Exploration of Science and Technology Interaction: A Case Study on Taxol

Arho Suominen, Samira Ranaei, and Ozgur Dedehayir

Abstract—Linkages between science and technology have been extensively studied using nonpatent literature citations or author-inventor matching. These methods suffer from limitations, such as the lack of citations to relevant documents or challenges with the disambiguation of author–inventor linkages. To mitigate these limitations, this paper uses Latent Dirichlet Allocation to create topic-based linkages between publications and patents based on the semantic content in the documents. The approach allows for the detection of topical overlap between patent and scientific publications, highlighting topical areas shared by research and application. Using a case study on “Taxol,” a cancer drug, with in total 26,475 documents retrieved from EuropePMC database the study illustrates the performance of the approach. The study offers qualitative and quantitative support that the approach is valuable in detecting patent and publication linkages.

Index Terms—Machine learning, science and technology (S&T) interaction, technology management, topic modeling, Taxol.

I. INTRODUCTION

The critical role of science and technology (S&T) interaction in explaining the pattern of economic growth on a macrolevel has been highlighted by a number of scholars [1]–[3]. Equally important has been the systematic analysis of S&T interaction upon the technology management domain, bestowed by increasing production of scientific and technological knowledge that acts as the main source of innovation and competitive advantage [4]. At the same time, theories such as the linear model of knowledge flow from basic science to practical technology (technology-push), coevolution of S&T [5], and the “Triple Helix” framework on “industry-university-government” interaction [6] as well as the concept of “innovation systems” at national or regional levels [7] rely on proxies for scientific [8] and technological development [9], [10].

Against the theoretical models, the development and operationalization of empirical methods to capture the interactions between S&T seems particularly important. Different S&T indicators have been extensively utilized in the technology management domain for analyzing the knowledge flow between scientific research and technological innovation. Common approaches for analyzing the interaction of S&T can be divided into five main categories:

1) simple count of university patents and scientific publication produced by industry [11];
2) author–inventor cooccurrence [12];
3) patent citations to scientific publications [13];
4) scientific papers referring to patent documents [14];
5) citation network analysis [15].

These methods offer different vantage points to uncover the complex S&T linkages. The first two methods address overall S&T related activities of institutions and individual actors, while the latter three patent citation approaches show the connections between patent and publication documents.

Current methodologies for analyzing S&T linkages offer a narrow window for an industry-level analysis. The main challenges facing existing approaches include citation practices, difficulties in name disambiguation, and lack of contextual information. The major limitations of linking S&T via citation data are as follows: first, only one-third of patents include nonpatent literature (NPL) [16] which limits the macrolevel analysis. Second, the concentration of NPL citation information varies across different technological fields [17] and patent offices around the world [18]. Therefore, nonhomogeneous NPL across different technology fields and patent authorities limit the possibility of generalization. The assumption behind citation network analysis is also limited to considering only the documents that are linked via citation. Therefore, documents not associated with any of the citation network components will be discarded. This preclusion logic ignores new or emerging knowledge areas that are not yet cited. The results of a recent survey [19] subsequently conclude that patent citations to scientific papers are not an accurate reflection of scientific contribution to the invention. This is a validity issue central to any operationalization of S&T linkage based on NPL citations. Meanwhile, author–inventor matching is heavily dependent on the accuracy of name disambiguation process, which might be labor intensive work for large data analysis. Even though automated means of name disambiguation have been developed (e.g., [20]) and widely utilized in the patent domain (e.g., [21]), this approach only considers situations where the same entity, inventor, or assignee has been published and patented. Hence, the approach takes an extremely narrow perspective to the interaction of S&T, when we should rather see
that the competence creating production of scientific and technological knowledge can to some extent be detached from the translation to working inventions. The author–inventor matching approach is practical in identifying key individuals, which, while useful for technology management, excludes capabilities of the whole system.

Mapping S&T interaction with empirical methods is an ongoing research trend [22]. While most of the recent studies rely on science related data sources, such as scientific publications [23]–[25] or scientific research proposals [26], [27], only few studies combine patent and publication data sources [15], [28], [29] to create S&T maps using citation information. A recent study by Magerman et al. [30] applied text analytics and semantic approach to study the linkage between both patent and paper. The study [30] evaluates the feasibility of linking patent and paper documents with latent semantic index (LSI) [31], a dimension reduction method based on singular value decomposition. Reporting that LSI does not yield the best results using small datasets, the study additionally applies the LSI clustering method on both patent–paper datasets separately and attempts to detect the S&T link using similarity measures. LSI clustering is based on a word-document matrix, whereby documents are clustered in the same class if their corresponding words are within each others proximity.

Motivated by the limitations of existing approaches, this paper examines the capabilities of a semantic approach for detecting the topical overlap between patent and scientific publications. To this end, this paper uses Latent Dirichlet Allocation (LDA) as an alternative approach to LSI used by [30] to identify the patent–publication linkage based on their topics. More specifically, LDA, introduced by [32], is a probabilistic function at both document and word levels, and extends the Probabilistic LSI method proposed by Hofmann [33] that provides a solid probabilistic foundation at word level alone. Contrary to [30], this study merges patent and publication data sets prior to applying the LDA algorithm, under the assumption that patents and publications are different in nature but that they may share similar features, such as topics. The objective of this investigation is therefore to examine if there are any topical overlaps between patents and publications. This paper concurrently extends the earlier usage of LDA to classify scientific publications [34], science mapping [35], and patent data analysis [36], illustrating its applicability to publications and patents separately. Our main focus in this study is to create topical classifications of a domain to identify document clusters that are similar based on their content, rather than having metadata linkages.

For the purpose of demonstration this study selected “Taxol,” a science driven medicinal invention (i.e., a drug) developed by the pharmaceutical industry to fight breast and ovarian cancer. This paper is structured as follows. Section II reviews previous literature on S&T interaction and the application of machine learning in Scientometrics. Section III then presents a brief history of the Taxol medicine as our case study, as well as the data collection process and methodological approach. Finally, Section IV discusses the results of our investigation, while Section V concludes this paper.

II. BACKGROUND

A. S&T Linkage

Monitoring S&T interaction has attracted significant scholarly attention, which has led to the introduction of a multitude of methods and indicators used to understand and measure the extent of S&T relationships. Early methods focused on counting the number of patents filed by universities [37] or the number of patents filed by university professors [38]. Matching metadata of patent documents and scientific articles has also been used to investigate S&T interaction. The author–inventor matching approach initially practiced by Coward and Franklin in 1989 [39] looks at individuals’ patenting as well as scientific publication activities, and uses these as proxies for S&T interaction. The authors’ examination of the joint development of S&T in the field of semiconductors shows that author–inventor name matching provides more matches than linking based on institutional names. The, to data, most common use of author–inventor networks is to thoroughly understand the knowledge flows [40]. For example, studying the author–inventor collaboration behavior by Chinese researchers, Wang and Guan [41] reveal that collaboration between authors and inventors facilitates the knowledge transfer and plays a key role in the emergence of nanotechnology research in China. An interesting insight acquired from author–inventor network analysis pertaining to nanotechnology is that academic researchers who are simultaneously active in patenting are more productive in terms of publication counts and citations than their academic peers who are not patenting [12]. Other fields of technology where the name matching technique has been practiced includes medicine research [42] and tissue engineering [43].

