Measuring the direction of innovation: frontier tools in unassisted machine learning

Version	Other
Citation (published version):	F. Teodoridis, J. Lu, J.L. Furman. "Measuring the Direction of Innovation: Frontier Tools in Unassisted Machine Learning." Strategic Management Journal (2nd R&R, Fall 2021),

https://hdl.handle.net/2144/44291

Boston University

Mapping the Knowledge Space:
Frontier Tools in Unassisted Machine Learning*

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

Jino Lu
Marshall School of Business, University of Southern California
701 Exposition Blvd., Los Angeles, CA 90089
jinhong.lu@marshall.usc.edu

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

Abstract

As strategy research has increasingly recognized the roles of innovation and knowledge as drivers of firm- and industry-level outcomes, greater attention has been given to the effort to identify relationships among ideas and the distances between knowledge bases. In this paper, we develop a methodology that infers the mapping of the knowledge landscape based on researchers' 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 strategy and innovation. We benchmark demonstrate the benefits of our approach in the context of 44 years of USPTO data.

Keywords: topic modeling, machine learning, knowledge landscape, distance in knowledge space, movement in knowledge space, diversity, knowledge trajectories.

** All authors contributed equally. We are grateful to the editorial team for extremely constructive suggestions on the manuscript. In addition, we thank Richard Freeman, Shane Greenstein, Keyvan Vakili and Rosemarie Ziedonis for helpful suggestions on the text as well as participants in the Harvard University Science-Based Business Seminar, the Max Planck Institute for Innovation & Competition, Innovation, & Entrepreneurship, and seminars at the London Business School, Tulane University, the University of Maryland, and the University of Colorado. Jeff Furman gratefully acknowledges financial support from NSF SciSIP grant, SES-1564368. Florenta Teodoridis gratefully acknowledges financial support from NSF SciSIP grant, SES-1564381.*

I. Introduction

As strategy research has increasingly recognized the roles of innovation and knowledge as drivers of firm- and industry-level outcomes, greater attention has been given to the effort to identify relationships among the knowledge sets of economic actors. In this paper, we adapt a natural language processing (NLP) technique to develop a novel 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. Our goal is to enable strategy scholars to extend research topics that benefit from a more comprehensive, flexible, and accessible approach to quantifying the relationships between knowledge areas as captured in text documents.

Strategy scholars have long been using text documents such as patents, academic publications, 10k statements and news articles as a paper trail of firms' knowledge. A method that maps the knowledge landscape as captured in such documents enables a more comprehensive measurement of *distance* and *movement* of the knowledge sets. This benefits numerous research topics. How does firms' knowledge, when compared to that of competitors, influence firms' business and corporate strategy, response to start-up entry, decisions to engage in mergers, acquisitions or licensing and the success of those actions (e.g., Tanriverdi and Venkatraman, 2005; Makri et al., 2010; Ahuja and Novelli, 2014)? How does the assortment of firms' knowledge areas and the relationships between them influence the evolution of industries (Klepper and Graddy, 1990; Helfat and Lieberman, 2002; Moeen and Agarwal, 2017)? How do firm characteristics, such as size or capabilities, or industry characteristics, such as concentration and evolution stage, influence firms' innovation productivity and the types of innovations produced (e.g., Ahuja et al.,

[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.

2008)? How do characteristics of a market for technologies influence the type of knowledge produced (Arora and Gambardella, 2010)? These and similar research questions would benefit from empirical measures that are developed from a comprehensive mapping of the knowledge landscape.

To date, research on these topics primarily relies on measures derived either from observable indicators of innovation output, such 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; Arts et al., 2018; 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., Kuhn et al., 2020)[2]. NLP-based similarity measures address several of these limitations (Arts et al., 2018). However, they are most useful in comparing pairs or groups of knowledge sets (e.g., Oxley and Sampson, 2004; Sampson, 2007; Furman and Stern, 2011; Mindruta et al., 2016; Ranganathan and Rosenkopf, 2018; Vakili and Zhang, 2018; Kelly et al., 2020). As a result, they apply best to research questions that do not directly require access to observing a flexible mapping of the entire knowledge landscape.

[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.

In this paper, we extend efforts exploiting 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),[3] 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. For example, capturing changes in diversity of knowledge (i.e., breadth of distance) over time is of interest to many studies in strategy and innovation (Cohen and Levinthal, 1990; Fleming and Sorenson, 2004; Carnabuci and Operti, 2013; Moreira et al., 2018). This approach enables researchers to measure changes in knowledge diversity more accurately and timely than earlier described techniques because it accounts for distance between all contemporaneous knowledge areas at once (Furman and Teodoridis, 2020).

In developing these techniques, we join others' efforts to democratize access to frontier developments in machine learning (Choudhury et al., 2019; Choudhury et al., 2021) and to data that can be applied to the study of several questions of interest in strategy and innovation research (Arts et al., 2018; 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

[3] For a comprehensive review of all topic modeling algorithms, please see Boyd-Graber and coauthors (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. How mapping the knowledge landscape can advance strategy research

Mapping the knowledge landscape enables researchers to develop measures of both *distance* and *movement* in knowledge space. These measures could be used as either explanatory variables for outcomes of interest to management scholars or as the outcome of interest.

The idea that firms possess different portfolios of knowledge and that doing so has implications for firm performance has been central to strategic management since the origins of the field (Penrose, 1959). For example, several studies have evaluated the relationship between firms' knowledge overlap and competitive behavior (e.g., Chung and Yeaple, 2008; Ranganathan and Rosenkopf, 2014), alliances and acquisitions (e.g., Oxley and Sampson, 2004; Villalonga and McGahan, 2005; Cassiman et al., 2005; Mindruta et al., 2016; Ghosh et al., 2016) and innovation performance (e.g., Sampson, 2007; Alcacer and Oxley, 2014). Firms' knowledge and the relationship between them have also played key roles in both theoretical models and empirical studies of industry life cycles and industry evolution (Klepper and Graddy, 1990; Utterback and Suarez, 1993; Klepper, 1996, 1997; Moeen et al, 2020), including employee-firm fit and spin-offs (e.g., Agarwal et al., 2016; Kim and Steensma, 2017; Byun et al., 2019). For example, heterogeneity in firm-level knowledge has played particularly important roles in recent analyses of industry evolution, including in studies of firm spin-offs (Agarwal et al., 2004) and pre-entry experience (Moeen, 2017).

Strategy and innovation scholars have also investigated firm knowledge production as an outcome of interest in its own right. For example, by recognizing the essential tradeoff between exploration and exploitation (March, 1991), studies have examined the antecedents of search, determinants of the nature of search, and the outputs associated with heterogeneous search behaviors (e.g., Tripsas and Gavetti, 2000; Rosenkopf and Nerkar, 2001; Hoang and Rothaermel, 2010; Caner et al., 2017). Scholars have also evaluated the role of inter-firm connections, including both formalized (e.g., mergers and alliances) and informal connections (e.g., ecosystems (e.g., Adner and Kupoor, 2010; Gawer et al, 2018; Tandon and Toh, 2021)), in order to understand firms' innovation investments and outputs (e.g., Ahuja, 2000; Makri et al., 2010; Jiang et al., 2011; Kneeland et al., 2020).

