Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03044-y OPEN

Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic

Keungoui Kim [1,2], Dieter F. Kogler [2] & Sira Maliphol[3]✉

Global scientific output is expanding exponentially, which in turn calls for a better understanding of the science of science and especially how the boundaries of scientific fields expand through processes of emergence. The present study proposes the application of embedded topic modeling techniques to identify new emerging science via knowledge recombination activities as evidenced through the analysis of research publication metadata. First, a dataset is constructed from metadata derived from the Web of Science Core Collection database. The dataset is then used to generate a global map representing a categorical scientific co-occurrence network. A research field is defined as interdisciplinary when multiple science categories are listed in its description. Second, the co-occurrence networks are subsequently compared between periods to determine changing patterns of influence in light of interdisciplinarity. Third, embedded topic modeling enables unsupervised association of interdisciplinary classification. We present the results of the analysis to demonstrate the emergence of global interdisciplinary sciences and further we perform qualitative validation on the results to identify what the sources of the emergent areas are. Based on these results, we discuss potential applications for identifying emergence through the merging of global interdisciplinary domains.

[1] School of Applied Artificial Intelligence, Handong Global University, Pohang, South Korea. [2] Spatial Dynamics Lab, School of Architecture, Planning & Environmental Policy & Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland. [3] Dept. of Technology & Society, the State University of New York, Songdo, South Korea. ✉email: sira.maliphol@sunykorea.ac.kr

Introduction

Science-driven research productivity and associated innovation processes have become increasingly complex for a number of reasons (Bloom et al. 2020; Boyack et al. 2017; Chen 2006; Chu and Evans 2021; Jones 2009; Kozlow 2023). Globally, over 2.6 million scientific articles were published in 2018 alone (White 2019). As scientific output increases over time, there has also been an increasing variety of sources of emergent topics as a result of the recombination of subjects and fields. Emergent topics that cross science fields are expected to be less path dependent than past patterns of scientific knowledge production. In line with this, Fortunato et al. (2018) emphasize the need to understand the science of science, especially as disciplinary boundaries break down.

> Contemporary science is a dynamical system of undertakings driven by complex interactions among social structures, knowledge representations, and the natural world. Scientific knowledge is constituted by concepts and relations embodied in research papers, books, patents, software and other scholarly artifacts, organized into scientific disciplines and broader fields. These social, conceptual, and material elements are connected through formal and informal flows of information, ideas, research practices, tools, and samples. Science can thus be described as a complex, self-organizing, and constantly evolving multiscale network. (Fortunato et al. 2018, p. 1)

While research output has risen, scientific productivity—or the value derived from that output—has fallen across fields (Bloom et al. 2020). The rate of innovation has slowed because the level of specialization (Jones 2009) and the size of teams (Kozlow 2023) needed to conduct science has increased. Intertwined with specialization and team size, the costs of research and development have sharply risen, reducing the rate of science productivity (Bloom et al. 2020). Another reason is how emergence has been measured. For instance, as the volume of scientific output increases, the ability to evaluate emerging research topics decreases because canonical literature is more likely to be cited (Chu and Evans 2021). "Could we be missing fertile new paradigms because we are locked into overworked areas of study?" (Chu and Evans 2021, p.5). Moreover, could we be misidentifying where emerging value is derived from science?

This has important implications considering the importance of scientific forecasting for understanding and developing effective science, technology, and innovation (STI) policy initiatives that aim to support science and to predict innovation trajectories (Börner et al. 2018). Essentially, innovative outcomes are frequently the result of converging technologies that often heavily depend on interdisciplinary scientific inputs (Kogler et al. 2022). Thus, and perhaps not surprisingly, contemporary attempts to address and to meet global grand challenges are directed toward interdisciplinary research where a deep integration of disciplines that combine different types of scientific and technological paradigms in genomic/biotechnology, nanotechnology, and information technology (e.g., blockchain, sensors, AI, and Big Data) are often believed to be the most promising avenues to pursue (Petersen et al. 2021). Recent examples, such as the mRNA vaccine for COVID-19, confirm this notion as they are usually the result of several decades of scientific research that might only become highly effective once the advances in various scientific fields are combined in a single applicable technological solution or innovation. Past convergence[1] stems from emergent interdisciplinary fields, e.g., biotechnology, which further catalyze innovations from other sectors (Feldman et al. 2015). Thus, changes at the interdisciplinary boundaries that are in flux may provide further insights into potential future convergence activities.

New discoveries, especially those with multi-disciplinary roots, are usually difficult to attribute to existing classification schemas (Fagerberg et al. 2012), but equally, they define the frontier of the innovation process as they combine existing forms of knowledge into something entirely novel (Eisenhardt and Martin 2000; Lee et al. 2015; Schumpeter 1934; 1942). Thus, interdisciplinary fields of science can be used to define the emergence of new topics (Chakraborty 2018; Khan and Wood 2015; Lee et al. 2015). Utilizing bibliometric network analysis on publication metadata, the present study proposes an approach capable of identifying from where interdisciplinary science fields emerge based on a global scientific map that indicates also changes in the growth of influence.

Specifically, the investigation employs topic modeling to classify scientific research topics from a large amount of data using unsupervised algorithms. The suggested embedded topic modeling approach then enables identification of emerging science topics in line with Schumpeterian notions of knowledge recombination processes where it is possible to observe how the combination of multiple disciplines or science categories unfolds over time. Unlike technology convergence that has been studied more systematically (Lee et al. 2019), few studies, to the best of our knowledge, have directed similar research efforts towards interdisciplinary knowledge recombination processes and how these might impact the overall evolution of the entire scientific knowledge landscape and subsequent innovation outcomes. Moreover, the application of topic modeling in natural language processing (NLP) environs to emerging interdisciplinary science studies holds the potential to provide important insights. The novel approach of combining embedded topic modeling and co-occurrence network analysis methods across global science maps can help with identifying emerging science topics before they consolidate into fields and predict those with potential value for knowledge recombination leading to global convergence.

The overarching goal is to analyze the complexity, self-organization, and evolution of scientific knowledge production while sifting through a large volume of scientific publications, and to understand how it might be possible to anticipate scientific innovations as they emerge from converging areas of research. The main objective of the present study is then to provide a novel approach to the bibliometric analyses toolkit by combining network analysis and embedded topic modeling techniques for the identification of emergent scientific topics of research interdisciplinarity.

Further, a novel measure for emergent topics is developed and employed, utilizing the network centrality index. Additionally, we leverage an embedded topic modeling technique, specifically BERTopic (Bidirectional Encoder Representations from Transformers), to gain insights into the emergent and globally domain-crossing profiles within interdisciplinary science fields. Through this comprehensive approach, we aim to illuminate the evolution of the science of science by investigating the changing boundaries of interdisciplinary research.

In the following sections, we provide an overview of the relevant literature in this line of inquiry, introduce the methodology followed by overall and detailed empirical findings, and finally offer a detailed discussion and some concluding thoughts.

Literature review

Science maps were developed to understand patterns related to the science of science, which include identifying topics of interest (Zahedi and van Eck 2018), identifying growth rates of science (Bornmann and Mutz 2015), identifying topic emergence (Jung and Segev 2022a), and detecting patterns and trends in the scientific literature (Kim and Chen 2015), especially through new combinations of interdisciplinary fields of science and technologies (Blei and Lafferty 2007; Eum and Maliphol 2023; Khan and Wood 2015; Lee et al. 2015). Science maps are network representations of the scientific literature that have evolved in research approaches (Chen 2006). Underlying these past approaches is an emphasis on finding radically new innovations *within* a specialized domain of science.

The evolution of the literature on emergence began with citation analysis and currently combines methods that identify network patterns using topic modeling techniques (Rotolo et al. 2015). Network analysis is commonly used to map the trends and patterns in the scientific literature, e.g., linked through citations, including the emergence of new seminal discoveries that change the course of a science specialization (Chen 2006). Science mapping linking research literature through citations can be used to demonstrate different evolutionary stages of scientific development over time, allowing the identification of transformative contributions through predictive analysis (Chen 2017). Models have been designed to include different aspects of the science of science. Science overlay maps represent subsets or networks of publications of global base maps, distinguishing different levels of research field categorization (Sjögårde 2022).