An alternative technique—coword analysis [44]—matches keywords derived from patent and publication documents, and uses cooccurrences for network mapping. The coword analysis method has been initially practiced based on author assigned keywords listed on the first page. The applicability of this technique has become easier through the use of online databases (e.g., Web of Science and Scopus), which provide helpful features such as keyword-plus (keywords assigned by the database) for conducting coword analysis. Furthermore, some studies [45] take advantage of natural language processing techniques and extract keywords from the title or abstract of patents and the scientific literature as inputs for coword analysis. For example, Moehrle and Caferoglu [46] look at using natural language processing to identify technological speciation within actors. More complex, hybrid approaches, that combine citation network analysis and natural language processing have also been used to detect S&T overlap [15].

The dominant approach used in the analysis of S&T linkage is patent references to NPL (nonpatent literature). NPL refers to any document cited by a patent that is not a patent. This includes scientific articles, books, and conference papers listed on patent documents. The potential of NPL as a measure of S&T linkage has been studied primarily based on Narin’s seminal work [13], which shows patent citations of scientific literature as proxies of knowledge flows from science to practice. Since Narin’s
contribution, NPL citations have emerged as a de facto standard for "science dependency" and S&T linkage and attracted significant scholarly attention. For example, Schmoeh [38] proposed the RNPL (relative NPL) index as a measure, which calculates the science intensity of technological domains. Verbeek et al. [47] meanwhile modeled universal S&T interaction by considering NPL the unit of analysis, and Glänzel and Meyer [14] demonstrated how technological advancement can push science by reviewing scientific paper citations of patent documents. Wagner and Wakeman[48] highlight the market outcomes on patents with NPL, indicating a speed-up in commercialization if patent has science linkages. Heinisch et al. [49] used NPL to show how mobile individuals are significant in transferring knowledge.

Notwithstanding the advancements in these techniques and the variety of analysis carried out, several limitations remain. First, the author–inventor matching methodology is difficult to implement on large-scale datasets. This is because the accuracy of the methodology is highly reliant on the process of author or inventor name disambiguation; name cleaning and spell checking, fixing name changes and variations, distinguishing names of certain cultural origin (e.g., Chinese), and considering middle names, among others. Furthermore, citation network analysis suffers from the fact that only large citation network components will be included in the study, and documents without any citation linkages will be eliminated from the study. The partial loss of data may subsequently have adverse effects on the detection of emerging S&T knowledge and the down side of co-word analysis is that the meaning of words are context dependent, while the reliably of results can be negatively impacted by the “indexer effect” that comes into play when using author-assigned keywords [50].

The use of NPL is not exempt from limitations either. The most significant is that NPL distribution is often skewed in patents, with the majority of patent documents absent of NPL reference [51]. More specifically, roughly one-third of patents include NPL of which only a part of are citations to scientific publications [16]. Moreover, the concentration of NPL varies across different technological fields. For instance, the domains of Chemistry and Pharmaceuticals have the highest numbers of NPL [17], but for other technological areas these numbers can be extremely low, rendering analysis difficult. This concern is emphasized by the variation in legal requirements for including citations among different patent offices around the world. For example, patents registered with the United States Patent and Trademark Office have greater number of NPL in comparison to those registered with the European Patent Office (EPO) [18].

Besides the challenges connected with operationalizing NPL as a science-linkage indicator [52], the definition of NPL additionally seems to be vague in the literature [53]. To reduce this ambiguity, some authors have highlighted the difference between science-based NPL (e.g., patent references to scientific articles) and nonscientific NPL (e.g., patent references to industry reports and manuals). The comparison between the two types of NPL shows a greater number of patents citing nonscientific NPL, thus signaling a shift from basic patenting to applied patenting. A further concern associated with using science-based NPL as an indicator is sourced from the need to disentangle self-cited NPL from nonself cited NPL, particularly as the latter is evaluated as a noisy and inadequate measure of science-linkage [54].

In addition to methodological challenges, author–inventor matching and citations have a conceptual limitation that relates to how we understand the interplay between S&T. While many studies apply science-dependence measures (e.g., NPL), thus, representing a linear direction of science to technology [13], [14], [55], some scholars [5], [56] argue that it is oversimplistic to consider scientific research to be directly translated to technological development. The cognitive relevancy of the NPL content to the patent has also been questioned [5], [57]. Indeed, some scholars propose that scientific knowledge is the source of inspiration for patent documents that do not contain NPL [19]. We can therefore argue that there is an interplay and multifaceted relationship between S&T [5], which requires broader selection of indicators and methods to capture the linkage.

B. Unsupervised Classification With LDA

Machine learning involves computers that act and learn without being explicitly programmed. The unsupervised learning method employed in this paper, namely, LDA, is used to learn and classify latent patterns in a collection of textual documents containing both patents and publications. The method’s objective is to distinguish if two documents of different type have similar knowledge origins. LDA is a topic model, which compared to supervised or reinforced learning algorithms, aims to create an outcome based on a formal framework.

Several studies have hitherto shown the applicability of LDA to classify textual data sourced from scientific publications [34], [35] or patents [36], [58]. For example, Yau et al. [34] have utilized a corpus of scientific publications based on expert opinion to evaluate a number of topic models, and critically evaluated the algorithm capabilities of distinguishing between classes. Departing from the scientometric tradition of journal classification based analysis, Suominen et al. [35] have in turn used LDA to create a map of science, premised on unsupervised learning based classification. Other recent studies include Venugopalan and Rai’s [59] application of LDA to uncover knowledge spillovers, Hu et al.’s [58] use of LDA to create an ontology of knowledge, Lee et al.’s [60], Song and Yuh [61] application of LDA to study technological convergence, and Suominen et al.’s [36] utilization of patent data to analyze company knowledge portfolios. This literature, nevertheless, has seldom merged patent and science publication corpuses to draw out joint latent patterns. The most significant analysis on a shared corpus has been taken in websites and patents [62] and tweets and patents [63].

Given that LDA is an unsupervised algorithm, the analyst has limited interaction and influence upon how documents are classified. This emphasizes the need to understand several basic assumptions prior to analyzing the results produced by LDA. First, documents are exchangeable in the corpus and words are exchangeable in a document. This to say that LDA does not consider the position or proximity of documents in the input, nor the
position or proximity of words embedded within a document as meaningful. This “bag of words,” or documents approach can be considered as a limitation, reduced by either preprocessing (words) or with, for example, dynamic topic modeling that considers a time series of documents. Second, the topic is defined by the probability distribution of words found within the corpus. For instance, high probability words in a particular research topic may be science, research, method, and so on. In addition, we assume that a few top words represent a given topic. Finally, we assume each word has a probability of belonging to each of the topics.

As an input to the analysis, LDA requires the researcher to select the number of topics created by the model. Selecting a practical number of topics has been discussed in the literature. Some researchers have seen that a trial-and-error method of testing a different number of topics with given input data will produce results that are most convenient for human interpretation [64]. Others look for approaches such as using Kullback–Leibler (KL) divergence [65] to estimate the input.

Fundamentally, each document in a corpus “is a random mixture over latent topics, and each latent topic is characterized by a distribution over words” [34]. This is to say that each word has a probability of belonging to a given topic (word–to–topic probability matrix) and similarly, each document has a probability of belonging to a topic (document–to–topic probability matrix). This results in LDA being a soft classification algorithm, where one input document has a probability of belonging to multiple topics. By comparison, other commonly used document clustering methods like k-means [66] is a hard partitioning method, where each document is assigned to only one cluster. Support vector machine (SVM) [67] is another popular document classification approach that can solve two-class pattern recognition problems [68]. LDA can additionally be described as a three layer Bayesian model, whereby learning the distributions for word and document probabilities is a Bayesian inference problem approached by, for example, collapsed Gibbs sampling [69]. In this paper, we use online variational Bayes [70] that enables both accurate and fast classification of even large collections.

