34(3), 1199-1235.

Stuart, Toby (2000) "Interorganizational alliances and the performance of firms: A study of growth and innovation," *Strategic Management Journal*, 21(8), 791-811

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei (2006) Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101(476), 1566-1581.,

Thompson, Peter and Melanie Fox-Kean (2005) "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 95(1) 450-460.

Vakili, Keyvan and Laurina Zhang (2018) "High on Creativity: The Impact of Social Liberalization Policies on Innovation," *Strategic Management Journal*, 39 (7): 1860–86.

Williams, Heidi L (2013) "Intellectual property rights and innovation: Evidence from the Human Genome," *Journal of Political Economy*, 121(1), 1-27.

Younge, Kenneth A and Jeffrey M. Kuhn (2019) "First Movers and Follow-on Invention: Evidence from a Vector Space Model of Invention," *Available at SSRN 3354530*.

Figure 1a. HDP patent implementation – likelihood metric

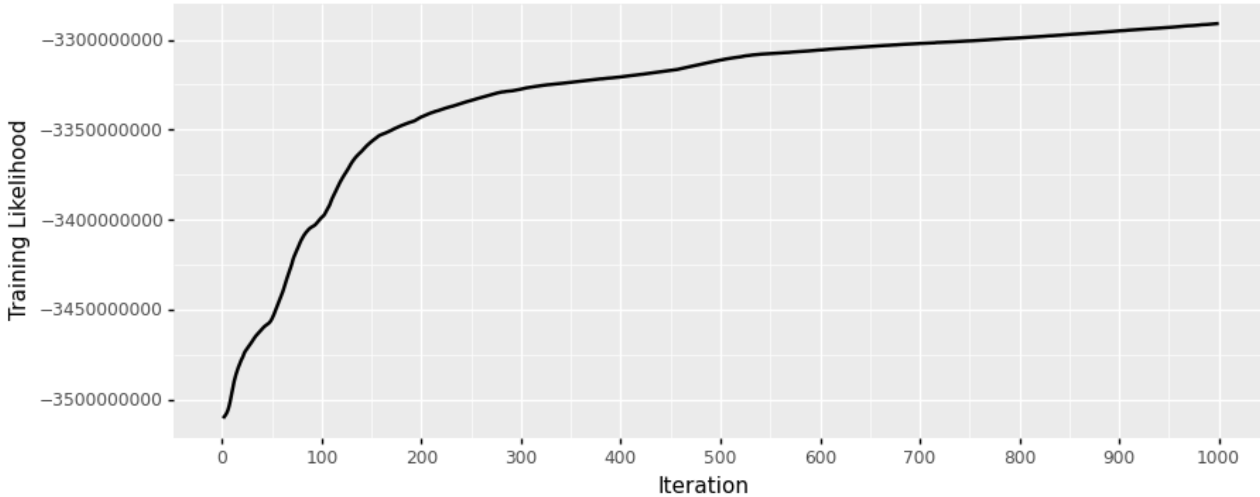


Figure 1b. HDP patent implementation – algorithm-identified number of topics

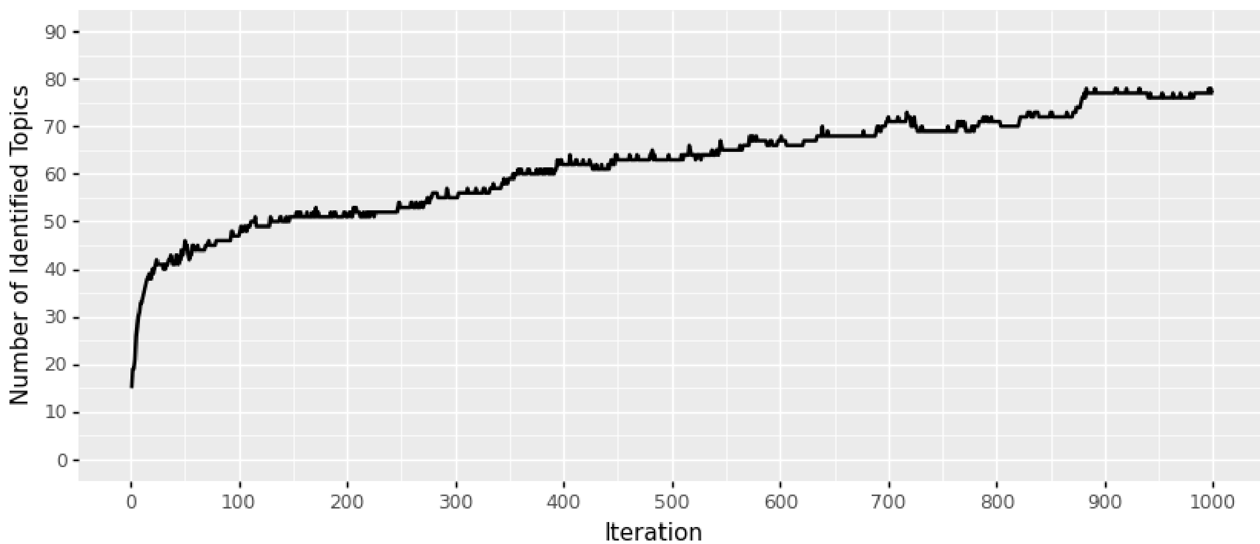