Emerging technologies from science can be defined by characteristics measured through bibliometric indicators and text analysis (Rotolo et al. 2015). By combining full-text analysis and bibliometric indicators, Glenisson et al. (2005) piloted a study that demonstrated the usefulness of data mining and bibliometric techniques that facilitate mapping fields of science. Patterns of scientific emergence have been modeled through clustering (Glänzel and Thijs 2012; Yau et al. 2014), national output (Suominen and Toivanen 2016), and using networks to demonstrate emergence (Khan and Wood 2015).

The emergent topics are expected to grow rapidly out of uncertain and ambiguous areas of research and converge to make a novel impact (Rotolo et al. 2015). Past studies on emergence focus on local maps or predefined areas of study, e.g. Curran and Leker (2011) on the nutraceuticals industry; Rey-Martí et al. (2016) on social entrepreneurship; and Song et al. (2017) on personalized medicine. Existing studies that demonstrate emergence have been carried out through bibliometric analyses using frequency-based topic modeling techniques that identified science topics (Griffith et al. 2004), topic coherence (Newman et al. 2011), topic "bursts" (Mane and Börner 2004), and patterns of scientific breakthrough (Winnink et al. 2019). Emergence is often identified through a measure of diversity within the local map, e.g., Rao-Stirling diversity and relative variety (Leydesdorff and Rafols 2011; Leydesdorff et al. 2019; Rafols and Meyer 2010).

The studies of emergent science are limited in scope by constraining fields of study through specific journals, articles, or authors. Once the science map is generated, topic modeling is analyzed based on network values generated from the map. The terms with higher frequency in the corpus are identified as emergent topic clusters. Thus, these studies examine the science of science generated within a science subject, category, or journal group based on measures of frequency and diversity within a local map. These approaches define the distance of interdisciplinarity through relative measures within the field of science. By relying on frequency, past approaches are more subject to canonical bias and may ignore context. Thus, the influence or importance of an interdisciplinary science pair in a science map offers an alternative approach to identifying emergence.

Novelty is also necessary to define emergence (Rotolo et al. 2015). Novelty can be identified through the merging of previously separate "streams of research" or fields of science (Day and Schoemaker 2000; Shin et al. 2022; Small et al. 2014). Thus, another measure of emergent organization is fast-growing multiple field or technology interdisciplinarity (Bornmann 2013; Bornmann and Marx 2014; Lee et al. 2021; Leydesdorff et al. 2013). Over time, research has become increasingly interdisciplinary (Chakraborty 2018). Research fields go through three stages: growth, maturity, and interdisciplinarity (Chakraborty 2018).

How disciplines are classified and differentiated, however, is still unsettled and still needs to be operationalized (Sugimoto and Weingart 2015). One method of defining disciplines is by using data-based publication indices such as Web of Science (WoS) categories (Sugimoto and Weingart 2015). Following this, interdisciplinarity can be modeled using keywords, authors' fields of study, and citations that cross multiple disciplines (Chakraborty 2018; Xu et al. 2018, 2019). Topic prediction using network analysis has been used to find emergent patterns across domains that are pre-defined and linked through co-occurrence frequency (Jung and Segev 2022b).

The measure of interdisciplinarity must balance variety and similarity (Leydesdorff 2018). When comparing against global data, limiting topic detection within a single discipline neglects to consider the increasingly interdisciplinary nature in which science is conducted (Boyack 2017). Using global maps leads to more accurate partitions and higher textual coherence of topics because the entire context is preserved. (Klavans and Boyack 2011). Moreover, long distances between interdisciplinary topics tend to have a higher scientific impact (Larivière et al. 2015). When scientific research incorporates new technological ideas, the convergent science tends to have a greater impact (Kwon et al. 2019). Further, humanities and social science research tends to have lower citation density which leads to lower measures of interdisciplinarity (Larivière et al. 2015).

While many investigations use interdisciplinary measures of emergence, past studies frequently restricted the analysis to local science maps that focus on a narrow field of science using relative measures for emergence. Furthermore, the formation of interdisciplinary research in the relevant literature has been mainly modeled through the evolution of keyword co-occurrence (Xu et al. 2018). Thus, one of the significant limitations of existing studies concerning the identification of thematic structures and dynamic patterns is that researchers constructed scientific maps around pre-defined topics (Gläser et al. 2017). By limiting the topic scope, the approaches resorted to using frequency-based measures of variety to determine relative novelty, and speed to define emergence. Frequency-based keyword evolution, however, can constrain our understanding of interdisciplinarity, disregard context, and intensify canonical bias. In contrast, global science maps can provide unbiased results if the size of the documents is sufficiently large (Rafols et al. 2010). While some studies differentiate between multi-, inter-, and trans-disciplinary (Chakraborty 2018; Leydesdorff et al. 2018), the operationalization of these distinctions remains limited. Thus, this study distinguishes the concept of growing and dominant sciences focused on broadly identifying the importance of interdisciplinarity across networks of STEM domains.

Methodology

The present study combines network analysis and BERTopic and applies it to understand cross-domain topic areas. BERTopic is an integrated topic modeling technique using embedding vector and c-TF-IDF to create dense clusters allowing interpretable topics from text data. Traditional text analysis is a labor-intensive activity that limits sample sizes to the speeds that human researchers are capable of reading, even ambitious studies are limited to a few hundred. For this reason, topic modeling

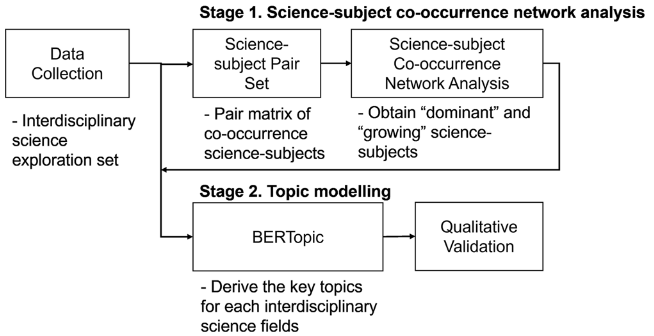

Fig. 1 Overall research process. The overall research process is performed in two stages: (i) defining a network of documents based on science-subject pairs and (ii) identifying topics from the network data.

techniques based on the frequency-based approach (ex. Latent Semantic Analysis, Latent Dirichlet Allocation, Dynamic Topic Model) were introduced to derive unobserved topics from a very large number of texts. However, frequency-based approaches remove context by relying only on term frequencies. New embedding-based approaches such as BERTopic, allow us to consider the contextual knowledge of large text data sets. The Web of Science Raw Data (WoS)[2], with over 63 million publication records found in 12,500 high-quality journals, is a common target of bibliometric analysis.

The data and methods used for the empirical analysis are introduced in accordance with the overall research process described in two stages (Fig. 1): data collection and pre-processing, network analysis of an interdisciplinary science dataset, and topic modeling of the newly constructed dataset. The first stage gathers and prepares the data from the journal publication metadata for network analysis and topic modeling. In stage 1, science category-subject network analysis is conducted to construct an interdisciplinary science network. In this constructed interdisciplinary science network, the science category-subjects that have greater network centrality, i.e. those that have greater potential value in terms of knowledge recombination, are defined. Here, the dataset is divided into two consecutive periods to create two interdisciplinary science networks. Comparing network values in two periods, science category-subjects that are more likely to grow (emerging science field) and that are more likely to have greater frequency (dominant science field) in the following period are selected to filter the final text dataset for topic modeling. Through this step, more precise and accurate data on publications can be extracted by filtering ones including such science category-subjects to restrict the data to the 'emerging science fields'. Utilizing the filtered list of publications, in the following subsection (Fig. 1, stage 2), topic modeling is conducted to explore the emerging topics in each interdisciplinary science field. This stage includes all the required processes for running the BERTopic model analysis. Through this process, latent topics representing each interdisciplinary science are derived. For qualitative validation, the publications that are the most representative of the emergent topics—which have been identified through the unsupervised learning process—are analyzed to identify what the topics of interest are for the given interdisciplinary categories.