C. Visualization of Soft Classification

As LDA is a soft classifier where a document or word does not belong fully to a single topic but has different probabilities of belonging to topics, we use network visualization to further reduce the dimensionality of the data. Rather than looking at individual documents being similar, we use the soft classification to create a network, in which we look for cluster, or communities, of documents.

The result of the LDA can be defined as an undirected bipartite graph. In a bipartite graph, nodes can be divided into two disjoint sets $U$ and $V$. Within the disjoint sets, each edge connects a node in $U$ to one or more in $V$. A bipartite graph by definition is a graph that does not contain odd-length cycles. Focusing on the documents to topic probabilities produced by LDA, we can define $G = (U, V, E)$, where $U$ are documents, $V$ topics, and $E$ document topic probabilities. In turn, focusing specifically on communities and interaction between either sets, we can use network analysis methods to transform the network to monopartite (one-mode) projection and find communities in either the bi- or monopartite graph. Community detection is a well-known network analysis problem, particularly important in being able to understand complex systems, and is supported by a plethora of available algorithms [71]. Some of the algorithms, for example, center on bipartite graphs [72], but for many applications the efficiency of calculation is an important factor [73], especially as the size of the network increases.

III. IMPLEMENTATION OF SEMANTIC APPROACH

A. Case Selection: Taxol

For the purpose of demonstrating the semantic approach to link patents and papers, we have chosen to study the case of Taxol, a medical drug developed to fight ovarian and breast cancer. Our case selection is premised on recent explorations [74], which elaborate on the early stages of drug development, patent activities, and commercialization. The history of Taxol dates back to the U.S. government’s screening program in 1960, which aimed to identify antitumor agents in plants [74]. Over the following years (1960–1964), scientists collected samples from a variety of plants in the Pacific Northwest on behalf of the U.S. Department of Agriculture (USDA), which contained samples of the Pacific Yew tree, the source of Taxol. At the USDA, other scientists demonstrated that the chemical agent, paclitaxel, acquired from the Pacific Yew tree, had a quality of being toxic to cells. This discovery eventually lead to the isolation of the chemical’s purified fraction in 1966, initiating the process of drug development at the same time.

The more than two decades of activity that culminated in the commercialization of Taxol witnessed several noteworthy roadblocks, each requiring resolution to ensure progression. The first of these appeared when the National Cancer Institute (NCI), the purveyor of early drug development, lost interest in the compound. The NCI’s indifference with paclitaxel was fueled by the low concentration of the active compound in the raw material, limited tree supply, and difficulties in the extraction and isolation of the agent. This barrier was nevertheless overcome by championing efforts of individuals such as Dr. M. Wall, as well as Dr. S. Horwitz, who, with colleagues, proved the unique effects and mechanism of the agent against cancerous cells in 1979. The removal of the NCI roadblock invigorated scientific work in the early 1980s, which nonetheless reached a plateau for a few years prior to the emergence of a second surge in 1992 (see below). Dedehayir et al. [74] reveal that the observed loss of momentum was caused particularly by concerns of material supply during drug trials. This material shortage sparked ongoing dialogue among stakeholders, including medical practitioners, politicians, and environmentalists, ultimately triggering further scientific work in search of alternative solutions. The NCI took a central role, targeting the problem of supply through a Cooperative Research and Development Award (CRADA) to motivate competition for higher quantities of supply. In 1991, the award was given to Bristol–Myers Squibb (BMS), a large pharmaceutical firm interested in the drug’s potential. As underscored by Dedehayir et al. [74], this proved to be a watershed event in
accelerating Taxol’s commercialization. After several years of monotonous development, a second surge in research was initiated as community doubts about the drug’s future were put to rest. BMS received approval from the FDA to use Taxol for refractory ovarian cancer in 1992. As revealed by quantitative data [74], patenting activity commenced almost simultaneously as BMS and members of the scientific community protected their intellectual property with expectations of commercialization. However, the core development of Taxol was undertaken in the public domain, with Bristol receiving five-year marketing exclusivity to Taxol [75].

B. Data Collection

To conduct our quantitative analysis, we gathered data from the EuropePMC database (see Fig. 1), containing articles from the life sciences, as well as patents and clinical guidelines, which collectively provided a corpus of roughly 31.9 million records. Using a Python programming language script created by the authors, the database was accessed programmatically and text-mined using an Application Programming Interface (API). The API was used to search for documents with terms taxol (Taxol’s commercial name) or paclitaxel (Taxol’s chemical name). The search returned 26 475 documents. The API was then used to retrieve metainformation on each document such as authors, inventors, publication year, and abstract.

C. Text Preprocessing

Prior to applying probabilistic models, the text of article abstracts was manipulated by removing punctuations, numbers, and stop words (see Fig. 1). The remaining raw text was then tokenized, such that sentences were reduced to single words or phrases (bigrams) using a tokenization technique. All bigrams with a minimum frequency of one have been included in the analysis. Morphological analysis has been done using the built-in lemmatization function from the NLTK Python library, where lemmatization is a process of identifying the word normalized format and removing any suffixes. We opted not to use stemming—another suffix removal method—since the generated terms often cannot be interpreted by the end user. A customized stop word list was additionally prepared to filter out domain specific phrases such as (e.g., this research, this study, and their report presents), or other words (e.g., use, analyze, and explore) that are more representative of content. In the last preprocessing step, words occurring in more than 50% percent of the documents were removed—setting up a 50% threshold is an aggressive preprocessing approach to address the inequality of patent and paper sample sizes. It is crucial to ensure that the algorithm can distinguish between patents and publications, rather than the content of the whole dataset. Furthermore, the documents with short or empty abstract were removed, resulting in the preprocessed dataset containing 23 430 documents. Finally, a dictionary of the filtered words was saved as a “bag of words” and in turn used as an input for the topic modeling algorithm.

D. Topic Modeling and Visual Analysis

Abstracts were analyzed using LDA, a machine-learning model that topically classifies documents based on their content (see Fig. 1). For this task, the preprocessed data were input to the LDA classifier. This process was implemented using a Python programming language, using the package Gensim [76]. Next, we estimated the number of topics to be generated by the algorithm, a requirement of LDA. In this effort, we considered different approaches that have been proposed to evaluate the topic count. We opted to employ the KL divergence to evaluate which could be a practical number of topics, but then used a qualitative technique for evaluating topic volume [77]. Evaluating the KL divergence values from 1 to 100 topics the authors tested a different number of topics, constantly assessing the interpretability and cohesiveness of the word-to-topic distribution. After multiple testing rounds with different topic numbers, the authors set the number of topics at 21.

To run the LDA model, the Gensim package allows for fast and robust analysis of semantic data, while producing a convenient
output for further analysis. Topic modeling with LDA produces two matrices; word probabilities and document probabilities. The word-topic matrix contains a probability distribution for each word that connects to a certain topic. In order to interpret the topics, the word probability distribution will be used to create word-clouds, thereafter used to qualitatively analyze the content of topics. Topic evaluation was done by cancer research experts \( (n = 5) \), who were independently given the task to label topics and to point out terms in the word-cloud that were out of place (if any).