By and large, research on these and similar topics has relied on codified, observable attributes of knowledge, such as patent technology classes and citations. While these approaches enable useful insights, strategy scholars acknowledge their multiple limitations and call for a more flexible and accurate mapping of the knowledge space. For example, Makri et al. (2010) note that "patent classification schemas […] were developed for another purpose" and that "broad data on science knowledge relatedness is difficult to obtain." Kneeland et al. (2020) note that "research that has used prior art citation measures […] suffers from issues of sequential interdependence (i.e., you can only cite what has come before you) and biases from patent examiners adding cites to patents (Alcacer and Gittelman 2006)." Ghosh et al. (2016) note that developing measures based on taxonomies suffers from an "implicit assumption that all classes are equidistant even though distance between classes likely varies both between and within industry," a sentiment echoed by several others (e.g., Sampson, 2007; Caner et al., 2017). Kornish and Ulrich (2011) note that studies that envision innovation as search, either in terms of exploration and exploitation (March, 1991) or NK models (Kauffman, 1993; Levinthal, 1997; Koput, 1997; Rivkin and Siggelkow 2003, 2007; Knudsen and Levinthal, 2007), struggle with empirical approaches to characterize the knowledge landscape.

We heed these calls by exploiting 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 (Hoberg and Phillips, 2016; Younge and Kuhn, 2019; Magerman et al., 2015; Arts et al., 2018). For example, Krieger, Li and Papanikolaou (2019) calculate Jaccard indexes to compare the similarity of firms' drug development projects. Most notably, Arts et al. (2018) apply this method to calculate a Jaccard index of similarity between any two USPTO patents or patent portfolios. However, despite its popularity and simplicity, the vector space approach suffers from several drawbacks[4] that

[4] 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.

prompted scholars to explore a more advanced NLP technique – topic modeling – for calculating pair-wise knowledge similarity (e.g., Giorgi and Weber, 2015; 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. We describe our approach in the next section.

The method has the potential to advance research on frontier topics in strategy and innovation, such as the role of heterogeneous firm knowledge in achieving competitive advantage, the impact of knowledge distance and knowledge complementarity on the ex-ante probability of alliances or acquisitions, as well as their ex-post performance. For example, a more accurate measure of knowledge distance could lead to a better understanding of how the knowledge of firms relative to that of competitors influences decisions to engage in mergers, acquisitions, or licensing events. It can also shed light on how the type of knowledge firms have relative to other incumbents influences their competitive strategy and hence facilitates market power or industry fragmentation. Related to this, knowledge distance and movement could be used to track the innovation trajectories of competitor firms and to ask whether firms that compete closely in particular markets benefit more or less by pursuing similar or divergent innovation trajectories or by pursuing a mixed strategy of overlap and distance. The same measures could be used to capture the heterogeneous innovation responses to competitive events, such as firm failure, firm entry, or the arrival of a breakthrough technology.

Our method could also help make progress in research on markets for technologies (Arora and Gambardella, 2010). A main focus of this literature involves understanding how firms' knowledge sets influence the development of technologies to be traded in the market and the success or failure of those transactions. For example, determinants of demand and supply in a market for technologies would depend on firms' knowledge and hence firms' ability to develop the technology internally or successfully integrate it through market transactions, like licensing. Thus, being able to measure distance and movement in knowledge space would benefit studies preoccupied with understanding the conditions under which markets for technologies exist or not, or function well or not.

Our approach could also be useful in a series of questions associated with strategic human capital. Measures of distance and movement in the knowledge landscape can be used to capture the distance between and movement among individuals' and firms' knowledge bases. These can be applied to predict knowledge workers' mobility patterns, the benefits firms might receive by locating near particular knowledge bases, and workers' exit to entrepreneurship.

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. Kaplan and Vakili (2015) were the first to introduce topic modeling to management scholars. Because their paper provides a primer on topic modeling, we assume the basics are known to this audience and focus on topic modeling characteristics that are directly relevant for our purposes.

Topic modeling algorithms are probabilistic models that employ a hierarchical Bayesian analysis of text (see e.g., Hofmann, 1999; Blei et al., 2003; Buntine and Jakulin, 2004). Specifically, topic modeling algorithms analyze the words of documents 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, 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, “a multinomial over a set of words” (Kaplan and Vakili, 2015, pg. 1441). Topics capture words that appear together in the text corpora with a certain probability;[5] 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 and coauthors (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.

[5] 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; Hofmann, 1999; Barde and Bainwad, 2017). This helps address, to a certain extent, issues of synonymity and polysemy.

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 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 that is worth attention 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). Stated somewhat differently, 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.[6] 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.

[6] 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.

We modify the off-the-shelf HDP algorithm to output (1) the top 30 words describing each topic[7], 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 strategy and innovation scholars. 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 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

[7] 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.

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.[8]

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 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; Arts et al., 2018; 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

[8] 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.

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. Author-assigned keywords, or any other set of keywords not pulled from a defined vocabulary, fare better in terms of capturing new knowledge directions, but lack structure, may lack consistency across contributors, and may be more subject to gaming. Similarly, citations as a method for tracing knowledge linkages (DeSolla Price, 1970; Griliches, 1990), is subject to the same concerns since measuring movement via citation maps requires some form of categorization. As is the case with author-assigned keywords, the selection of backward and forward citations is a result of social processes and is subject to strategic behaviors that complicate their interpretation.

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 facilitates 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 loses at least some information about movement in the knowledge space. It is then easy to see how taxonomies, citations, keywords and other similar approaches are less likely to meet all these criteria when compared to our technique.

Lastly, a key advantage of our approach is flexibility. The 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, and business press articles. 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 are just starting to be systematically documented (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 keep in mind that it is not superior to all other approaches of capturing knowledge attributes in all instances. In some cases, fields are relatively well-defined and taxonomies can form the 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, even if limitations remain (e.g., Arts et al., 2018; Righi and Simcoe, 2019). We advise researchers to consider the benefits of our method in the context of their research topic. 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 lamented 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); it is difficult to infer robust patterns when the available information represents the topic of interests sparingly. Second, increasing the amount of data is not sufficient. The data needs to be relevant. A lot of text that captures relatively redundant information about actors' knowledge bases does not support uncovering information about patterns not captured in the data.

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

We provide a step-by-step guide on how to implement our technique. 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 in quantitative studies. Just because several are available, with some based on newer techniques than others, it does not mean the most recent or the most novel approach is the most relevant. 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 with its own advantages and disadvantages, but collectively stronger.