Figure 2a. Example of HDP mapping of a patent abstract

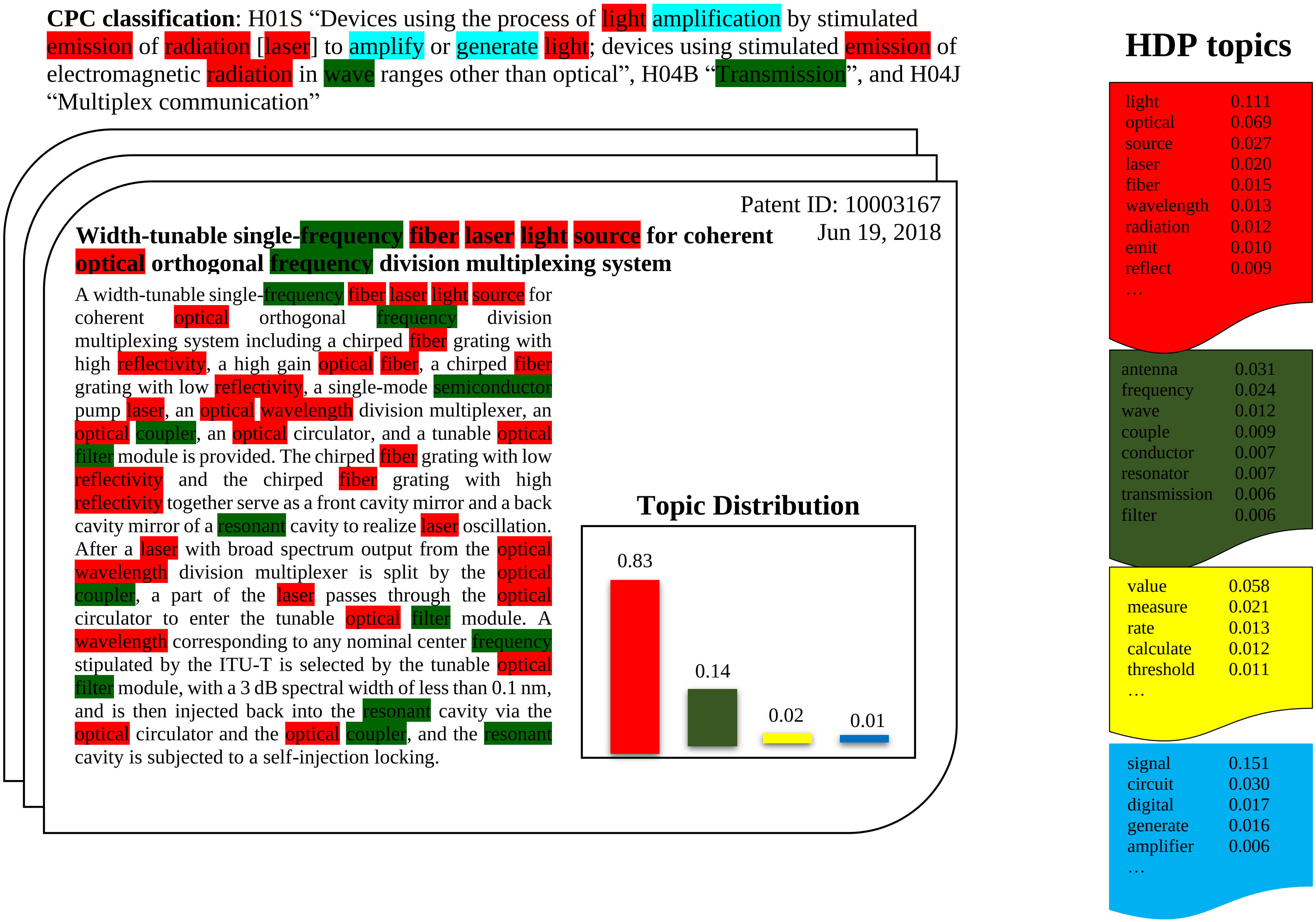


Figure 2b. Example of HDP mapping of a patent abstract

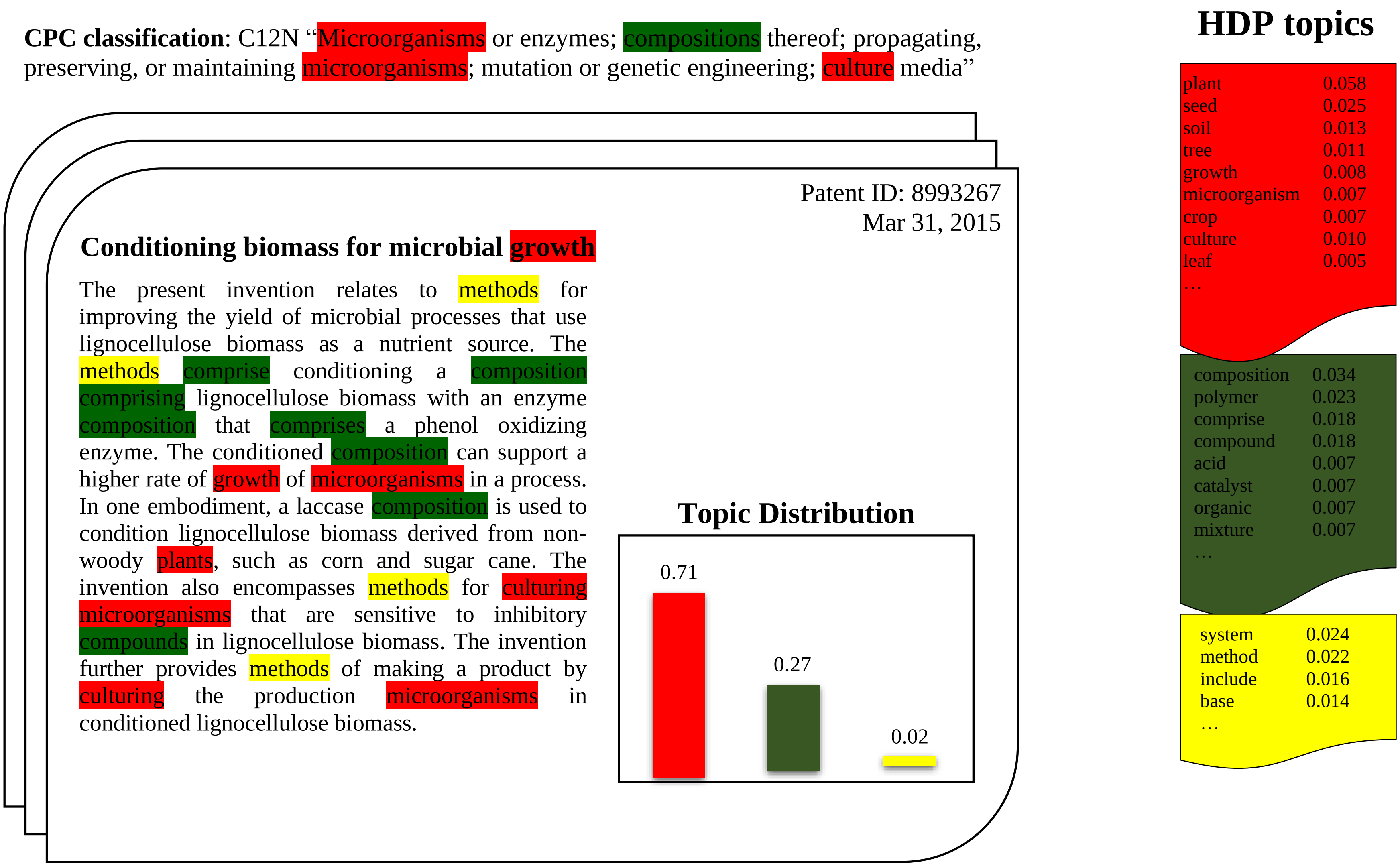
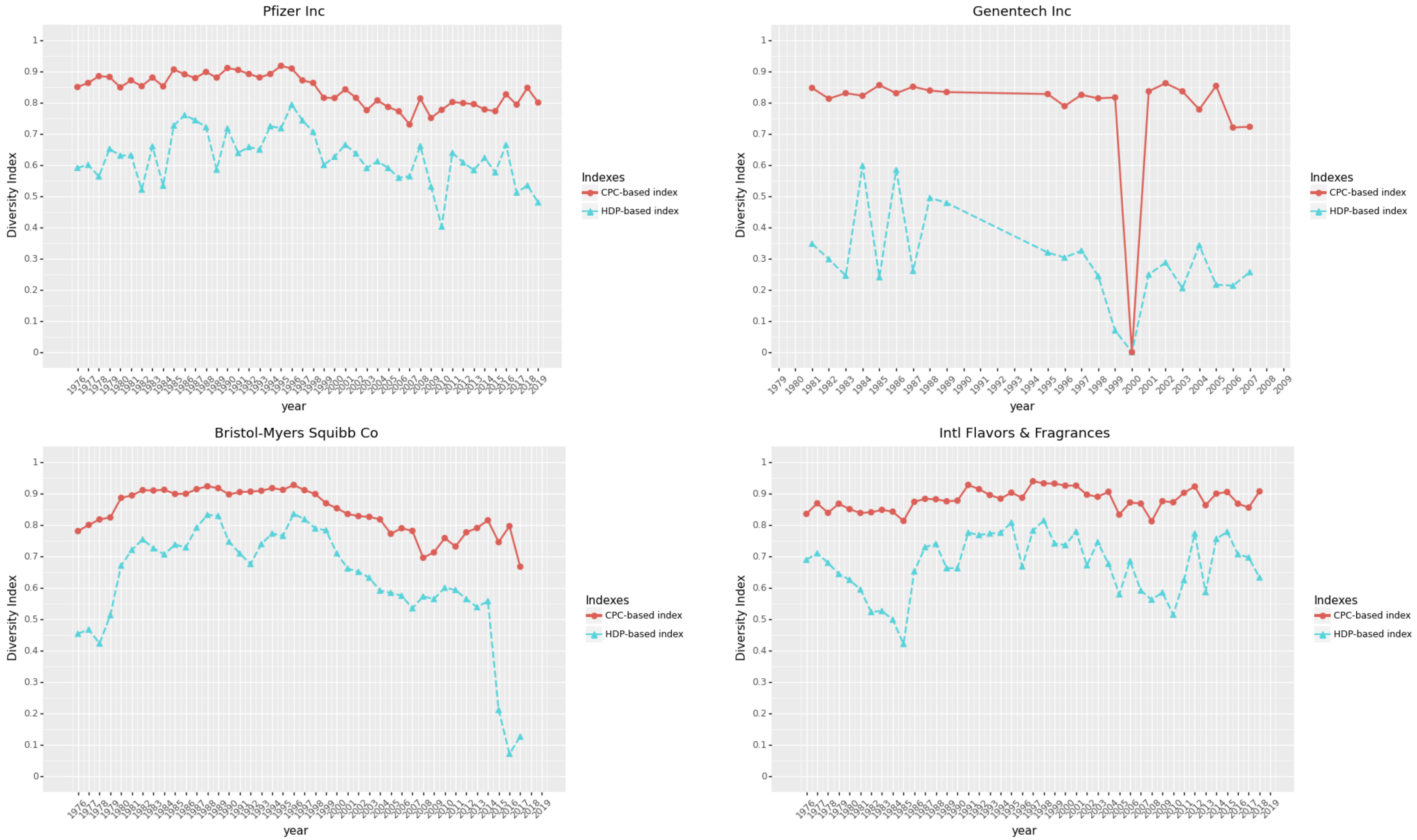
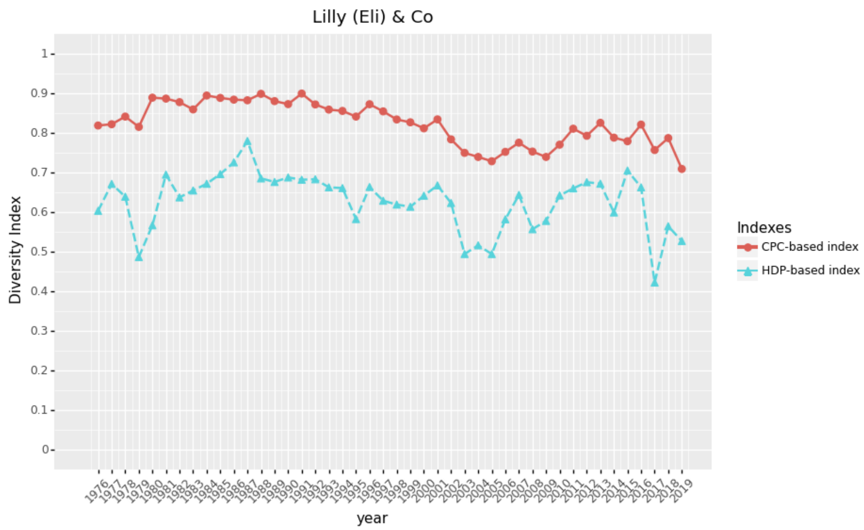