The document-topic matrix, by contrast, denotes soft clustering, which represents a probability distribution for each document to belong to several topics—for example, a document can have large probabilities of belonging to three different topics. The document-topic matrix forms a bipartite network, where nodes are defined by documents and topic. In turn, the bipartite network edges form links between a document and different topics that it has a probability of belonging to. The soft classification approach allows us to see documents as a mixture of content, such as in the case of Taxol embedding information on drug testing and a specific type of cancer. We utilize this soft clustering technique for the network analysis in this study.

Using a Python code, we transformed the document-topic matrix into two flat files (see Fig. 1); an edge file and a node file. The edge file contains information on links between documents and topics, and their probability (weight). The node file, meanwhile, contains the metadata of each node (document and topic). Final analysis was subsequently undertaken for the 21 476 Medline publications and 1954 patent documents (see Table I).

These files were inputted to the network visualization application Gephi [78] for further analysis, following preliminary visual inspection of the bipartite network \( G = (U, V, E) \), where \( U \) are documents, \( V \) topics, and \( E \) document topic probabilities.

In our analysis, we reduced the noise in the dataset by removing extremely small probabilities below 1%. Looking at the bipartite graph descriptives, LDA would produce a complete graph \( G_{21,23430} \), and in the complete graph \( G \) the degree sequence would be \( (x_1, x_2, \ldots, x_{21}), (y_1, y_2, \ldots, y_{23430}) \). Due to the removal of small probabilities, the graph in this study is not complete.

The network was transformed to a monopartite projection \( T \), which contains only the set \( V \) from \( G \). The transformation was done using a matrix multiplication approach, embedded in the Gephi plug-in multimode networks. This approach reduces \( G = (U, V, E) \) to \( T = (V, TE) \), where \( V \) are topics and \( TE \) edges connecting topics. Using the transformation, we are able to analyze the interaction of topics based on how documents share their probability distribution among multiple topics. This exercise essentially draws out interaction between high volume science and patent topics. To identify clusters within both \( G \) and \( T \), we use a modularity optimization algorithm [73] in Gephi to find latent communities.

The validation of the analysis was done by evaluating acquired results from existing S&T linkage methods. Namely, the study focused on using NPL citations to find out if our sample data contained patent to publication linkages. The analysis process searched each patent for all NPL citations, which were then matched to publications in the sample. The matching was done by creating an authors, title, and journal concatenated publication reference for each publication and matching NPL citations to these. Matching was done by first evaluating if an author is present in each and then calculating the cosine similarity for the NPL citation and publication reference. The cosine similarity values were then used to manually select matching publications and patents.

In addition to the quantitative estimation, for a qualitative validation of the results, we acquired data from the EPO developer portal for 20 randomly selected patents with NPLs. These randomly selected patents were analyzed if the NPL citations potentially had science references in the gathered EuropePMC dataset. If so, the probability distribution of the patent document and reference science publication were further analyzed to qualitatively evaluate if the patents that cite science are similarly classified.

IV. RESULTS

Fig. 2 compares the patent and publication records proportion throughout the period of analysis. The proportion of Taxol
related publications are significantly higher than patent applications, which is understandable given the significantly higher volume of scientific publications compared to patenting, in general. The share of patents in the collections is roughly 8.4%, which renders them underrepresented in the sample. The impact of this underrepresentation is reduced by the preprocessing steps that focus on removing patent or publication specific terms and phrases from each document.

At the same time, we observe that publication documents (red line) lead patenting activity (black line) with respect to increasing–decreasing trends. This time lag between the two trend lines is commensurate with Suominen and Seppänen [79] findings, which show that patent and publication data grow simultaneously. The time series presented in Fig. 2 also allow us to identify possible watershed events in the development of the drug, as marked by inflection points or sudden upward–downward turns. Publication documents citing Taxol started to increase in number in 1980 until approximately 1985. This spurt was likely fueled by championing actions (e.g., [80], [81]), which served to alleviate the skepticism of key stakeholders such as the NCI, who made vital decisions on whether to continue or kill such projects [81], [82]. The removal of the NCI roadblock thus resulted in a surge in scientific work in the early 1980s. The period stretching from 1985 to 1991 depicts a plateau in scientific research and publications. We believe that this loss of momentum was caused by several significant problems. First there were concerns over material supply during clinical phases I and II trials (conducted between 1983 and 1986), which limited and also demotivated investment in scientific work. Furthermore, poor stability, hypersensitivity, and toxicity of the drug were deemed to cause safety issues and the abandonment of many trials. Meanwhile, the supply problem reappeared between 1987 and 1989, when the observed success of the drug against ovarian cancer created a buzz within the scientific community and the general public, albeit with a cause for concern as to the availability of stock and raw materials [80]. Ultimately, the second surge in scientific article publications that began in 1991 and lasted until 2005, was sparked by the awarding pharmaceutical company, Bristol-Myers Squibb, with the CRADA [83], [84]. After some years of stifled progress, this surge in research was initiated as the scientific community’s doubts about the drug’s future were put to rest. Following the logic of the linear model of innovation, the patenting trend is slower to show activity, with notable growth arriving in 1995. We interpret this observation to suggest that strong increases in scientific research influenced industry R&D resulting in a strong wave of patenting. It should also be noted that the increases in Taxol related patents coincide with the expiration of core patents and subsequent litigation.

| Topic          | Sum of probabilities | Ratio (%) | Mod class | Topic labels                        |
|----------------|----------------------|-----------|-----------|-------------------------------------|
|                | Publications | Patents | Publications % | Patents % |                                    |
| Topic 1        | 958.37       | 210.52  | 4.5 %      | 11.3 %    | 3          | Chemotherapy                        |
| Topic 2        | 751.54       | 297.46  | 3.6 %      | 15.9 %    | 3          | Pre-clinical testing                |
| Topic 3        | 417.19       | 297.30  | 2.0 %      | 15.9 %    | 7          | Intercellular transport             |
| Topic 4        | 1435.51      | 51.39   | 6.8 %      | 2.7 %     | 6          | Paclitaxel in breast and ovarian cancer |
| Topic 5        | 740.51       | 20.97   | 3.5 %      | 11.1 %    | 5          | Cancer diagnosis                   |
| Topic 6        | 371.82       | 18.67   | 1.8 %      | 10.0 %    | 5          | Drug metabolism                    |
| Topic 7        | 721.72       | 38.38   | 3.4 %      | 21.1 %    | 1          | Tumor progression                  |
| Topic 8        | 349.79       | 18.49   | 1.7 %      | 10.0 %    | 4          | Paclitaxel mechanism of action      |
| Topic 9        | 892.42       | 32.69   | 4.2 %      | 17.7 %    | 2          | Paclitaxel synthesis               |
| Topic 10       | 2188.44      | 73.76   | 10.3 %     | 3.9 %     | 2          | Cell response                      |
| Topic 11       | 2211.21      | 65.30   | 10.5 %     | 3.5 %     | 4          | Resistance of treatment            |
| Topic 12       | 1367.92      | 26.15   | 6.5 %      | 14.4 %    | 5          | Combination therapy                |
| Topic 13       | 1041.24      | 355.93  | 4.9 %      | 19.0 %    | 3          | Treatment regimen                  |
| Topic 14       | 770.85       | 41.45   | 3.6 %      | 22.2 %    | 5          | Neural adveres effects of paclitaxel |
| Topic 15       | 1201.16      | 82.41   | 5.7 %      | 4.4 %     | 6          | Treatment of ovarian cancer         |
| Topic 16       | 298.38       | 29.78   | 1.4 %      | 1.6 %     | 1          | Paclitaxel effects on intracellular vesicular traffic |
| Topic 17       | 782.76       | 33.59   | 3.7 %      | 1.8 %     | 1          | Paclitaxel effects on the cytoskeleton and its consequences |
| Topic 18       | 761.99       | 18.70   | 3.6 %      | 1.0 %     | 1          | Molecular determinants of cancer response to treatment |
| Topic 19       | 832.92       | 39.83   | 3.9 %      | 21.1 %    | 5          | Anti-cancer drug combinations       |
| Topic 20       | 1087.92      | 83.16   | 5.1 %      | 4.4 %     | 7          | Local/topical paclitaxel drug delivery systems |
| Topic 21       | 1963.61      | 35.27   | 9.3 %      | 1.9 %     | 2          | Identification of paclitaxel as an anti-cancer agent |
Table II illustrates the soft-partitioning of documents (patent documents and Medline scientific literature) into 21 topics. Results shows that all topics contain a mixture of both document types, but with different propositions. The first two columns of Table II show the sum of probabilities of documents associated with each of the 21 topics. The association is based on the probability value of a document classified under corresponding topics.