Data collection. For the empirical analysis, the metadata is collected from the Web of Science Database. The database provides bibliometric information of scientific publications including the publication title, year, journal title, author, institution, institution's address, broad category, subject field, funding, citations, etc. The metadata should also include fields that enable differentiation by document type (ex. Article, editorial material, review, biographical item, letter, bibliography, correction, book review, meeting abstract, or proceedings paper) and publication type (journal, book in series, or book). These criteria allow us to restrict our sample to publications that are written for the same purpose, to maintain the quality of articles, and to avoid duplication. The dataset employed here is limited to journal articles by filtering its document and publication types.

Then, the list of publications that meet the definition of interdisciplinary science is selected and divided into three-year periods, which helps to stabilize dataset rankings (Archambault et al. 2009). By definition, interdisciplinary science refers to the cases where the scientific outcome is based on different research areas. In the WoS database, the research areas are defined by the scientific classifications, subheadings, and subjects. The broad global science category ('subheading' in WoS) indicates the top-level classification for the scientific fields including life-science & biomedicine (LSB), technology (TE), physical sciences (PS), arts & humanities, and social sciences. These categories are mutually exclusive. The subject field refers to a lower-tier classification of science that is assigned to an accordant category subheading. Here, all classifications are provided by WoS, as all journals and books included in WoS are categorized accordingly. In this study, an interdisciplinary science field is defined as the scientific outcome based on at least two subheadings, which are science categories.

In our WoS publication sample dataset, publications with technology- and science-based subheadings (LSB, TE, and PS) are used to maintain the consistency of the scientific fields. A total of 7,453,987 publications (from 10,138 journals) with 226 subjects are first collected over the reference period of 2012 and 2017. From this data set, global interdisciplinary science publications are filtered, which gives us 1,194,332 publications (from 1137 journals) with 172 subjects. Our final sample is restricted to publications that are classified as Journal Article (doc_type = ' Article' and pub_type = 'Journal') without missing abstracts. Table 1 presents the basic descriptive statistics on the number of publications, subjects, and journals for each interdisciplinary science field included in our final sample. Among all the interdisciplinary sciences, PS-TE has the greatest number of publications, subject, and journals, showing that it is the most active interdisciplinary science field. The increments of publication from all interdisciplinary science activities reflect the global trend of technology convergence as more heterogeneous technologies and industrial fields are used together over time.

Science category-subject co-occurrence network analysis

Science category-subject pair set. Prior to the science category-subject co-occurrence network analysis, a science category-subject co-occurrence pair set is constructed. In the interdisciplinary science dataset, a list of science category-subjects that are relevant to the category subheadings are assigned for each publication. Each science category-subject represents a node in the network connected by publications. To conduct co-occurrence network analysis, the combinations of category-subjects for each publication are transformed into a pair-form dataset for each interdisciplinary science field that defines the edges between nodes. We illustrate science category-subjects by signifying their categories with a capital letter (A, B, or C) and a number (1–9) to differentiate the science category-subjects within the categories. If publication X contains three science category-subjects A3, B6, and C9, it will have three rows of pair sets: A-B, B-C, and A-C. If a publication Y contains three science category-subjects of A1, A2, and B5, it will have two duplicate rows of interdisciplinary pair sets: A-B, A-B. Once the data set is transformed, the numbers of science category-subject pairs are aggregated by counting the

Table 1 Descriptive Statistics of Interdisciplinary Science Exploration Sets.

	2012-2014			2015-2017		
	Publication	Subject	Journal	Publication	Subject	Journal
LSB-TE	68,768	80	162	79,112	81	175
LSB-PS	115,499	67	228	120,161	67	248
PS-TE	345,520	85	584	414,010	86	637
LSB-PS-TE	25,447	43	40	25,805	43	43

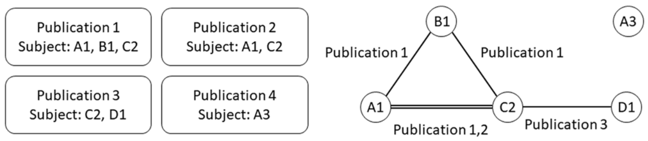

Fig. 2 Science category-subject co-occurrence network. The Science category-subject co-occurrence network shows an example network of publication nodes, e.g., Publication 1, linked by listed subjects, e.g., A1.

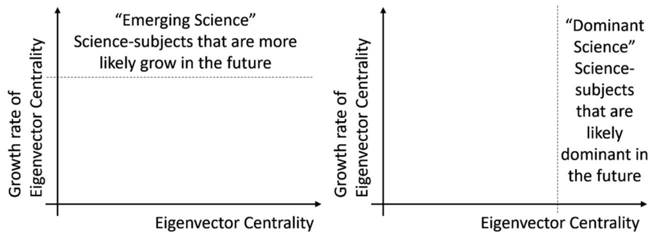

Fig. 3 Concept of growing and dominant interdisciplinary subjects. The graphs demonstrate how emerging science differs from dominant science as measured by Eigenvector centrality and the growth rate of Eigenvector centrality.

number of publications including such science category-subject pairs. The aggregated science category-subject pair set, therefore, presents the number of publications of science category-subject pairs in each interdisciplinary science in the respective period.

Science category-subject co-occurrence network analysis. Using subject pair sets, subject co-occurrence network analysis is conducted for interdisciplinary science fields in each period. A co-occurrence network is an effective method for analyzing the structural relationship between elements. A similar approach has been used with patent data for technology convergence analysis (Curran and Leker 2011; Kogler et al. 2017; Kim et al. 2018, 2019). In this regard, a co-occurrence network using publication data can provide greater understanding of how science category-subjects are being used and related to each other across interdisciplinary science fields. In a subject co-occurrence network, science category-subjects are used as nodes, and publications are used as edges. For the linkage rule, undirected and weighted networks are adopted. As shown in Fig. 2, science category-subjects are connected only if they were used in the same publication. For instance, subjects A and C have a total of two edges because they are used in publications 1 and 2.

Once the global network map is constructed for interdisciplinarity, the Eigenvector centrality (EIG) values of all nodes (in this network, science category-subjects) are measured. In this science category-subject co-occurrence network of interdisciplinary science, a science category-subject that is more important or influential can be regarded as a key science category-subject in an interdisciplinary science field, and those with a greater network value should be highlighted as they are the ones leading science category interdisciplinarity. Here, EIG measures the influence of network nodes beyond mere frequency counts by considering the centrality of connected nodes (West et al. 2013). For instance, a science category-subject connected to important science category-subjects is considered to have greater influence in the network. Rather than assuming equal importance, this measure differentiates the weight of edges by the importance of connected nodes. Unlike degree centrality, which solely focuses on the number of connections, EIG assesses a node's importance by evaluating the significance of its connections. This approach captures the qualitative aspect of network relationships. Furthermore, while PageRank is specifically tailored for directed networks, EIG's versatility allows it to be effectively applied to undirected networks as well. In this aspect, EIG can be used as an indicator for measuring the importance or influence of the emergent field interdisciplinarity (Heo & Lee, 2019; Qian et al., 2017; Rapach et al., 2015). With EIG, a network index that measures the influence of a node in a network by assigning weights to each connection based on the centrality of the connected node (Bonacich 2007), the key science category-subject in terms of being comparatively more important can isolated.

Using EIG, the conceptual framework of dominant and emerging science fields is proposed for the following purposes. First, by using EIG and its growth rate (EIG.GR), either dominant- or growing-sciences in terms of knowledge recombination can be determined. The threshold for dominant and growing interdisciplinary science is set to the top 10% of science category-subjects. Essentially, only those that are ranked in the top 10% in each measure are selected and named as dominant- and growing-sciences, respectively. Choosing the top 10% threshold for EIG and EIG.GR as criteria for identifying dominant or emerging science subjects is a deliberate methodological decision. This threshold is designed to selectively highlight the most influential or rapidly evolving fields, accounting for the skewed distribution of scientific networks where a few nodes accumulate the majority of connections. It allows for the identification of both established and emerging fields, reflecting on the dynamic nature of scientific research. A conservative approach like this minimizes false positives due to statistical fluctuations, ensuring that only subjects with consistently high metrics are considered. Furthermore, setting a clear benchmark facilitates comparative analysis over time and across disciplines, providing a consistent and reliable method for tracking changes in the scientific landscape. This choice underscores a strategic approach to recognizing significant trends and shifts within the realm of scientific research, emphasizing the importance of both sustained influence and notable growth in determining the prominence of science subjects.