If our technique is deemed appropriate for 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 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. Specifically, we suggest eliminating all stop words and words with only one character, lowercasing all text, and 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).[9]

More advanced vocabulary processing steps could be taken, however, as previously explained, are not crucial if the focus is on obtaining a mapping of the knowledge space as captured in your text document, rather than on interpreting the 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. 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

[9] 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.’ 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.

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).[10] 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-

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

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.

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

Our main goal is to demonstrate that our algorithm performs as claimed. 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, strategy and entrepreneurship 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;[11] 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.

[11] 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."

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.

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; Arts et al., 2018). 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, as discussed, 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 to confirm relatability. To demonstrate, we include two examples of patent abstracts and their 3-digit CPC categorization alongside their assigned HDP topics. Figures 2a and 2b provide the examples. First, 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. Second, in both cases, the HDP-assigned topics intuitively align with the 3-digit CPC classification of the patents.

Step 4: Generate measures

The intended goal for the HDP generated data is to develop covariates for empirical analysis. We choose to focus on a measure of firm-level diversification given the widespread interest in the role of diversification among management scholars. As such, we calculate a yearly measure of diversity of knowledge at the assignee level.[12] We define diversity as the breadth of a

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

knowledge portfolio at each point in time i.e., the set of knowledge topics tackled at a given point in time. 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 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. We chose to focus on the Herfindahl-Hirschman because is a widely used index in strategy, innovation and entrepreneurship research (Cool and Dierickx, 1993; Byun et al., 2018).

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 underestimating 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 NAIC 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 NAICs to 3-digit NAICs).

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 is 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." Of all, 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, research in strategy and innovation has begun to exploit 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 researchers 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 (2020) 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 similar efforts to democratize access to tools and techniques that enable more nuanced measures of innovation activity, such as Arts et al. (2018) that explore patent similarity using a Jaccard index and share the code that enables users of USPTO data (from 1976-2013) to measure the similarity between any two patents or any two groups of patents. Our algorithm enhances such efforts from at least two broad perspectives. First, we apply HDP, a probabilistic topic model that better addresses the issues of dimensionality, synonymy and polysemy, and has a more realistic assumption about the distribution of words. We also include lemmatization, a step that helps the algorithm characterize the relationship between documents more accurately. Second, while efforts such as Arts et al. (2018) focus on applying NLP techniques to measure similarity between text documents, our main objective in this study is to exploit the unique advantage of topic modeling algorithms in uncovering the hidden topics that characterize a corpus of documents. This unique advantage of topic modeling algorithms allows us to make progress in addressing a central yet challenging broad issue in strategy and innovation: measuring changes in knowledge trajectories. Finally, because the approach we describe can be applied to other 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 firms' knowledge bases, it can help develop a more complete picture of a technological landscape. 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.

Adner, Ron and Rahul Kapoor (2010) “Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations,” *Strategic Management Journal*, 31(3), 306–333.

Agarwal, Rajshree, Raj Echambadi, April M. Franco and M. B. Sarkar (2004), “Knowledge transfer through inheritance: spin-out generation, development, and survival,” *Academy of Management Journal*, 47(4), 501–522.

Agarwal, Rajshree, Alfonso Gambardella, and Daniel M. Olson (2016) “Employee mobility and entrepreneurship: A virtual special issue,” *Strategic Management Journal*, 37(13), E11-E21.

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

Ahuja, Gautam, and Elena Novelli (2014) "Mergers and acquisitions and innovation," In *The Oxford Handbook of Innovation Management*

Ahuja, Gautam, Curba Morris Lampert, and Vivek Tandon (2008) "1 moving beyond Schumpeter: management research on the determinants of technological innovation." *Academy of Management annals*, 2(1), 1-98.

Ahuja, Gautam (2000) "Collaboration networks, structural holes, and innovation: A longitudinal study." *Administrative science quarterly*, 45(3), 425-455.

Alcacer, Juan and Joanne Oxley (2014) “Learning by Supplying,” *Strategic Management Journal*, 35 (2): 204–23.

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.

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.

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

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.

Byun, Heejung, Frake, Justin, and Rajshree Agarwal (2018) “Leveraging who you know by what you know: Specialization and returns to relational capital,” *Strategic Management Journal*, 39(7), 1803–1833.

Byun, Heejung, Joseph Raffiee, Martin Ganco (2019) "Discontinuities in the value of relational capital: The effects on employee entrepreneurship and mobility," *Organization Science*, 30(6), 1368-1393.

Carnabuci, G., & Operti, E. (2013). Where do firms’ recombinant capabilities come from? Intra-organizational networks, knowledge,and firms’ ability to innovate through technological recombination. Strategic Management Journal, 34(13), 1591–1613.

Caner, T., Cohen, S. K., & Pil, F. (2017) "Firm heterogeneity in complex problem solving: A knowledge-based look at invention," *Strategic Management Journal*, 38(9), 1791-1811.

Carnabuci, Gianluca, and Elisa Operti (2013) "Where do firms' recombinant capabilities come from? Intraorganizational networks, knowledge, and firms' ability to innovate through technological recombination." *Strategic management journal*, 34(13), 1591-1613.

Cassiman Bruno, Massimo Colombo, Paola Garrone, and Reinhilde Veugelers (2005) “The impact of M&A on the R&D Process: An empirical analysis of the role of technological and market relatedness,” *Research Policy*, 34, 195–220.

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, Ryan Allen, and Michael G. Endres (2021) “Machine learning for pattern discovery in management research,” Strategic Management Journal, 42, 30-57.

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.

Chung, Wilbur and Stephen Yeaple (2008) “International Knowledge Sourcing: Evidence from US Firms Expanding Abroad,” *Strategic Management Journal*, 29 (11): 1207–24.

Cohen, M. Wesley, and Daniel A. Levinthal (1990) “Absorptive capacity: A new perspective on learning and innovation,” *Administrative Science Quarterly*, 35, 128-152.

Cool, Karel and Ingemar Dierickx (1993) "Rivalry, Strategic Groups and Firm Profitability," Strategic Management Journal, 14(1) 47-59.

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.

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.

Fleming, Lee and Olav Sorenson (2004) “Strategic Management Journal,” *Strategic Management Journal*, 25, 909–928.

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.

Gawer, Annabelle, Carmelo Cennamo, and Michael Jacobiddes (2018) “Towards a theory of ecosystems” *Strategic Management Journal* 39(9), 2255-2276.

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.

Giorgi, Simona and Klaus Weber (2015) "Marks of Distinction: Framing and Audience Appreciation in the Context of Investment Advice," *Administrative Science Quarterly,* 60 (2): 333–67.

Griliches, Zvi (1990) "Patent Statistics as Economic Indicators: A Survey," *Journal of Economic Literature*, 28(4): 1661–707.

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.

Helfat Constance E and Marvin B. Lieberman (2002) "The birth of capabilities: Market entry and the importance of pre-history," *Industrial and Corporate Change*, 11(4): 725–760.