The next two columns of Table II contain the percentage ratio of documents divided by the total number of patent documents as well as Medline documents. According to the table, a high percentage of patents—15.9%, 15.9%, and 19%—are classified under Topics 2, 3, and 13, respectively. It can be observed that the treatment of particular cancers (e.g., brain, ovarian, and breast cancer) using Taxol medicine is the focus of majority of patent documents. On the science side, we can see that scientific publications are high in Topic number 10, 11, and 21, where the topics are more about general tumor and cancer treatment processes.

The last column of Table II provides the topic’s thematic labels generated by the authors, based on consultation with five cancer research experts. The consultation was done using an email questionnaire asking the experts to label the word-clouds. Consensus among the evaluation has been reached by the authors relying on knowledge from previous qualitative studies on Taxol [74], [85], [86]. In practice, topic labeling is based on a qualitative evaluation of the word-to-topic probability matrix by experts visualized as a word-cloud based on the word-to-topic matrix (see Fig. 3). The experts were each given a word-cloud individually to label. The authors thereafter created a consensus label based on their background knowledge. In Fig. 3, Topic 1 is represented with highly frequent keywords like “chemotherapy, treatment, lung cancer” and labeled as Chemotherapy. The prominent appearance of keywords (“tumor, mouse, human, in vivo, in vitro”) connects the second topic to a particular kind of preclinical testing. In contrast, Topic 3 focuses on a more detailed area, intercellular transport, the movement of vesicles and substances within the cell. Topic 4 (with top keywords of “expression, ovarian cancer, breast cancer”) concerns the use of paclitaxel in treatment of ovarian and breast cancer, while Topic 5 focuses on a clinical process of cancer diagnosis. The
distinguished keywords of Topic 6 are “alkyl, treatment” highlighting drug metabolism. The focus of Topic 7 is again in the clinical process with understanding tumor progression, as the words “cancer, tumor,” and “growth” are highlighted in the word-cloud. The top keywords for Topic 8 are “microtubule” and “tubulin,” thus connected with information on paclitaxel’s mechanism of action. Topic 9 contains keywords such as “drug, compound, agent” associated with the Taxol drug preparation particularly synthesis. Topic 10 is represented by top keywords such as “protein, microtubule, kinesin, spindle.” This topic is related to a research stream of evaluating cell response. Topic 11 focuses on the resistance to treatment, with terms such as “apoptosis,” “programmed cell death,” “expression” and “protein.” In Topic 12, high probability terms such as “drug” and “concentration” labeled the topic as a combination therapy, an approach that uses more than one medication.

In Fig. 4, Topic 13, labeled as treatment regimen, is linked to the clinical process, similar to Topic 12. Having only a few high probability terms, Topic 14 was labeled as the “neural adverse effects of paclitaxel.” Topic 15, similar to Topic 4, focuses on “treatment” and “ovarian cancer” and is labeled as “Treatment of ovarian cancer in patients.” Topic 16 is labeled “Paclitaxel effects on intracellular vesicular traffic” and is based on terms such as “TNF alpha,” a cell signaling protein regulating immune cell, “vesicle,” a large structure within a cell. Topic 17 is similarly labeled as “Paclitaxel effects” but now “on the cytoskeleton and its consequences.” Similar to Topic 12, Topic 19 focuses on “Anticancer drug combinations” with terms such as “docetaxel, combination,” and “treatment.” Topic 20 focuses on drug delivery and is labeled “Local/topical paclitaxel drug delivery systems.” Finally, Topic 21 is labeled as “Identification of paclitaxel as an anticancer agent” with terms such as “plant, concentration,” and “drug” explaining the topic content.

A. S&T Overlap

We used a network approach to analyze the thematic overlaps of patents and publications in each topic. In this effort, soft classification output has been utilized to illustrate the network of documents and generated topics. Table III contains the information about nodes and edges of generated networks, while Fig. 5 shows the network view of full bipartite network $G$, where documents and topics represent nodes and are connected based on the document probability distributions. Fig. 6 represents a filtered bipartite network based on only patent documents. In both figures the nodes and topics are colored using the Modularity algorithm [73], which resulted in seven modularity classes tabulated in Table I. The monopartite projection graph $T$ is also visualized in Fig. 7, where the links between topics can be observed. The layout in these graphs is based on Open Ord [87].
The bipartite network in Fig. 5 shows 23,451 nodes, where 51 are topics and 23,430 are Medline and patent documents. The graph visualization is partial as small probabilities are omitted, with 144,106 edges. The average weighted degree for the full network is 1.96 ($s = 36.61, N = 23,451$), while for Medline documents it is 0.98 ($s = 0.0098, N = 21,476$) and for patents 0.96 ($s = 0.028, N = 1,954$). This highlights that patents are more concentrated to a few topics, while publications can share probabilities more broadly. For the projected graph in Fig. 7, the average degree is 20 with an average weighted degree of 716.769. Even though the bipartite network is not a complete graph, each of the projected topic nodes are linked—the strongest link is between Topic 21 and Topic 10 (weight 360.358).

The visual inspection of the two networks (see Figs. 5 and 6) highlights the different positions of publications and patents in the graph. However, there is a significant overlap between patent and publication in Modularity classes 3 (Topics 1, 2, and 3), Modularity class 7 (Topics 3 and 20), and Modularity class 6 (Topics 4 and 15). This is of particular interest, as these areas have been quantitatively labeled as ones where overlap is expected. It seems the major patenting activities in the Taxol domain is about drug development and its chemical compound. On the contrary, Topic 4 shows little topical overlap between science and patent documents. The underlying content of Topic 4 corresponds with the use of Taxol in the care of breast and ovarian cancer and might be irrelevant to patenting.