As illustrated in Fig. 3, if the EIG (or EIG.GR) value of a science category-subject falls within the top 10%, it is considered to be a dominant (or emerging) science. If the values of both EIG and EIG.GR are within the top 10%, then the science category-subject can be classified as both dominant and emerging, signifying not only its current influence but also a significant increase in its impact. Conversely, if neither value falls within the top 10%, the science category-subject is not considered either

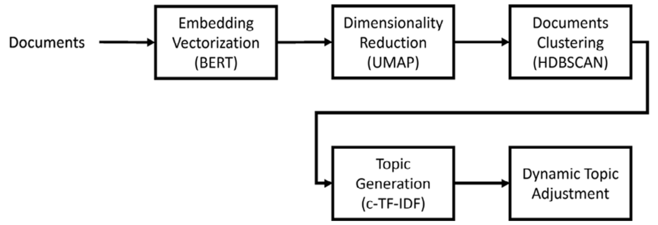

Fig. 4 Process of BERTopic modeling. The process of BERTopic modeling involves transforming document data into vectorized data, reducing the dimensionality, organizing the data into clusters and topics.

dominant or emerging. This allows us to focus on the specific list of publications that are more valuable in interdisciplinary science activity. Also, this contributes to improving the computation process for running text analysis by reducing the sample size. Rather than running text analysis for the whole sample, focusing on the selected publications that can be assumed to have more potential and to be consistent in terms of science subjects can improve the precision of our analysis. In this regard, selected growing interdisciplinary science category-subjects can be used as a reference for potential ones in the future. Due to the path-dependent nature of knowledge, a strong tendency or preference to follow such a trajectory is often observed, especially in knowledge-intensive activities. In other words, either a present network position or current network growth is very likely to be consistent also in the following period. This will be discussed in more detail with empirical findings in the following section.

Since the main interest of this study is exploring new rising topics in interdisciplinary science fields, we focus on growing-sciences rather than dominant-sciences. For the following step, publications representing growing interdisciplinary science category-subjects are filtered.

Embedded topic modeling

BERTopic. To derive topics for growing-sciences of each interdisciplinary science document, the BERTopic model is used. BERT, also known as Bidirectional Encoder Representations from Transformers, is a deep learning-based language model built on Transformer architecture developed by Google (Devlin et al. 2019). As presented in Fig. 4, the BERTopic is an integrated topic modeling technique that incorporates BERT embeddings, Unified Manifold Approximation and Projection (UMAP), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and a class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) (Grootendorst 2022).

The first step is embedding vectorization, which transforms target documents into vectors. Unlike conventional topic modeling methods that rely on Bag-of-Words (BoW) approaches, focusing solely on the frequency of terms, BERTopic utilizes embedding vectors. These embeddings represent documents in a space that, while lower in dimension compared to the vast potential vocabulary of BoW, is rich in capturing the deep semantic information inherent in the text. This allows for a higher contextual understanding of documents. By leveraging pre-trained word embeddings, BERTopic enables the analysis of documents with nuanced insights into their contextual meanings, surpassing the limitations of traditional encoding vectorization methods. Here, we utilized the default text representation model, "all-MiniLM-L6-v2", for our analysis. This model, designed as an all-purpose model, functions by converting sentences and paragraphs into a 384-dimensional dense vector space. It's versatile, suitable for tasks like clustering or semantic search, especially for English language text. Compared to the "all-mpnet-base-v2" model, one that is known to provide the best quality, it operates five times faster without compromising on quality[3], and its effectiveness has led to its adoption in various relevant studies (Samsir et al. 2023; Wang et al. 2023).

The second step in BERTopic involves dimensionality reduction. This is crucial because clustering algorithms, which are integral to topic modeling, perform better with lower-dimensional data. The primary challenge addressed here is the 'curse of dimensionality,' where high-dimensional spaces can negatively impact the efficiency and effectiveness of clustering algorithms. By reducing the dimensionality of the embedding space, BERTopic effectively mitigates this issue, facilitating more coherent and accurate topic clusters. This approach emphasizes the importance of tailoring data preprocessing steps to enhance the performance of specific algorithms used in the topic modeling process. For this reason, the UMAP algorithm is used to reduce the complexity of the embedding vector while preserving its essential structure. Assuming that high dimensional data lies on a lower dimension, UMAP maps highly complex data onto a simpler space efficiently by preserving the comparative distance and density and makes it easier to identify the cluster of similar documents (McInnes et al. 2016).

The following step is document clustering using HDBSCAN, which generates clusters based on the density of data points by using the hierarchical tree method. One of the strengths of HDBSCAN is that it can effectively identify and handle noise, which can help to derive more meaningful clusters. In addition, the combination of UMAP and HDBSCAN shows better performance in text clustering (Asyaky and Mandala 2021), and the clustering results can be modified by adjusting the hyperparameters regarding cluster generation.

The last step is topic generation with c-TF-IDF. c-TF-IDF is an adaptation of TF-IDF, which is designed to capture the representative terms from documents for each topic. TF-IDF is known as an effective measure for finding representative terms by combining term frequency and inverse document frequency (Salton and Buckley 1988). Under the assumption that a representative term of a document should be a distinctive one that represents the document, this measure simply captures the terms that not only occur more frequently in a document but also occur less frequently in other documents. By using c-TF-IDF[4] (Eq. 1), the importance of a term within a specific class can be found.

$$c - TF - IDF_{i,c} = \frac{tf_{i,c}}{w_c} \times \log \frac{N}{Docs(w)} \tag{1}$$

Qualitative validation of results. Once the interdisciplinary science maps have been analyzed, a list of representative publications for each interdisciplinary category can be generated based on the topics defined through BERTopic. Reliance on machine learning, however, can lead to misclassification (Lyutov et al. 2021), so we examine the results of the topic modeling to identify from where the newly emergent topic stems and describe them. Many recent studies that apply BERTopic have performed qualitative or manual validation of the results (Balcı et al. 2023; Capra, 2024; de Lima et al. 2023; Kasperiuniene et al. 2020; Wang et al. 2023). Using qualitative analysis, we review the results of the BERTopic process to validate them. First, the topic keywords are considered to determine if they provide a common theme for the articles under the topics. A qualitative approach is used to examine the topics to identify characteristics of emergent topics. After BERTopic is performed on the data sets, a list of topic keywords and representative articles emerge through the unsupervised process, e.g. topic-1. Additionally, traceability requires parsimony that the representations are unnecessarily complex such that even non-experts should be able to interpret them