Hoang Ha and Frank T. Rothaermel (2010) "Leveraging internal and external experience: exploration, exploitation, and R&D project performance," *Strategic Management Journal,* 31(7), 734–758.

Hoberg, Gerard and Gordon Phillips (2016) "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 124 (5): 1423–65.

Hofmann, 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.

Jiang, Lin, Justin Tan, and Marie Thursby (2011) "Incumbent Firm Invention in Emerging Fields: Evidence from the Semiconductor Industry," *Strategic Management Journal*, 32 (1): 55–75.

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.

Kim JY and HK Steensma (2017) "Employee mobility, spin-outs, and knowledge spill-in: How incumbent firms can learn from new ventures," *Strategic Management Journal,* 38: 1626–1645.

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.

Kneeland. Madeline, M. A. Schilling, and Barak S. Aharonson (2020) "Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation," Organization Science, 31(3), 535-795.

Knudsen, Thorbjørn and Daniel A. Levinthal (2007) "Two Faces of Search: Alternative Generation and Alternative Evaluation," Organization Science, 18(1), 39-54.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, Noah Stoffman (2017) "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics,* 132(2), 665–712.

Koput, Kenneth W (1997) "A chaotic model of innovative search: some answers, many questions." *Organization science*, 8(5), 528-542.

Kornish, Laura J., and Karl T. Ulrich (2011) "Opportunity spaces in innovation: Empirical analysis of large samples of ideas." *Management Science*, 57(1), 107-128.

Krieger, Joshua L., Danielle Li, Dimitris Papanikolaou (2019), "Missing Novelty in Drug Development," NBER Working Paper #24595.

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.

Lu, Jino (2020) "Innovation incentive and diversity of research lines," 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.

Makri, Marianna, Michael A. Hitt, Peter J. Lane (2009) "Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions," 31(6), 602-628.

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.

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

Mindruta, Denisa, Mahka Moeen, and Rajshree Agarwal (2016) "A Two-sided Matching Approach for Partner Selection and Assessing Complementarities in Partners' Attributes in Inter-firm Alliances," *Strategic Management Journal*, 37 (1): 206–31.

Moeen, Mahka (2017) "Entry into nascent industries: disentangling a firm's capability portfolio at the time of investment versus market entry," *Strategic Management Journal*, 38(10), 1986-2004.

Moeen, Makha and Rajshree Agarwal (2017) "Incubation of an industry: Heterogeneous knowledge bases and modes of value capture," *Strategic Management Journal*, 38: 566–587.

Moeen, Mahka, Rajshree Agarwal, and Sonali Shah (2020) "Building Industries by Building Knowledge: Uncertainty Reduction Over Industry Milestones," Strategy Science, 5(3) 218–244.

Moreira, Solon, Arjan Markus and Keld Laursen (2018) "Knowledge diversity and coordination: The effect of intrafirm inventor task networks on absorption speed," *Strategic Management Journal*, 39, 2517–2546.

Nelson, Andrew J. (2009) "Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion," *Research Policy*, 38(6), 994-1005.

Oxley, Joanne E and Rachelle C Sampson (2004) "The Scope and Governance of International R&D Alliances," *Strategic Management Journal*, 25 (8-9): 723–49.

Pakes, Ariel and Griliches, Zvi (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.

Ranganathan, Ram, Anindya Ghosh, and Lori Rosenkopf (2018) "Competition–Cooperation Interplay during Multifirm Technology Coordination: The Effect of Firm Heterogeneity on Conflict and Consensus in a Technology Standards Organization," *Strategic Management Journal*, 39 (12): 3193–3221.

Ranganathan, Ram and Lori Rosenkopf (2014) "Do ties really bind? The effect of knowledge and commercialization networks on opposition to standards," *Academy of Management Journal* 57(2), 515-540.

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.

Rosenkopf Lori and Atul Nerkar (2001) "Beyond Local Search: Boundary-Spanning, Exploration, and Impact in the Optical Disk Industry," *Strategic Management Journal,* 22, 287–306.

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

Sampson, Rachelle C. (2007) "R&D Alliances and Firm Performance: The Impact of Technological Diversity and Alliance Organization on Innovation," *Academy of Management Journal,* 50 (2): 364–86.

Scotchmer, Suzanne (1996) "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.

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

Tandon V., Toh, PK. 2021. "Who Deviates? Technological Opportunities, Career Concern, and Inventor's Distant Search." *Strategic Management Journal*, forthcoming.

Tanriverdi H and Venkatraman N (2005) "Knowledge relatedness and the performance of multibusiness firms," Strategic Management Journal 26(2): 97–119.

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.

Tripsas Mary and Giovanni Gavetti (2000) "Capabilities, cognition, and inertia: evidence from digital imaging," *Strategic Management Journal*, 21, 1147–1161.

Utterback, James M. and Fernando F. Suarez (1993), "Innovation, competition, and industry structure," *Research Policy*, 22(1), 1–21.

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

Villalonga Belen and Anita M. McGahan (2005) "The choice among acquisitions, alliances, and divestitures," *Strategic Management Journal*, 26(13), 1183–1208.