### B. Evaluation of Topical Overlaps Between S&T

To evaluate the topical overlap quantitatively, the connection between NPLs and Medline publications in our dataset is examined. First, the NPL list is retrieved from the EPO database using 2189 patent applications, although only 378 out of 2189 patent documents (about 17%) contained NPLs in our Taxol dataset.
The format of retrieved NPLs is incomplete and not unified. The standard format of a reference usually contains title, authors names, source of publication, year, volume, number, and page number. However, more than half of the NPL list does not contain title of publication. Using a cosine similarity measure, the similarity between the NPL list and Medline publications is calculated. The cosine similarity script in Python first searches for title similarity if the title exists. Then, it moves to match author names and finally matches the source of publication. The calculation process resulted in 150 matches based on a high cosine similarity value (higher than 0.7). The sample of 150 is significant enough to assess overlap resulting through LDA.

Then, the distance between each matched patent and cited Medline publication, based on their topic probabilities, was calculated using Hellinger distance (see Table IV). A similar analysis was done for 150 randomly selected paired documents in the sample. For matched patents and publications via NPL, the average distance is 0.77, and for the randomly selected paired documents the average distance is 0.86. This shows that matched documents via NPL have a smaller distance than randomly selected documents. In other words, matched document pairs via NPL are conceptually closer to each other.

For the purpose of describing the overlap qualitatively, a random sample of nine U.S. patents that contain NPL to scientific articles in the Taxol dataset were qualitatively analyzed. The objective was to detect if the patent-papers that are linked based on NPL are correctly classified in the same topical category. The sample is partially illustrated in Table V. Out of the sample patents, almost half of the pairs were classified under the same Modularity class. The patent-paper pairs that were not classified in similar modularity classes were mostly due to the cited NPL not being at the core of the invention. For patent-paper pair number 2, the cited publications were mentioned early in the description as a probable application area for the invention. For patent-paper pair number 4, the publication seems to form a clear background for the invention. The publication has been cited as “Taxol has also been tested in certain ophthalmological applications” and the patent proposes “The method of preventing opacification after extracapsular cataract extraction.”

The reason of the publication and patent not being classified
similarly most probably lies with the patent having an extremely short abstract.

The remaining patents were also clustered in similar topics, but with lower probability rates. For instance, patent document US5850032 is clustered under Topic 2 with the probability of 22% similar to its publication pair at 19.2%, respectively (see Table V). We can argue that the document pair is correctly classified based on their semantic similarities. A challenge is US5310672, which only shares Topic 12 and the publication remains weakly associated with it. Interestingly enough the publications, while given as NPL citations were not cited patent description text. Overall, our findings suggest that patents and publications associated with NPL reference are more closely classified, using Hellinger distance, but at a qualitative level, we can question whether NPL matches always give an indication of true transfer of knowledge. Collectively, these findings suggest...
## TABLE V
EXAMPLE OF NINE PATENTS WITH NPL CITATIONS

| #   | Title                                                                 | Topic (Probability) | Mod |
|-----|----------------------------------------------------------------------|---------------------|-----|
| 1   | US$965718 - Analogs of sarcodictyin and eleutherobin                  | 1 (21.4 %); 3 (24.0 %); 5 (4.2 %); 13 (31.2 %); 14 (8.0 %); 20 (8.7 %) | 2   |
|     | Eleutherobin, a novel cytotoxic agent that induces tubulin polymerization, is similar to paclitaxel (Taxol). | 1 (32.6 %); 5 (11.4 %); 6 (5.0 %); 8 (2.6 %); 13 (36.7 %); 14 (5.8 %); 16 (4.5 %) | 2   |
| 2   | US$6080733 - Thioureido-cyclodextrins, utilized in particular to solubilize anti-tumor, and antiparasitic agents and their preparation processes | 1 (25.0 %); 4 (17.0 %); 10 (6.9 %); 13 (39.1 %); 16 (7.3 %) | 2   |
|     | Experimental antitumor activity of taxotere (RP 56976, NSC 628503), a taxol analogue. | 1 (7.7 %); 2 (7.9 %); 3 (16.3 %); 4 (5.8 %); 7 (11.7 %); 10 (11.3 %); 15 (6.4 %); 20 (31.7 %) | 6   |
| 3   | US$231600 - Stents with hybrid coating for medical devices          | 2 (49.1 %); 4 (6.1 %); 8 (12.3 %); 14 (13.0 %); 14 (18.0 %) | 2   |
|     | Regression of collagen-induced arthritis with taxol, a microtubule stabilizer. | 1 (2.5 %); 7 (6.6 %); 19 (10.6 %); 20 (5.5 %); 13 (2.4 %); 15 (36.7 %); 20 (4.7 %); 21 (29.8 %) | 5   |
|     | In vitro and in vivo antitumoral activity of free, and encapsulated taxol. | 1 (54.3 %); 7 (15.2 %); 19 (3.8 %); 20 (24.0 %) | 2   |
| 4   | US$376345 - Method and means for inhibiting posterior capsule opacification | 2 (82.3 %); 18 (10.2 %) | 2   |
|     | Taxol treatment of experimental proliferative vitreo-retinopathy    | 2 (3.6 %); 11 (3.7 %); 12 (39.9 %); 19 (3.4 %); 20 (47.6 %) | 6   |
| 5   | US$310672 - Induction of somatic embryogenesis in taxus, and the production of taxane-ring containing alkaloids therefrom | 22 (25.4 %); 5 (14.5 %); 8 (26.6 %); 12 (24.5 %) | 3   |
|     | Cell culture of Taxus as a source of the antineoplastic drug taxol and related taxanes. | 1 (55.3 %); 2 (19.2 %); 12 (3.7 %); 13 (13.4 %); 14 (3.0 %); 17 (3.9 %) | 2   |
| 6   | US$665576 - Callus cell induction and the preparation of taxanes     | 1 (93.7 %) | 2   |
|     | Cell culture of Taxus as a source of the antineoplastic drug taxol and related taxanes. | 1 (55.3 %); 2 (19.2 %); 12 (3.7 %); 13 (13.4 %); 14 (3.0 %); 17 (3.9 %) | 2   |
| 7   | US$336684 - Oxidation products of cephalomannine                     | 3 (29.7 %); 11 (34.4 %); 13 (15.4 %); 14 (15.2 %) | 3   |
|     | Synthesis of biologically active taxol analogues with modified phenylisoserine side chains. | 2 (4.0 %); 11 (11.4 %); 12 (12.6 %); 13 (37.4 %); 19 (21.5 %); 20 (11.1 %) | 2   |
| 8   | US$450032 - Method for production of plant biological products in precocious neomorphic embryos | 2 (22.0 %); 5 (7.7 %); 12 (12.6 %); 19 (9.1 %); 20 (43.5 %) | 6   |
|     | Cell culture of Taxus as a source of the antineoplastic drug taxol and related taxanes. | 1 (55.3 %); 2 (19.2 %); 12 (3.7 %); 13 (13.4 %); 14 (3.0 %); 17 (3.9 %) | 2   |
| 9   | US$258121 - Stent coating                                           | 2 (59.4 %); 8 (3.8 %); 10 (17.2 %); 13 (16.9 %) | 2   |
|     | In vitro and in vivo antitumoral activity of free, and encapsulated taxol. | 2 (54.4 %); 7 (15.2 %); 19 (3.8 %); 20 (24.0 %) | 2   |

Rows are patents, with cited publication nested below.
that graph number 5 can be used to find patent and publication similarities.

C. Linkage Between S&T Topics

In an attempt to group the topics based on their conceptual proximity, we have reduced the bipartite network to one monopartite projection (see Fig. 7), which divides the topic network into communities based on the strength of relationships between topics. This graph gives an understanding of the size of different Topics and how the Topics form latent structures. For example, Topics 10, 21, and 9 are highly publication dominated areas, focusing on the discovery of Taxol as an anticancer agent. On the other hand, the cluster formed by Topics 1, 13, and 2 focuses on a clinical treatment phase, where patents take up a larger share of the documents.