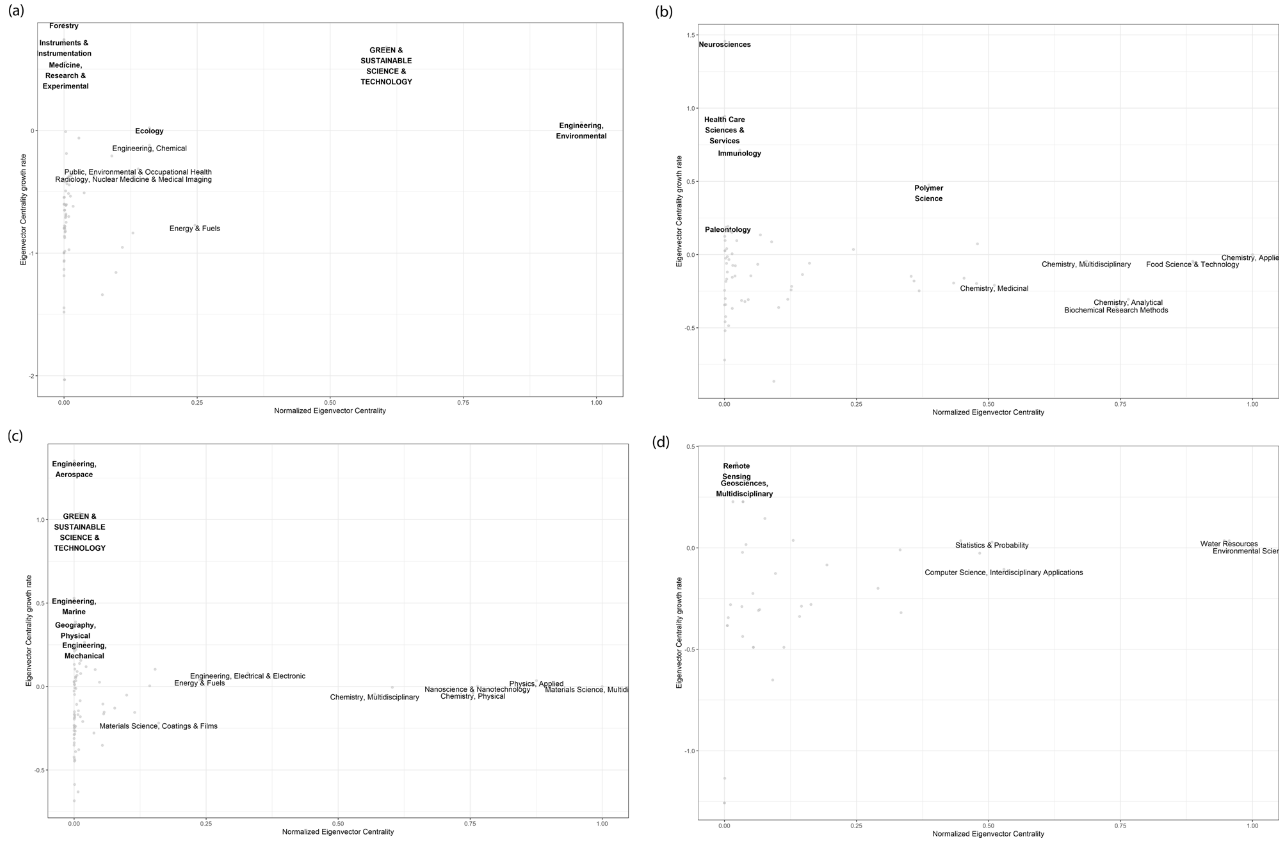

Fig. 5 Subject co-occurrence network analysis result. a LSB-TE. **b** LSB-PS. **c** PS-TE. **d** LSB-PS-TE. Note: The growing interdisciplinary science subjects are in bold.

(Rafols et al. 2010). The results are compared to check that they are rational or "make sense" to non-experts. Additionally, the journal lists are evaluated to discern the characteristics of the topics. Nonsensical topics would be expected to be random or not fit our definition of global interdisciplinary.

Case Study on Interdisciplinary Science in the Web of Science

Preparing the interdisciplinary science dataset. Following previous bibliometric studies using topic modeling techniques (Suominen and Toivanen 2016; Velden et al. 2017; Yau et al. 2014), we use the Web of Science Core Collection (WoS),[5] which is a database of peer-reviewed scholarly journals published worldwide. The WoS database provides the necessary metadata required for pre-processing, e.g. selecting peer-reviewed journal articles.

Results of science category-subject co-occurrence network analysis. In this section, the results of science category-subject co-occurrence network analysis are presented. Figure 5 illustrates the dominant- and growing-interdisciplinary science using the conceptual framework presented in Fig. 3, and Table 2 presents the full list of dominant- and growing-sciences. All nodes represent the science category-subjects included in each interdisciplinary science field, and dominant- (located further to the right on the x-axis) and growing-science (located higher on the y-axis) are labeled. One interesting point is that a clear distinction between dominant- and growing-interdisciplinary science is observed in all cases. Considering the path-dependent nature of knowledge, the dominant-sciences are likely to remain dominant in the following period. The prediction of key emergence trends, however, focuses on new interdisciplinary science category-subject merging that is expected to be more influential, rather than those that are already well-known. The gap between two types of science category-subjects justifies our approach to distinguishing promising science category-subjects in the future from those that already prevail, and more importantly, indicates that focusing on the emerging topics fits more into the purpose of this research.

This study focuses on the growing influence of interdisciplinary science to investigate the key topics that are likely to rise in the near future. In this regard, the publications including growing-interdisciplinary science are used for the following step of analysis. As shown in Table 2 and Fig. 6, EIG values of growing cross-domain science category-subjects in the following period tend to be greater than that of other science fields. This reflects that growing interdisciplinary science category-subjects in the current period have the greatest increases in the following period. With few exceptions, these subjects are different than those in the dominant-science fields. For BERTopic modeling, therefore, a set of cross-domain publications including growing-science are used.

Unsupervised classification of the emergent interdisciplinary science topics

BERTopic setting. While conventional topic modeling approaches consider the number of topics as an important hyperparameter to run analysis, BERTopic does not necessarily require it because UMAP and HDBSCAN ease the optimization of the clustering process, and automatically generate the list of topics. However, setting the number of topics is still important because a fully automated learning process may end up with an incomprehensible result. For instance, if BERTopic is conducted with its default settings and HDBSCAN optimization algorithms, it will automatically generate a list of topics, but this does not guarantee

Table 2 List of dominant and growing science category-subjects in interdisciplinary science fields.

Science	Dominant science category-subjects	Growing science category-subjects
LSB-TE	Environmental Sciences	Forestry
	Engineering, Environmental	Materials Science, Textiles
	Green & Sustainable Science & Technology	Instruments & Instrumentation
	Energy & Fuels	Pharmacology & Pharmacy
	Engineering, Chemical	Green & Sustainable Science & Technology
	Ecology	Medicine, Research & Experimental
	Public, Environmental & Occupational Health	Engineering, Environmental
	Radiology, Nuclear Medicine & Medical Imaging	Ecology
LSB-PS	Chemistry, Applied	Neurosciences
	Biochemistry & Molecular Biology	Health Care Sciences & Services
	Food Science & Technology	Immunology
	Chemistry, Analytical	Polymer Science
	Biochemical Research Methods	Paleontology
	Chemistry, Multidisciplinary	Microbiology
	Chemistry, Medicinal	Fisheries
PS-TE	Materials Science, Multidisciplinary	Engineering, Aerospace
	Physics, Applied	Green & Sustainable Science & Technology
	Nanoscience & Nanotechnology	Engineering, Marine
	Chemistry, Physical	Geography, Physical
	Physics, Condensed Matter	Water Resources
	Chemistry, Multidisciplinary	Engineering, Mechanical
	Engineering, Electrical & Electronic	Acoustics
	Energy & Fuels	Engineering, Ocean
	Materials Science, Coatings & Films	Automation & Control Systems
LSB-PS-TE	Environmental Sciences	Remote Sensing
	Water Resources	Imaging Science & Photographic Technology
	Engineering, Environmental	Geosciences, Multidisciplinary
	Computer Science, Interdisciplinary Applications	Crystallography
	Statistics & Probability	

Note: The list of science category-subjects are arranged in descending order.