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

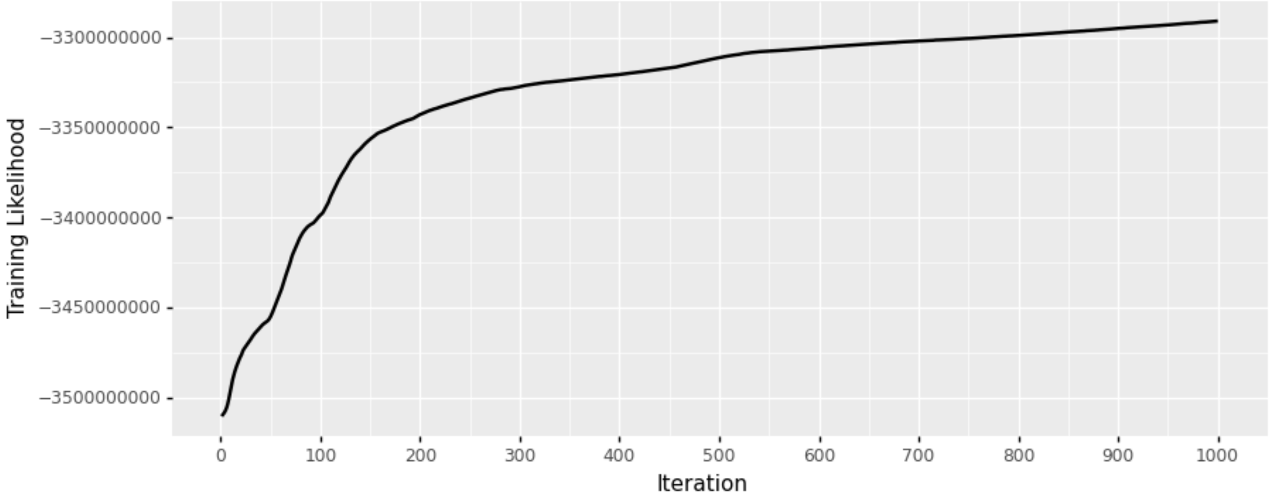


Figure 1b. HDP patent implementation – algorithm-identified number of topics

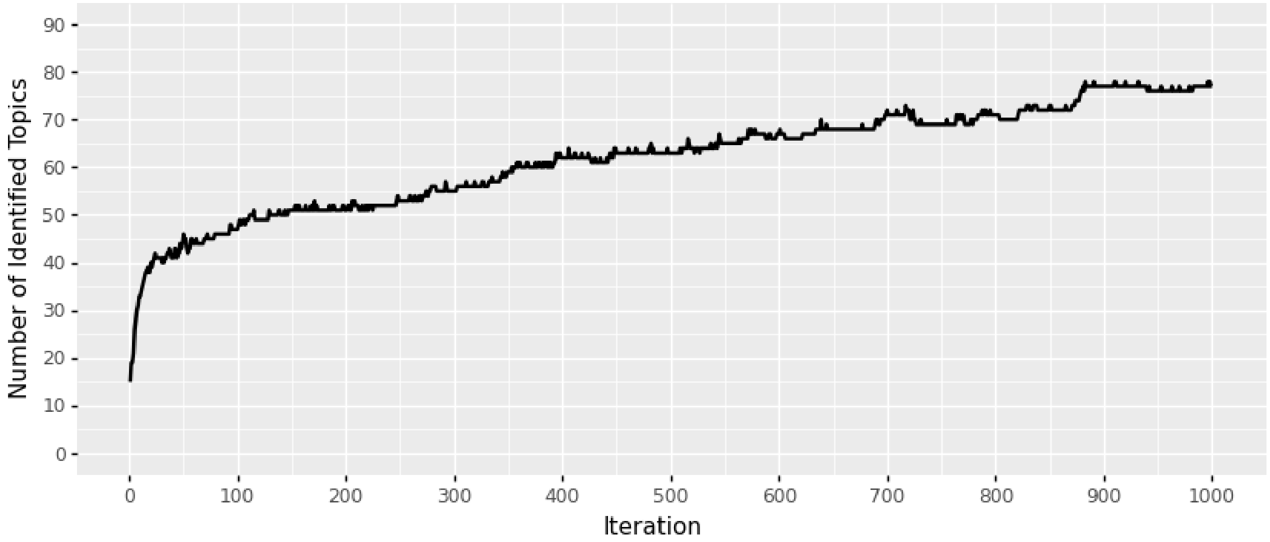
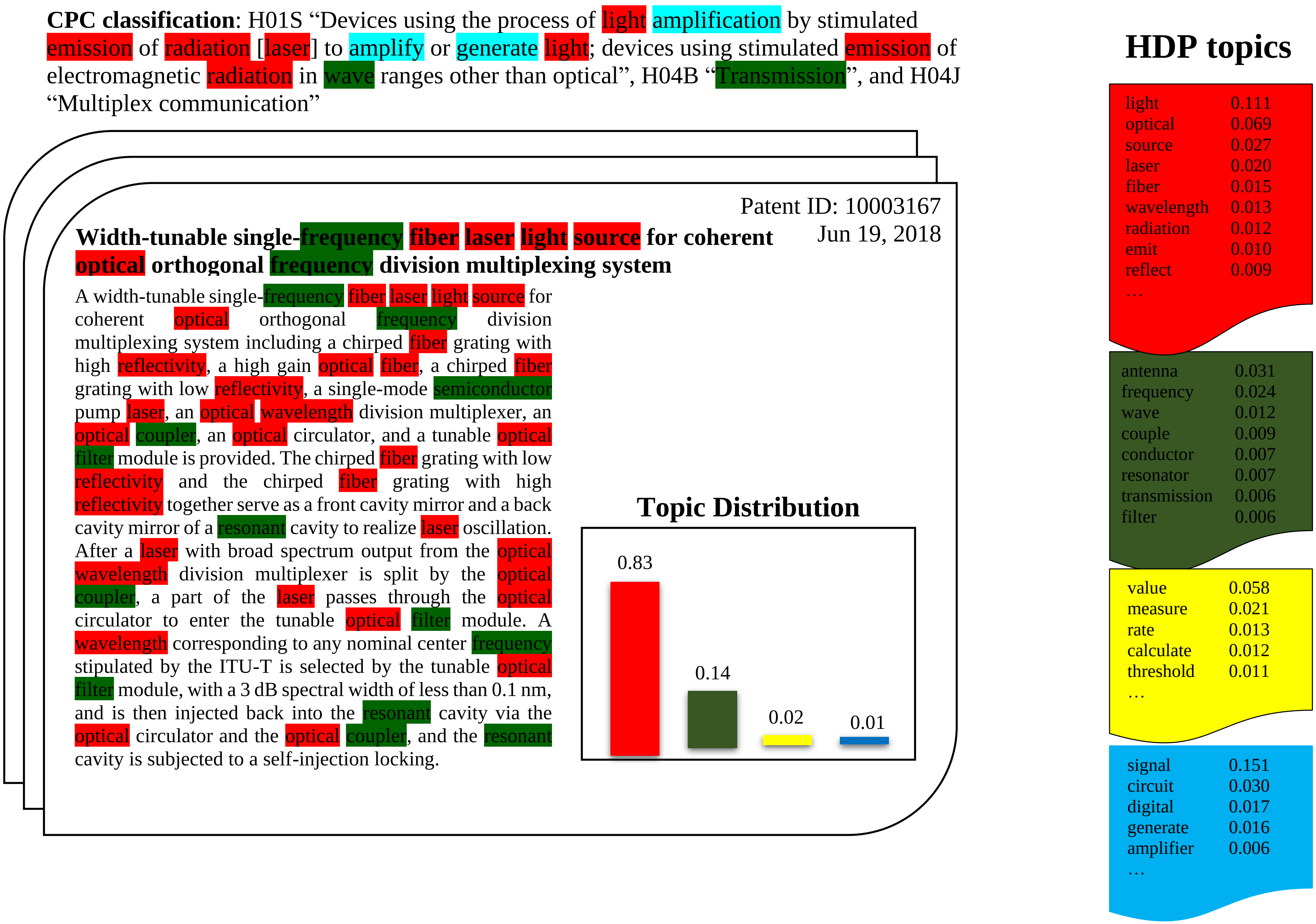


Figure 2a. Example of HDP mapping of a patent abstract



CPC classification: H01S “Devices using the process of light amplification by stimulated emission of radiation [laser] to amplify or generate light; devices using stimulated emission of electromagnetic radiation in wave ranges other than optical”, H04B “Transmission”, and H04J “Multiplex communication”