An interesting finding from the figure is the large Topic 11 addressing resistance to treatment. A highly publication driven area, this Topic focuses on understanding the mechanism of Taxol and the impact of treatment.

V. Conclusion

The literature provides a diverse set of indicators and methodologies that have been developed and utilized for the detection of S&T interactions. Majority of the methods stem from the Scientometric research field, where many studies use NPL, citation network analyses, word cooccurrences, and author-inventor analyses to match the patent-paper pairs. Less attention has been paid, however, to the semantic content of documents.

This paper examined the applicability of an unsupervised learning algorithm on the detection of S&T linkage. The pharmaceutical industry was selected as the empirical setting for such application, focusing specifically on Taxol, a cancer treatment medicine. The selection of this setting was motivated by the science-driven nature of the field, thus expecting a strong NPL citation linkages to emerge. This is of particulate importance as we cannot expect NPL citations to be broadly present across domains—a problem this paper tries to contribute to. The topic modeling results showed a set of 21 topics occupied by both document types of patents and publications. The proportion of scientific publications was an order of magnitude greater than patents between 1980 and 2015, which meant that topics were higher in number of publications as well.

To visualize the overlap between documents, a bipartite network of document and topics was generated using the Gephi software. By filtering the network based on document types, the overlap between the patent and paper sets was observed. Furthermore, the interaction between topics was illustrated by projecting the clusters onto a monopartite network. In the study, NPL citations and a qualitative analysis has been performed to validate the performance of the algorithm in clustering similar documents. Using distance measures for quantitative analysis, we showed that citations matched patents, and that publications were classified closer than random documents. In addition, almost half of the patent-paper that were selected for qualitative evaluation were classified under the same Topics with the highest probability values. The remaining document pairs were also classified into at least one similar Topic with lower probability values.

Our findings propose a new approach to examine S&T linkages. Claiming no superiority over existing methods, we argue that the unsupervised classification approach used in this study allows for a practical method to chart new technology areas, where both publications and patents play a significant role in the landscape. This is very much the case in, for example, pharmaceuticals. The process proposed here requires relatively little preanalysis information from the researchers, but creates insights through a visual representation of the thematic overlap. The primary advantage of the method is its ability to overcome the lack of explicit linkages (such as citations) and challenges in identifying author-inventors. The best insights are nevertheless created by merging different approaches.

This paper is not without limitations, which might inspire future research. First, even though our study introduces a promising machine learning approach in detecting S&T linkages, the method has been applied to a single science-driven invention. It is worth examining the applicability of these advanced methods on multiple case studies to strengthen our confidence in the method. The validation of clustering accuracy was limited with only 17% of patent documents that contained (NPLs) references to scientific articles. The evaluation process can be improved further when the clustering precision and recall can be compared with human expert knowledge.

This paper has implications for the academic community through the methodology it has demonstrated to measure the science–technology linkage. S&T research has been limited by the low number of NPL linkages and the challenges of name disambiguation. The proposed LDA approach allows researchers to complement the limitations of existing methods. In particular, the approach allows us to look for implicit linkages where, for example, university knowledge creation is implemented in industry and subsequently patented. Rather than basing the analysis on a technology as in this paper, the approach allows us to study regional or national knowledge production and epistemic changes in S&T.

For practitioners, our results extend traditional means of analyzing intellectual capital of organizations [88], and the technological opportunities that exist within industries [89]. By identifying implicit science–technology linkages, companies can extend existing technology maps [90] to contain scientific knowledge and patent portfolios. This allows practitioners to create one process to create a competitive intelligence landscape [91]. Policy makers will also be more attuned to making the most beneficial investments toward strengthening university–industry collaboration [92]. Building a mapping at the level of an innovation system allows for a holistic view and policies on developing the Triple–Helix interactions [6].

REFERENCES

[1] R. R. Nelson and S. G. Winter, An Evolutionary Theory of Economic Change, vol. 93, no. 2. Cheltenham, U.K.: Edward Elgar Publishing, 2000.
[2] G. Dosi, Innovation, Organization and Economic Dynamics. Cheltenham, U.K.: Edward Elgar Publishing, 2000.
[3] C. Freeman, “The economics of technical change,” Cambridge J. Econ., vol. 18, no. 5, pp. 463–514, Oct. 1994.

[4] Keith Pavitt, The Oxford Handbook of Innovation, J. Fagerberg, D. C. Mowery, and R. R. Nelson, Eds. Oxford, U.K.: Oxford Univ. Press, Jan. 2005.

[5] M. Meyer, “Does science push technology? Patents citing scientific literature,” Res. Policy, vol. 29, no. 3, pp. 409–434, 2000.

[6] H. Ettzokowitz and L. Leydesdorff, “The dynamics of innovation: From national systems and “mode 2” to a triple helix of university–industry–government relations,” Res. Policy, vol. 29, pp. 109–123, 2000.

[7] R. Nelson, National Innovation Systems: A Comparative Analysis. Oxford, U.K.: Oxford Univ. Press, 1993.

[8] A. Van Raan, Handbook of Quantitative Studies of Science and Technology. Amsterdam, The Netherlands: Elsevier, 1988.

[9] Z. Grilliches, “Patent statistics as economic indicators: A survey,” J. Econ. Literature, vol. XXVIII, pp. 1661–1707, Mar. 1990.

[10] J. Schmookler, Invention and Economic Growth. Cambridge, U.K.: Harvard Univ. Press, 1958.

[11] M. Meyer, “Tracing knowledge flows in innovation systems—An informetric perspective on future research science-based innovation,” Econ. Syst. Res., vol. 14, no. 4, pp. 323–344, 2002.

[12] M. Meyer, “Are patenting scientists the better scholars?: An exploratory comparison of inventor-authors with their non-inventing peers in nanoscience and technology,” Res. Policy, vol. 35, pp. 1646–1662, 2006.

[13] F. Narin, K. Hamilton, and D. Olivastro, “The increasing linkage between us technology and public science,” Res. Policy, vol. 26, pp. 317–330, Oct. 1997.

[14] W. Glänzel and M. Meyer, “Patents cited in the scientific literature: An exploratory study of “reverse” citation relations,” Scientometrics, vol. 58, no. 2, pp. 415–428, 2003.

[15] “Extracting the commercialization gap between science and technology: Case study of a solar cell,” Technological Forecasting and Social Change, vol. 77, no. 7, pp. 1147–1155, Sep. 2010.

[16] J. Callaert, B. Van Looy, A. Verbeek, K. Debackere, and B. Thijs, “Traces of prior art: An analysis of non-patent references found in patent documents,” Scientometrics, vol. 69, no. 1, pp. 3–20, 2006.

[17] B. D. Van Looy et al., “Do science-technology interactions pay off when developing technology? An exploratory investigation of 10 science-intensive technology domains,” Scientometrics, vol. 57, no. 3, pp. 355–367, 2003.

[18] J. Michel and B. Bettels, “Patent citation analysis. A closer look at the basic input data from patent search reports,” Scientometrics, vol. 51, pp. 185–201, 2001.

[19] J. Callaert, M. Pellens, and B. Van Looy, “Sources of inspiration? Making sense of scientific references in patents,” Scientometrics, vol. 98, no. 3, pp. 1617–1629, 2014.