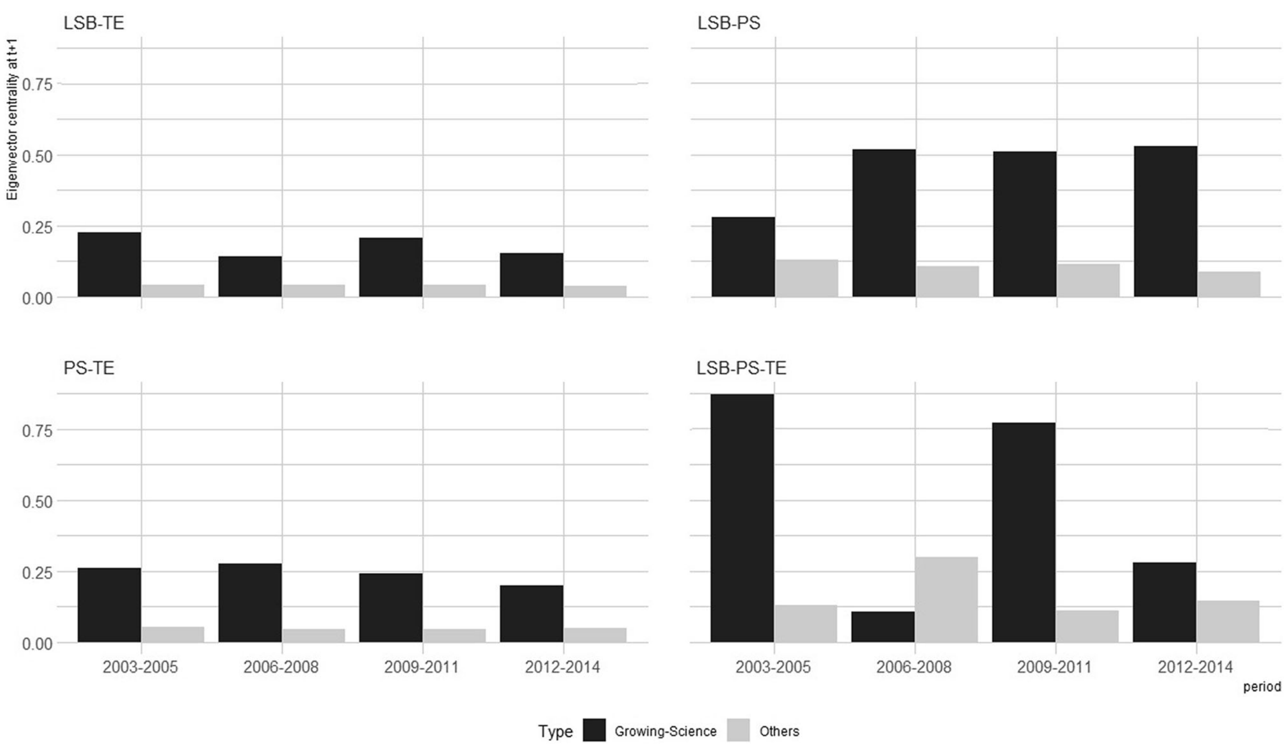

Fig. 6 Comparison of the EIG in following period between Growing-Interdisciplinary Science category-subjects and others. Note: On average, Eigenvector centrality in the following period of Growing-Interdisciplinary Science category-subjects (0.348) is higher than others (0.093).

Table 3 Hyperparameter testing of BERTopic.

	Publication	n-gram range	Number of topics	Minimum topic size
Growing-science of LSB-TE	26,164	(1,1) or (1,2) or (1,3)	50 ~ 1000	130–780
Growing-science of LSB-PS	10,577			50–300
Growing-science of PS-TE	49,042			240–1440
Growing-science of LSB-PS-TE	904			5–50

that the result is also acceptable in terms of application and obtaining insights.

For this reason, the three hyperparameters of n-gram range, number of topics, and minimum topic size are tested within ranges to find the best BERTopic model results (Table 3). The n-gram range determines whether the term should cover unigrams, bigrams, or trigrams, the number of topics sets the initial number of topics when running BERTopic, and the minimum topic size sets the minimum number of documents that each topic should contain. While the first two hyperparameter values were tested with the same range (n-gram range: unigram, bigram, trigram; number of topics: 5–1000), minimum topic size values proportional to the total number of publications were used. Minimum topic size values can be strongly affected by the size of documents, which may lead to topic sizes that are too broad or narrow for different cases. This especially largely influences the creation of outlier topics and an inexplicable number of topics. Thus, applying a proportional minimum topic size can help us minimize the size of outlier topics and maintain an explainable number of topics. For this reason, an integer value is used for the minimum topic size for each case that represents 0.5–3% of total publications. To help us consider a combination of different hyperparameters with wide ranges, a random search method is used to find an optimized parameter with random combinations, limited to no more than 100 iterations.

For each iteration, the information entropy value is measured (Eq. (2)) (MacKay 2003). By finding cases with uneven distribution of words in the topic, a set of topics with explicit semantic expression can be found (Wang et al. 2023). Known as a measurement of uncertainty, information entropy provides a means to determine whether topics can be clearly distinguished. In this regard, a model with the lowest information entropy value (Eq. (2)) is selected as the best model.

$$Entropy_i = -K \sum_{i=1}^{m} P\left(W_i|T\right) \log\left(P\left(W_i|T\right)\right) \tag{2}$$

BERTopic results. Once the dataset has been divided into different interdisciplinary sciences, the BERTopic process identifies articles that have similar topics, limited to the number of topics defined. The topics are defined through an unsupervised algorithm that identifies common lists of keywords that describe the topics.

Table 4 presents the groups of topics that appear in the greatest number of articles for each pairing of the subheadings: LSB-TE, LSB-PS, PS-TE, and LSB-PS-TE. The list of topic keywords identified in the interdisciplinary text set is used to define the topics. Outlier groups are used to prevent the formation of nonsensical or isolated topic groups.

Qualitative validation of results. Following recent studies that apply BERTopic (Balcı et al. 2023; Capra 2024; de Lima et al. 2023; Kasperiuniene et al. 2020; Wang et al. 2023), this study performed qualitative or manual validation of the results. While topic modeling may allow for the analysis of a large corpus of data, the results of the topic modeling should remain decipherable to non-experts (Rafols et al. 2010). Thus, we perform small-scale, qualitative analysis to verify that this condition holds.

While all articles are matched with the topic that is the most likely fit, not all articles that fall under the topic are equally representative of the topic. The representative articles are identified through the topic modeling technique, which means that they have the highest probability of matching the topic. The top 3 representative articles that fit the topics defined through topic modeling are provided in Table 5. All of the representative articles can be readily fit with the topics with which they are matched.

When considering the LSB-TE case, "Mechanical Properties and Composition of Natural Fibrous Materials" is most represented by articles LSB-TE-0-A through C. The article titles contain the phrases that are recognizably appropriate for the emergent topic: "tree bark," "insulation material," "manufacturing," "green-glued plywood panel," "resistance of thermally modified," "under extreme pressure," and "ash wood." Moreover, the journal titles are also representative of the topic: *Forest Products Journal* and *European Journal of Wood and Wood Products* (appears twice). Similar patterns are found for the other emergent topics listed in Table 5. Therefore, we find that the emergent topics that have been defined represent an easily recognizable theme. More broadly, many of the emergent topics are related to green technologies and sciences and to a lesser extent health-related technologies.

The journals with the greatest number of emergent interdisciplinary topic publications can be identified from the list of identified topics (Table 6). Yet, the journals in which the topics appear are clustered among a small portion of all publications; the distribution of publications with emergent interdisciplinary topics is skewed towards a small share of all journals in the dataset. Half of all publications were published in the top quintile of all journals for each interdisciplinary category group: 14th percentile (LSB-TE), 13th percentile (LSB-PS), 10th percentile (PS-TE), and 18th percentile (LSB-PS-TE). Additionally, when considering the top journals that emerge from the ranking of interdisciplinarity results, the categories become clearer when considering the emergent topics. For PS-TE, the emergent topics can only be seen in *Desalination and Water Treatment* and *International Journal of Hydrogen Energy*. The other titles are suggestive of the science and technologies involved: physical chemistry, sensors, and materials.

Discussion and conclusion

As science continues to expand its research output, the science of science emergence provides an opportunity to understand where new knowledge—the source of innovation—originates from by examining global interdisciplinarity. Most previous studies have focused on breakthroughs or identifying popular directions within narrow fields of study measured by frequency size. These past approaches apply the logic of identifying patterns of frequency-based dominant topics within a specific field of science. In contrast, the present study provides an alternative perspective in understanding the science of science emergence with a focus on the influence of the changing boundaries of conjoining science across categories. The main contributions of our research are (i) to expand the definition of interdisciplinarity to include global

Table 4 Topics and keyword lists for science category co-occurrence pairs[6].