Figure 3: Examples of CPC- vs. HDP-based diversification index for SIC 28 firms




Lilly (Eli) & Co
Diversity Index
year
Indexes
CPC-based index
HDP-based index

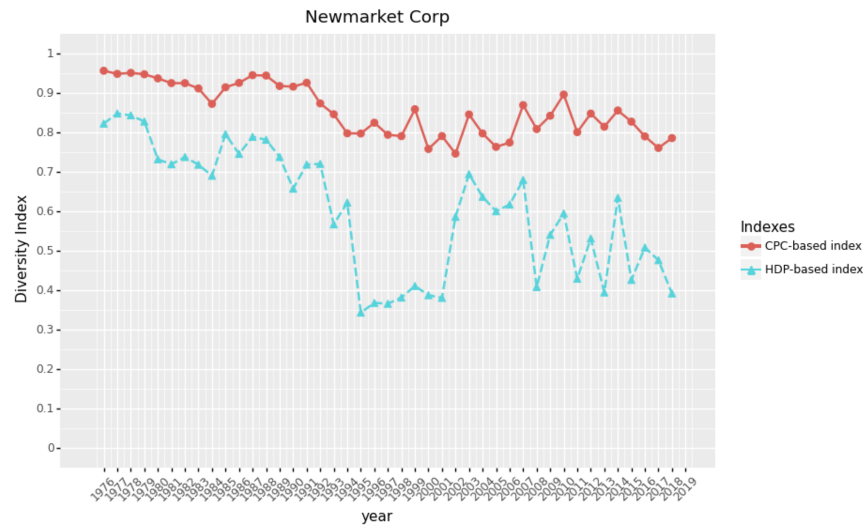

Newmarket Corp
Diversity Index
year
Indexes
CPC-based index
HDP-based index

Figure 4. Pfizer Inc. patent example

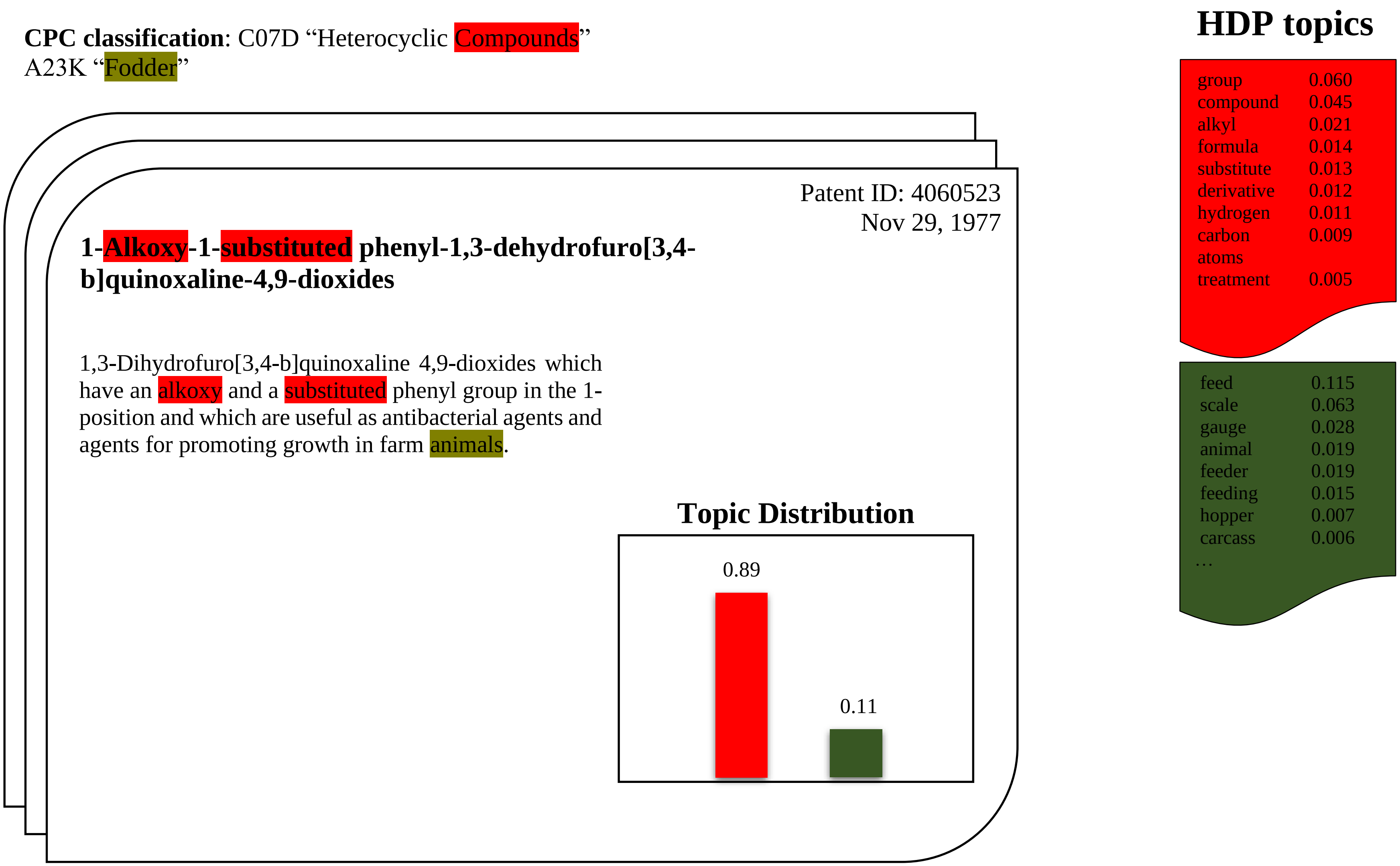


Table 1: Correlation between the CPC- and HDP-based indexes of diversification.

DV= CPC diversification index	**(1)**	**(2)**	**(3)**
	Overall	**Top patenting assignee (90^{th} percentile)**	**Top patenting assignee (99^{th} percentile)**
HDP diversification index	0.353***	0.212***	0.334***
	(0.003)	(0.003)	(0.003)
HDP diversification index * top patenting assignee		0.567***	0.806***
		(0.005)	(0.018)
Assignee FE	Yes	Yes	Yes
R-squared	0.031	0.121	0.077
Observations	843,752	843,752	843,752

Note: Data is a panel at the assignee-year level, 1976-2019. All models are OLS with assignee fixed effects and robust standard errors. We display the estimated coefficient alongside the standard error and the p-value – coefficient (st.error) ***significant at 1%, **significant at 5%, *significant at 10%.

Table 2: Correlation between the CPC- and HDP-based indexes of diversification by SIC industry code.

DV= CPC diversification index	**(1)**	**(2)**
HDP diversification index	0.647***	0.931***
	(0.030)	(0.032)
HDP diversification index x SIC 13 "Oil and Gas Extraction"	-0.015	
	(0.049)	
HDP diversification index x SIC 20 "Food and Kindred Products"	-0.056	