[20] V. I. Torvik, M. Weeber, D. R. Swanson, and N. R. Smalheiser, “A probabilistic similarity metric for medline records: A model for author name disambiguation,” J. Assoc. Inf. Sci. Technol., vol. 56, no. 2, pp. 140–158, 2005.

[21] G.-C. Li et al., “Disambiguation and co-authorship networks of the us patent inventor database (1975–2010),” Res. Policy, vol. 43, no. 6, pp. 941–955, 2014.

[22] A. F. van Raan, “Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research,” J. Data Inf. Sci., vol. 2, no. 1, pp. 13–50, 2017.

[23] L. Leydesdorff and I. Rafols, “Local emergence and global diffusion of research technologies: An exploration of patterns of network formation,” J. Assoc. Inf. Sci. Technol., vol. 62, no. 5, pp. 846–860, 2011.

[24] H. Small, K. W. Boyack, and R. Klavans, “Identifying emerging topics in science and technology,” Res. Policy, vol. 43, no. 8, pp. 1450–1467, Oct. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0040733014000298

[25] S. P. Upham and H. Small, “Emerging research fronts in science and technology: Patterns of new knowledge development,” Scientometrics, vol. 83, no. 1, pp. 1–15, 2010.

[26] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, “Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics,” J. Assoc. Inf. Sci. Technol., vol. 68, no. 8, pp. 1925–1939, 2017.

[27] Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, and J. Lu, “Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research,” Technological Forecasting and Social Change, vol. 105, pp. 179–191, 2016.

[28] J. Winnink and R. J. Tijssen, “R&D dynamics and scientific breakthroughs in HIV/AIDS drugs development: The case of integrase inhibitors,” Scientometrics, vol. 101, no. 1, pp. 1–16, 2014.
[53] H.-Y. Sung, C.-C. Wang, M.-H. Huang, and D.-Z. Chen, “Measuring science-based science linkage and non-science-based linkage of patents through non-patent references,” J. Informetrics, vol. 9, no. 3, pp. 488–498, Jul. 2015.

[54] R. Li, T. Chambers, Y. Ding, G. Zhang, and L. Meng, “Patent citation analysis: Calculating science linkage based on citing motivation,” J. Assoc. Inf. Sci. Technol., vol. 65, no. 5, pp. 1007–1017, May 2014.

[55] F. Narian and D. Olivastro, “Status report: Linkage between technology and science,” Res. Policy, vol. 21, no. 3, pp. 237–249, 1992.

[56] L. W. Tussen, R. K. Butler, and T. N. van Leeuwen, “Technological relevance criteria: An assessment of citation linkages between patents and research papers,” Scientometrics, vol. 47, no. 2, pp. 389–412, 2000.

[57] M. Meyer, “What is special about patent citations? Differences between scientific and patent citations,” Scientometrics, vol. 49, no. 1, pp. 93–123, 2000.

[58] Z. Hu, S. Fang, and T. Liang, “Empirical study of constructing a knowledge organization system for patent documents using topic modeling,” Scientometrics, vol. 100, no. 3, pp. 787–799, 2014.

[59] S. Venugopalan and V. Rai, “Topic based classification and pattern identification in patents,” Technological Forecasting Social Change, vol. 94, pp. 236–250, 2015.

[60] W. S. Lee, E. J. Han, and S. Y. Sohn, “Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents,” Technological Forecasting Social Change, vol. 100, pp. 317–329, 2015.

[61] B. Song and Y. Suh, “Identifying convergence fields and technologies for industrial safety: LDA-based network analysis,” Technological Forecasting Social Change, vol. 138, pp. 115–126, 2019.

[62] S. L. Woltmann and L. Alkærsig, “Tracing university–industry knowledge transfer through a text mining approach,” Scientometrics, vol. 117, no. 1, pp. 449–472, 2018.

[63] X. Li, Q. Xie, J. Jiang, Y. Zhou, and L. Huang, “Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of Perovskite solar cell technology,” Technological Forecasting Social Change, to be published.

[64] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, “Reading tea leaves: How humans interpret topic models,” PNAS, vol. 31, 2009, pp. 1–9.

[65] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. Murty, “On finding the natural number of topics with Latent Dirichlet Allocation: Some observations,” Proc. Pacific-Asia Conf. Knowl. Discover. Data Mining, 2010, pp. 391–392.

[66] A. K. Jain, “Data clustering 50 years beyond k-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[67] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[68] S. Ranaei, M. Karvonen, A. Suominen, and T. Kssi, “Patent-based technology forecasting and social change: A case study on Taxol,” Scientometrics, vol. 100, no. 3, pp. 787–799, 2014.

[69] A. Suominen and M. Seppänen, “Bibliometric data and actual development of ecosystem genesis: A tale of two drugs,” in Proc. ISPIM Innov. Symp., 2015, pp. 1–11.

[70] A. Suominen and O. Dedehayir, “Pathways to a drug: A mixed methods analysis of emergence,” Int. J. Innov. Manag., vol. 21, no. 08, 2017, Art. no. 1740011.

[71] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack, “Openend: An open-source toolbox for large graph layout,” ProC. SPIE, vol. 7868, 2011, Art. no. 788606.

[72] A. Hess and F. T. Rothaermel, “Intellectual human capital and the emergence of biotechnology: Trends and patterns, 1974–2006,” IEEE Trans. Eng. Manag., vol. 59, no. 4, pp. 65–76, Feb. 2012.

[73] T. A. Hemphill, “Economic considerations in cooperative research and development agreements (CRADA): The case of Taxol, NIH, and technology transfer,” Technol. Soc., vol. 28, no. 3, pp. 321–331, 2006.

[74] K. C. Nicolau and R. K. Guy, “The conquest of taxol,” Angewandte Chemie Int. Ed., vol. 34, no. 19, pp. 2079–2090, 1995.

[75] A. Dedehayir, R. J. Ortt, S. J. Mäkinen, and A. Chakrabarti, “Towards a framework of ecosystem genesis: The story of Taxol,” in Proc. ISPIM Innov. Symp., 2015, pp. 1–11.

[76] J. Goodman and V. Walsh, “The story of taxol: Nature and politics in the pursuit of an anti-cancer drug,” Biomed. J., vol. 323, 2001, Art. no. 7304.

[77] S. Fortunato, “Community detection in graphs,” in Proc. Neural Inf. Process. Syst., 2007, pp. 593–600.

[78] A. Porter and M. J. Detampel, “Technology opportunities analysis,” in Proc. Neural Inf. Process. Syst., 2009, pp. 288–296.

[79] R. Li, T. Chambers, Y. Ding, G. Zhang, and L. Meng, “Patent citation analysis: Calculating science linkage based on citing motivation,” J. Assoc. Inf. Sci. Technol., vol. 65, no. 5, pp. 1007–1017, May 2014.

[80] F. Narian and D. Olivastro, “Status report: Linkage between technology and science,” Res. Policy, vol. 21, no. 3, pp. 237–249, 1992.

[81] R. J. W. Tussen, R. K. Butler, and T. N. van Leeuwen, “Technological relevance criteria: An assessment of citation linkages between patents and research papers,” Scientometrics, vol. 47, no. 2, pp. 389–412, 2000.

[82] M. Meyer, “What is special about patent citations? Differences between scientific and patent citations,” Scientometrics, vol. 49, no. 1, pp. 93–123, 2000.