Science Categories	Topic	Keywords	Generated Labels	Number of Documents
LSB-TE (26,164 publications)	Outlier	cutting, machining, leather, grinding, radon, asbestos, lubrication, tanning, lead, mql	-	272
	0	wood, lignin, properties, strength, mechanical, bamboo, moisture, cellulose, modulus, specimens	Mechanical Properties and Composition of Natural Fibrous Materials	22314
	1	water, study, energy, results, environmental, waste, model, using, production, based	Sustainable Environmental Technologies and Resource Management	1339
	2	patients, group, lt, groups, clinical, control, cancer, expression, cells	Cancer Biomarker Expression in Clinical Patient Groups	2239
LSB-PS (10,577 publications)	Outlier	brain, fnirs, imaging, optical, neurons, cortex, cortical, neural, attribution, stimulation	-	176
	0	species, water, sea, early, data, late, marine, formation, climate, new	Marine Biodiversity and Climate Impact Studies	5129
	1	model, data, models, proposed, methods, trial, regression, method, simulation, clinical	Clinical Trial Modeling and Simulation Techniques	397
	2	showed, activity, properties, protein, cells, chitosan, cell, ph, acid, drug	Chitosan Bioactivity and Drug Delivery Applications	4785
PS-TE (49,042 publications)	Outlier	forum, journal, views, readership, essays, speculation, editorial, provoking, asce, founded	-	11
	0	adsorption, removal, membrane, ph, process, concentration, water, treatment, mg, acid	Adsorption and Membrane Processes for Water Treatment	10,873
	1	heat, model, flow, results, data, based, water, method, temperature, transfer	Heat Transfer Modeling and Analysis in Fluid Systems	38,158
LSB-PS-TE (905 publications)	0	data, study, land, area, using, flood, spatial, based, model, used	Flood Risk Assessment and Spatial Modeling	334
	1	binding, molecular, protein, energy, interactions, docking, structure, dynamics, molecules, results	Protein-Molecule Docking and Interaction Dynamics	570

Table 5 Representative articles for each interdisciplinary emergent topic.

Science Categories-Emergent Topic	Representative Article
LSB-TE	
Mechanical Properties and Composition of Natural Fibrous Materials	0-1: Kain et al. (2015) 0-2: Lavalette et al. (2016) 0-3: Candelier et al. (2017)
Sustainable Environmental Technologies and Resource Management	1-1: Egle et al. (2015) 1-2: Palma-Rojas et al. (2017) 1-3: Harijani et al. (2017)
Cancer Biomarker Expression in Clinical Patient Groups	2-1: Liu et al. (2017) 2-2: Liu and Li (2017) 2-3: Qi et al. (2017)
LSB-PS	
Marine Biodiversity and Climate Impact Studies	0-1: Chen et al. (2016) 0-2: Bataille et al. (2016) 0-3: Lowery et al. (2017)
Clinical Trial Modeling and Simulation Techniques	1-1: French et al. (2016) 1-2: Luo et al. (2016) 1-3: Lu (2017)
Chitosan Bioactivity and Drug Delivery Applications	2-1: Zhao et al. (2016) 2-2: Berah et al. (2017) 2-3: Gomes et al. (2017)
PS-TE	
Adsorption and Membrane Processes for Water Treatment	0-1: Ahmed (2016) 0-2: Zhang et al. (2016) 0-3: Saadati et al. (2017)
Heat Transfer Modeling and Analysis in Fluid Systems	1-1: Colombo and Fairweather (2016) 1-2: Wu et al. (2017). 1-3: Daabo et al. (2017)
LSB-PS-TE	
Flood Risk Assessment and Spatial Modeling	0-1: Ding et al. (2017) 0-2: Chian and Wilkinson (2015) 0-3: Rizeei et al. (2016)
Protein-Molecule Docking and Interaction Dynamics	1-1: Shamim et al. (2015) 1-2: Khan et al. (2017) 1-3: Bobovská et al. (2016)

Note: The representative articles are preceded by the topic number and a number index, e.g., "0-1".

domain-crossing science categories, (ii) to use Eigenvector centrality as a measure of influence on emergent topics, and (iii) to demonstrate the use of embedded topic modeling over a dataset the represents a global science map. This study provides an early foray into applying unsupervised classification using BERTopic modeling on interdisciplinary science datasets. This approach is one of the few contemporary studies that apply text-embedding-based topic modeling techniques to the science of science emergence, and the only one to focus on the influence of existing science topics on emergence.

Furthermore, the present investigation provides a simple model to achieve the desired analysis and, in addition, demonstrates that the originating subjects of interdisciplinary topics can be identified using embedded topic modeling. Using the Schumpeterian definition of knowledge creation based on recombination processes, the model examines the intersection of interdisciplinary sciences to identify the most influential topics related to emergent scientific knowledge based on science topics that are projected onto a global science map. The results can be used to identify trend profiles of the interdisciplinary sources of emergent topics over time.

Since dominant science is subject to the bias of size and canonical fields, emergent science based on the influence of co-occurring science domains provides an alternative measure. The Eigenvector centrality value can be used as a measure for the growth of interdisciplinarity that is different from approaches that focus on dominant science in a co-occurrence network of interdisciplinary emergence. Dominant science subjects are different than the topics related to growing interdisciplinary science, differentiating the results of this study from prior studies that emphasize frequency-based, dominant science. The approach that we used allows us to retain contextual knowledge in text analysis. Nonetheless, those science subjects that appear in both emergent growing and dominant interdisciplinary sciences such as "Green & Sustainable Science & Technology" may indicate greater influence on research for society and have greater potential for applications.

This study suggests that identifying emergent topics may help us better understand how to direct and use innovative research. This study detected green- and health-related topics are emergent across many of the global interdisciplinary science categories. As global challenges emerge, more efficient and effective means to identify emergent research to address them are necessary; yet, it has become increasingly difficult to meet this aim (Petersen et al. 2021). Bloom et al. (2020) posit that if firms are shifting towards defensive research activities, then government policy must reconsider how research is publicly funded. In order to increase economic productivity, the sources (and barriers) of innovation need to be detected within sectors and individuals. Although this may help when focusing on economic-related challenges, there may be the need for additional measures of research productivity when considering socially oriented innovation demands. Thus, an alternative explanation for the decline in science productivity is that social innovation may be driving research rather than economic imperatives.

Although the present study has departed from prior studies in several aspects, further research is needed to address its limitations. First, the number of topics that were automatically generated was small, which means that there are likely additional emergent topics that can be identified in follow-up studies. Nevertheless, the current investigation adopted a conservative approach to ensure that the topics identified were meaningful, especially when considering that the distributions are highly skewed. Future research should also consider how to refine the level at which emergent topics are still acceptably defined, e.g., recursive clustering on large-scale bibliometric data (cf. Mejia and Kajikawa 2020) while balancing the diversity of domains and similarity of emergent topics. Additionally, the NLP approach adopted here requires a comparably large amount of computing power, which, in turn, might pose a challenge for universal day-to-day applications and policy purposes.

Another limitation is that our data is constrained to scientific journal articles in the WoS. Not all innovations—especially social innovations—may be derived from science and technology fields. This approach may also ignore disciplines that tend to produce other types of publications. A broader approach that considers these types of interdisciplinarity may provide alternative sources of identifying social innovation. Lastly, while this study focused on specific characteristics of emergence defined through interdisciplinarity in the WoS, future research assessments should "consider the value and impact of all research outputs" and "consider a broad range of impact measures," as stated in the San Francisco Declaration on Research Assessment (Cagan 2013). Rather than redefine emergence through science maps, this study aimed to explore a different approach to understanding emergence by providing an alternative perspective on emergence.

The science of science can link existing knowledge reservoirs for technology development, especially as global challenges influence the direction of science emergence that can be applied to the innovation of new technologies. A better understanding of the existing topics that are cross-domain and, as such, generate new innovative outcomes and solutions can help to apply the

Table 6 Top 10 journals by interdisciplinary category pairs.