	(0.043)	
HDP diversification index x SIC 28 "Chemical and Allied Products"	-0.348***	-0.631***
	(0.030)	(0.028)
HDP diversification index x SIC 30 "Rubber and Miscellaneous Plastics Products"	0.166***	
	(0.060)	
HDP diversification index x SIC 33 "Primary Metal Industries"	0.019	
	(0.059)	
HDP diversification index x SIC 34 "Fabricated metal Products, Except Machinery and Transportation Equipment"	0.241***	
	(0.055)	
HDP diversification index x SIC 35 "Industrial and Commercial Machinery and Computer Equipment"	0.371***	0.087**
	(0.041)	(0.041)
HDP diversification index x SIC 36 "Electronic and Other Electrical Equipment and Components, Except Computer Equipment"	0.349***	0.066*
	(0.036)	(0.035)
HDP diversification index x SIC 37 "Transportation Equipment"	0.407***	
	(0.047)	
HDP diversification index x SIC 38 "Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks"	0.284***	
	(0.037)	
HDP diversification index x SIC 48 "Communications"	0.290***	
	(0.051)	
HDP diversification index x SIC 49 "Electric, Gas, and Sanitary Services"	-0.285***	
	(0.067)	
Individual covariates for 2-digit SIC (omitted from this table for brevity)	Yes	Yes
Assignee FE	Yes	Yes
R-squared	0.271	0.256
Observations	43,360	29,914

Note: Data is a panel at the assignee-year level, 1976-2019. All models are OLS with assignee fixed effects and robust standard errors. We display the estimated coefficient alongside the standard error and the p-value – coefficient (st.error) ***significant at 1%, **significant at 5%, *significant at 10%.

.