Patent ID: 10003167
Jun 19, 2018

Width-tunable single-frequency fiber laser light source for coherent optical orthogonal frequency division multiplexing system

A width-tunable single-frequency fiber laser light source for coherent optical orthogonal frequency division multiplexing system including a chirped fiber grating with high reflectivity, a high gain optical fiber, a chirped fiber grating with low reflectivity, a single-mode semiconductor pump laser, an optical wavelength division multiplexer, an optical coupler, an optical circulator, and a tunable optical filter module is provided. The chirped fiber grating with low reflectivity and the chirped fiber grating with high reflectivity together serve as a front cavity mirror and a back cavity mirror of a resonant cavity to realize laser oscillation. After a laser with broad spectrum output from the optical wavelength division multiplexer is split by the optical coupler, a part of the laser passes through the optical circulator to enter the tunable optical filter module. A wavelength corresponding to any nominal center frequency stipulated by the ITU-T is selected by the tunable optical filter module, with a 3 dB spectral width of less than 0.1 nm, and is then injected back into the resonant cavity via the optical circulator and the optical coupler, and the resonant cavity is subjected to a self-injection locking.

Topic Distribution

HDP topics

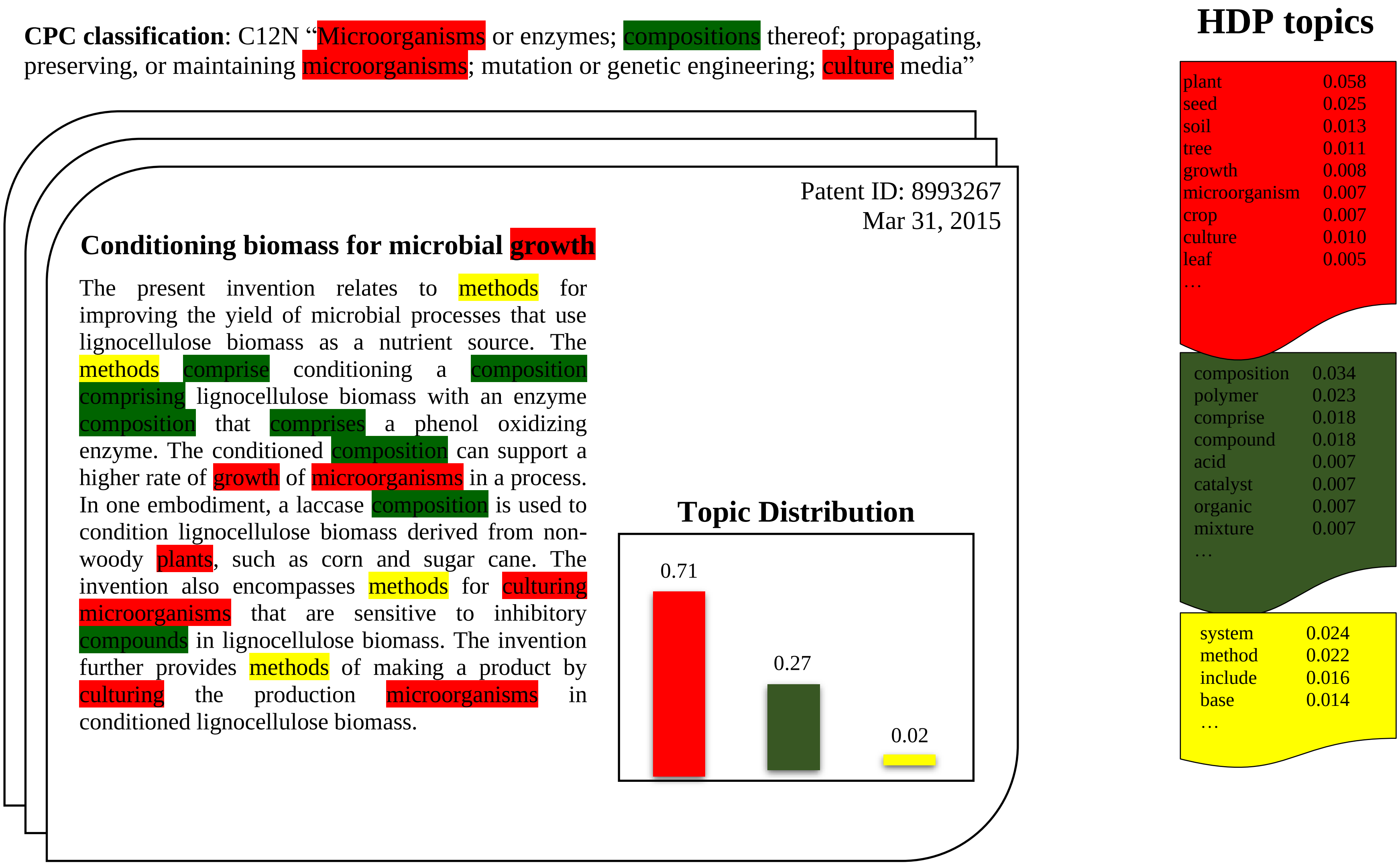
light	0.111
optical	0.069
source	0.027
laser	0.020
fiber	0.015
wavelength	0.013
radiation	0.012
emit	0.010
reflect	0.009
…	

antenna	0.031
frequency	0.024
wave	0.012
couple	0.009
conductor	0.007
resonator	0.007
transmission	0.006
filter	0.006

value	0.058
measure	0.021
rate	0.013
calculate	0.012
threshold	0.011
…	

signal	0.151
circuit	0.030
digital	0.017
generate	0.016
amplifier	0.006
…	

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



CPC classification: C12N “Microorganisms or enzymes; compositions thereof; propagating, preserving, or maintaining microorganisms; mutation or genetic engineering; culture media”

Patent ID: 8993267
Mar 31, 2015

Conditioning biomass for microbial growth

The present invention relates to methods for improving the yield of microbial processes that use lignocellulose biomass as a nutrient source. The methods comprise conditioning a composition comprising lignocellulose biomass with an enzyme composition that comprises a phenol oxidizing enzyme. The conditioned composition can support a higher rate of growth of microorganisms in a process. In one embodiment, a laccase composition is used to condition lignocellulose biomass derived from non-woody plants, such as corn and sugar cane. The invention also encompasses methods for culturing microorganisms that are sensitive to inhibitory compounds in lignocellulose biomass. The invention further provides methods of making a product by culturing the production microorganisms in conditioned lignocellulose biomass.

Topic Distribution

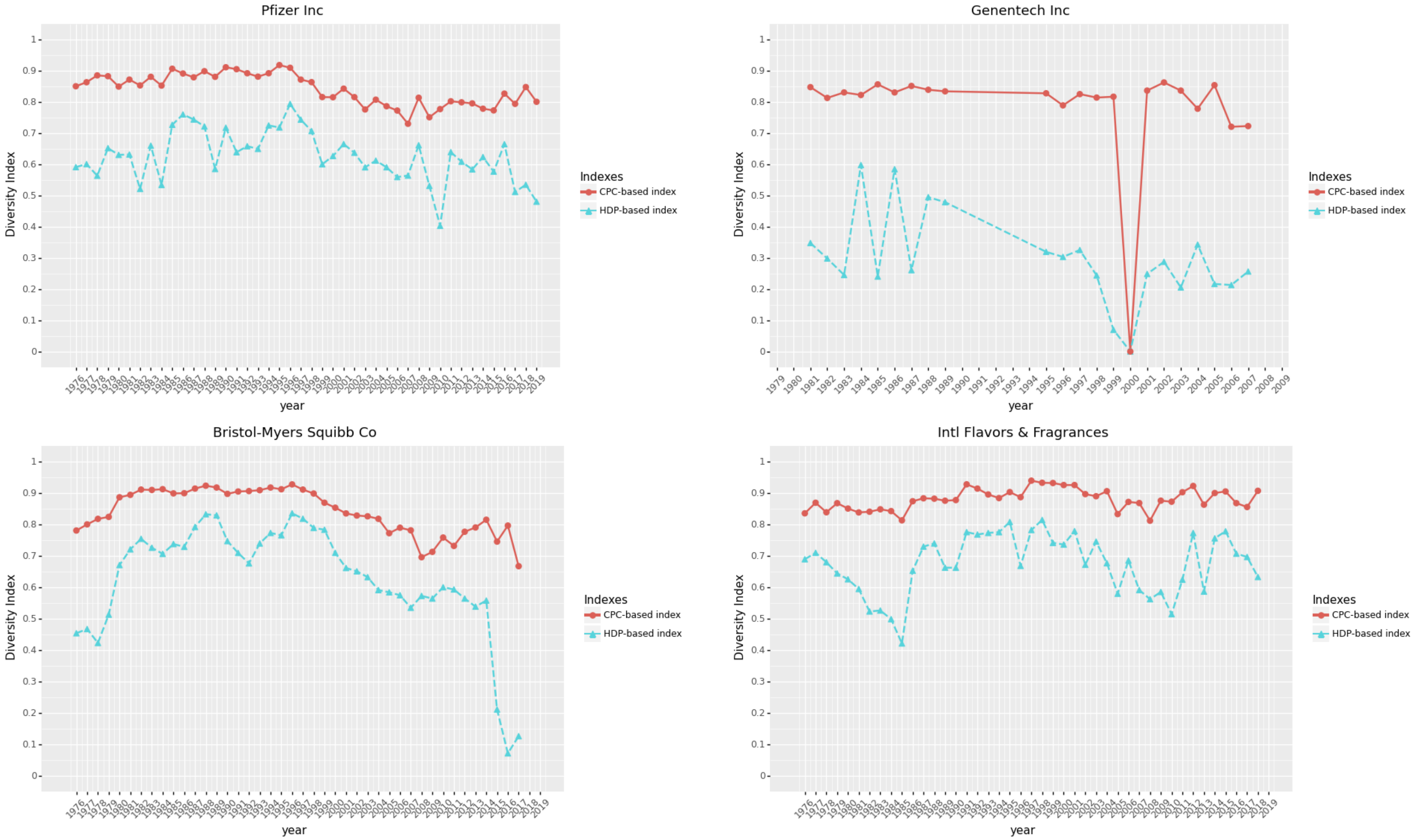
HDP topics

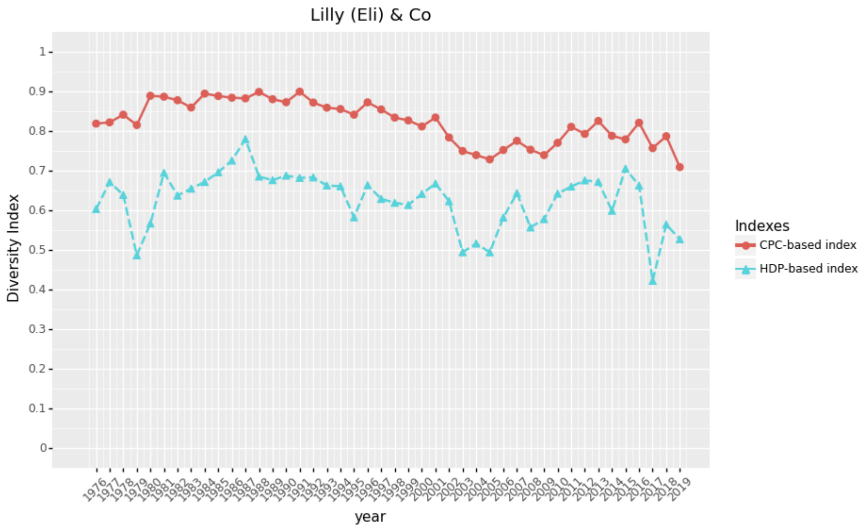
plant	0.058
seed	0.025
soil	0.013
tree	0.011
growth	0.008
microorganism	0.007
crop	0.007
culture	0.010
leaf	0.005
…	

composition	0.034
polymer	0.023
comprise	0.018
compound	0.018
acid	0.007
catalyst	0.007
organic	0.007
mixture	0.007
…	

system	0.024
method	0.022
include	0.016
base	0.014
…	

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




Lilly (Eli) & Co
Diversity Index
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
year
1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Indexes
CPC-based index
HDP-based index

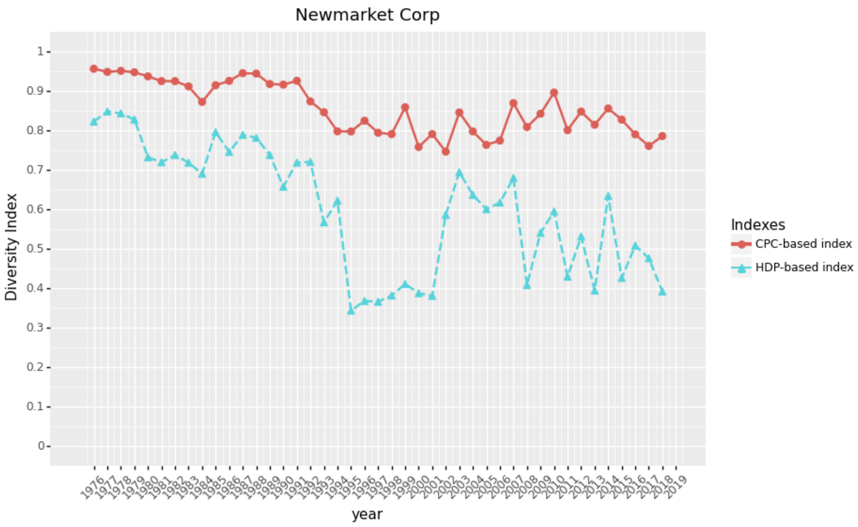
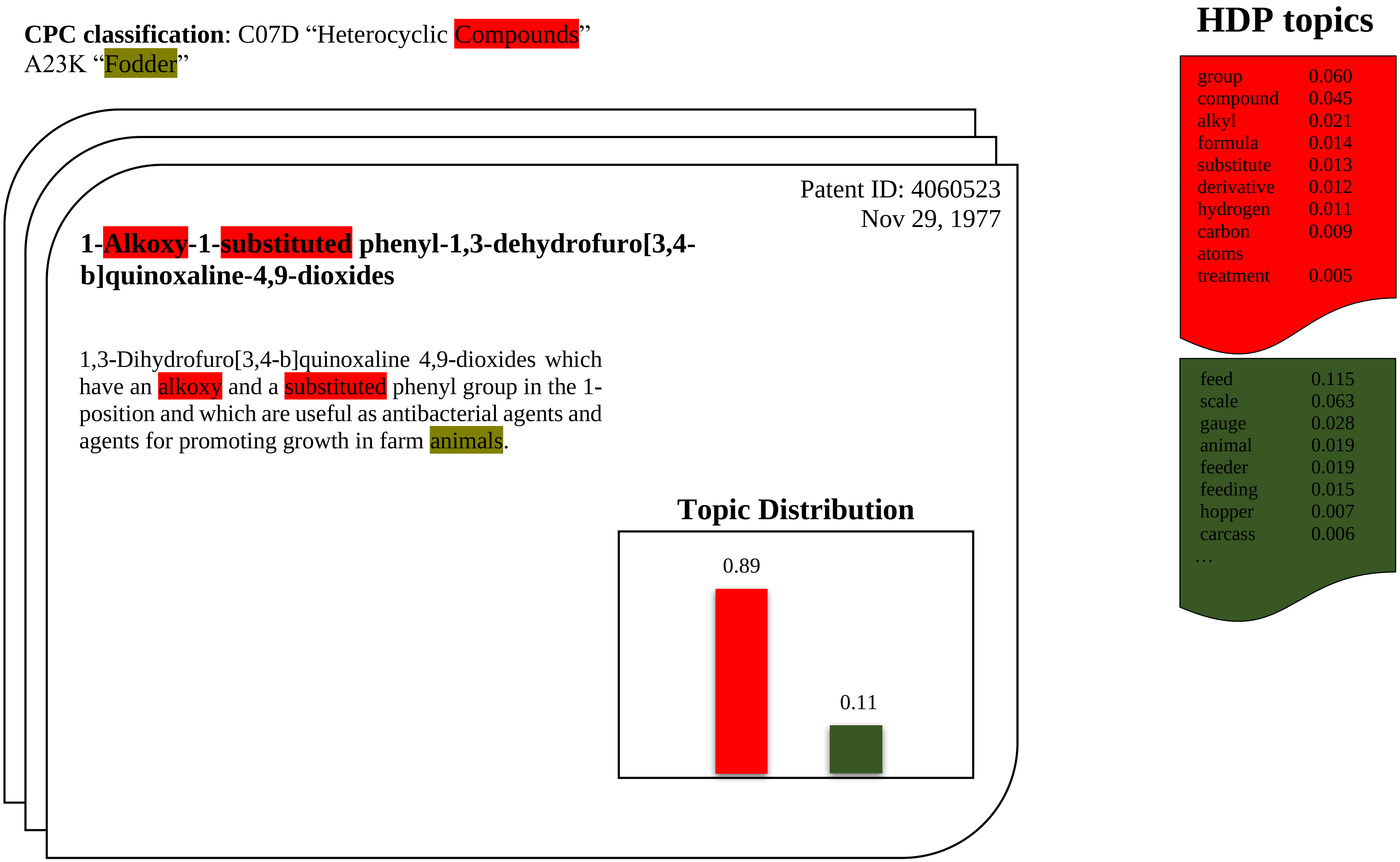

Newmarket Corp
Diversity Index
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
year
1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Indexes
CPC-based index
HDP-based index

Figure 4. Pfizer Inc. patent example



CPC classification: C07D “Heterocyclic Compounds”
A23K “Fodder”

Patent ID: 4060523
Nov 29, 1977

1-Alkoxy-1-substituted phenyl-1,3-dehydrofuro[3,4-b]quinoxaline-4,9-dioxides

1,3-Dihydrofuro[3,4-b]quinoxaline 4,9-dioxides which have an alkoxy and a substituted phenyl group in the 1-position and which are useful as antibacterial agents and agents for promoting growth in farm animals.

HDP topics

group	0.060
compound	0.045
alkyl	0.021
formula	0.014
substitute	0.013
derivative	0.012
hydrogen	0.011
carbon	0.009
atoms	
treatment	0.005

feed	0.115
scale	0.063
gauge	0.028
animal	0.019
feeder	0.019
feeding	0.015
hopper	0.007
carcass	0.006
…	

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

DV= CPC diversification index	**(1)**	**(2)**	**(3)**
	Overall	**Top patenting assignee (90th percentile)**	**Top patenting assignee (99th percentile)**
HDP diversification index	0.353	0.212	0.334
	(0.003/0.000)	(0.003/0.000)	(0.003/0.000)
HDP diversification index * top patenting assignee		0.567	0.806
		(0.005/0.000)	(0.018/0.000)
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/p-value).

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 (0.030/0.000)	0.931 (0.032/0.000)
HDP diversification index x SIC 13 “Oil and Gas Extraction”	-0.015 (0.049/0.762)	
HDP diversification index x SIC 20 “Food and Kindred Products”	-0.056 (0.043/0.197)	
HDP diversification index x SIC 28 “Chemical and Allied Products”	-0.348 (0.030/0.000)	-0.631 (0.028/0.000)
HDP diversification index x SIC 30 “Rubber and Miscellaneous Plastics Products”	0.166 (0.060/0.008)	
HDP diversification index x SIC 33 “Primary Metal Industries”	0.019 (0.059/0.752)	
HDP diversification index x SIC 34 “Fabricated metal Products, Except Machinery and Transportation Equipment”	0.241 (0.055/0.000)	
HDP diversification index x SIC 35 “Industrial and Commercial Machinery and Computer Equipment”	0.371 (0.041/0.000)	0.087 (0.041/0.039)
HDP diversification index x SIC 36 “Electronic and Other Electrical Equipment and Components, Except Computer Equipment”	0.349 (0.036/0.000)	0.066 (0.035/0.064)
HDP diversification index x SIC 37 “Transportation Equipment”	0.407 (0.047/0.000)	
HDP diversification index x SIC 38 “Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks”	0.284 (0.037/0.000)	
HDP diversification index x SIC 48 “Communications”	0.290 (0.051/0.000)	
HDP diversification index x SIC 49 “Electric, Gas, and Sanitary Services”	-0.285 (0.067/0.000)	
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/p-value).

.