Interdisciplinary categories	Journal Title	Number of documents	Share
LSB-TE	*Journal of Cleaner Production*	5589	21.4%
	Environmental Science & Technology	4524	17.3%
	Journal of Hazardous Materials	2417	9.2%
	Biomedical Research-India	1933	7.4%
	Ecological Engineering	1615	6.2%
	Waste Management	1368	5.2%
	Environmental Modeling & Software	735	2.8%
	Environmental Progress & Sustainable Energy	652	2.5%
	Resources Conservation and Recycling	541	2.1%
	Clean Technologies and Environmental Policy	512	2.0%
LSB-PS	*International Journal of Biological Macromolecules*	3347	31.6%
	Paleogeography Paleoclimatology Paleoecology	1257	11.9%
	Biomacromolecules	1250	11.8%
	ICES Journal of Marine Science	698	6.6%
	Cretaceous Research	586	5.5%
	Marine and Freshwater Research	501	4.7%
	Statistical Methods in Medical Research	399	3.8%
	Journal of Water and Health	282	2.7%
	Paleoceanography	268	2.5%
	Food and Agricultural Immunology	259	2.4%
PS-TE	*Desalination and Water Treatment*	5622	11.5%
	Applied Thermal Engineering	5040	10.3%
	International Journal of Heat and Mass Transfer	3791	7.7%
	ACS Sustainable Chemistry & Engineering	2468	5.0%
	Journal of Hydrology	2206	4.5%
	Advances in Mechanical Engineering	1931	3.9%
	Ocean Engineering	1639	3.3%
	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	1447	3.0%
	Combustion and Flame	1093	2.2%
	Ultrasonics Sonochemistry	1089	2.2%
LSB-PS-TE	*Journal of Molecular Graphics & Modeling*	570	63.0%
	Geocarto International	212	23.4%
	Natural Hazards Review	115	12.7%
	Geocarto International	8	0.9%

science of science to applicable and effective STI policy initiatives that incorporate social innovation objectives as well.

Data availability

The data that support the findings of this study are available from the Web of Science but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Web of Science.

Received: 22 June 2023; Accepted: 12 April 2024;
Published online: 10 May 2024

Note

1 For clarity, in this paper we refer to 'convergence' as technological convergence with respect to the realization of new technologies unless otherwise stated.
2 https://clarivate.libguides.com/c.php?g=593069&p=4101845.
3 https://www.sbert.net/docs/pretrained_models.html.
4 i is the term and c refers to the class, $tf_{i,c}$ is the freqeuncy of term i extracted from class i, w_c is total number of terms from class i, N is the total number of documents.
5 Full list of science classification: https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-List-of-Subject-Classifications-for-All-Databases?language=en_US.
6 Following prompt has been used with ChatGPT (GPT-4): I have topic that contains the scientific publications related to ["Name of Interdisciplinary Science"]. The topic is described by the following keywords: ["List of keywords"] Based on the above information, can you give a short label of the topic?

References

Ahmed SA (2016) Removal of lead and sodium ions from aqueous media using natural wastes for desalination and water purification. Desalination Water Treat. 57(19):8911–8926

Archambault É, Campbell D, Gingras Y, Larivière V (2009) Comparing bibliometric statistics obtained from the Web of Science and Scopus. J Am Soc Inf Sci Technol. 60(7):1320–1326

Asyaky MS, Mandala R (2021) Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. Proc 2021 8th Int Conf Adv Inform Concepts Theory Appl 2021:1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640285

Balcı U, Sirivianos M, Blackburn J (2023) A data-driven understanding of left-wing extremists on social media. Preprint. arXiv preprint arXiv:2307.06981

Bataille CP, Watford D, Ruegg S, Lowe A, Bowen GJ (2016) Chemostratigraphic age model for the Tornillo Group: A possible link between fluvial stratigraphy and climate. Palaeogeogr Palaeoclimatol Palaeoecol 457:277–289

Berah R, Ghorbani M, Moghadamnia AA (2017) Synthesis of a smart pH-responsive magnetic nanocomposite as high loading carrier of pharmaceutical agents. Int J Biol Macromol 99:731–738

Blei DM, Lafferty J (2007) A correlated topic model of science. Annals Appl Stat 1(1). https://doi.org/10.1214/07-aoas114

Bloom N, Jones CI, Van Reenen J, Webb M (2020) Are ideas getting harder to find? Am Econ Rev 110(4):1104–1144

Bobovská A, Tvaroška I, Kóňa J (2016) Using DFT methodology for more reliable predictive models: Design of inhibitors of Golgi α-mannosidase II. J Mol Graph Model 66:47–57

Bonacich P (2007) Some unique properties of eigenvector centrality. Soc Netw 29(4):555–564. https://doi.org/10.1016/j.socnet.2007.04.002

Börner K, Rouse WB, Trunfio P, Stanley HE (2018) Forecasting innovations in science, technology, and education. Proc Natl Acad Sci 115(50):12573–12581

Bornmann L (2013) What is societal impact of research and how can it be assessed? A literature survey. J Am Soc Inf Sci Technol 64(2):217–233

Bornmann L, Marx W (2014) How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons. Scientometrics 98:211–219
Bornmann L, Mutz R (2015) Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol 66(11):2215–2222
Boyack K, Glänzel W, Gläser J, Havemann F, Scharnhorst A, Thijs B, van Eck NJ, Velden T, Waltmann L (2017) Topic identification challenge. Scientometrics 111:1223–1224
Boyack KW (2017) Investigating the effect of global data on topic detection. Scientometrics 111(2):999–1015
Cagan R (2013) The San Francisco declaration on research assessment. Dis Models Mech 6(4):869–870
Candelier K, Hannouz S, Thévenon MF, Guibal D, Gérardin P, Pétrissans M, Collet R (2017) Resistance of thermally modified ash (Fraxinus excelsior L.) wood under steam pressure against rot fungi, soil-inhabiting micro-organisms and termites. Eur J Wood Wood Prod 75:249–262
Capra L (2024) A computational linguistic approach to study border theory at scale. ACM Trans Comput-Hum Interaction 37(4):1–23
Chakraborty T (2018) Role of interdisciplinarity in computer sciences: quantification, impact and life trajectory. Scientometrics 114:1011–1029
Chen C (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 57(3):359–377
Chen C (2017) Science mapping: a systematic review of the literature. J Data Inf Sci 2(2):1–40
Chen J, Shen SZ, Li XH, Xu YG, Joachimski MM, Bowring SA, Mu L (2016) High-resolution SIMS oxygen isotope analysis on conodont apatite from South China and implications for the end-Permian mass extinction. Palaeogeogr Palaeoclimatol Palaeoecol 448:26–38
Chian SC, Wilkinson SM (2015) Feasibility of remote sensing for multihazard analysis of landslides in Padang Pariaman during the 2009 Padang earthquake. Nat Hazards Rev 16(1):05014004
Chu JS, Evans JA (2021) Slowed canonical progress in large fields of science. Proc Natl Acad Sci 118(41):e2021636118
Colombo M, Fairweather M (2016) Accuracy of Eulerian–Eulerian, two-fluid CFD boiling models of subcooled boiling flows. Int J Heat Mass Transf 103:28–44
Curran CS, Leker J (2011) Patent indicators for monitoring convergence - examples from NFF and ICT. Technol Forecast Soc Change 78(2):256–273. https://doi.org/10.1016/j.techfore.2010.06.021
Daabo AM, Al Jubori A, Mahmoud S, Al-Dadah RK (2017) Development of three-dimensional optimization of a small-scale radial turbine for solar powered Brayton cycle application. Appl Therm Eng 111:718–733
Day GS, Schoemaker PJ (2000) Avoiding the pitfalls of emerging technologies. Calif Manag Rev 42(2):8–33
de Lima BC, Baracho RMA, Mandl T, Porto PB (2023) Reactions to science communication: discovering social network topics using word embeddings and semantic knowledge. Soc Netw Anal Min 13(1):119
Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), 4171–4186
Ding Q, Chen W, Hong H (2017) Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping. Geocarto Int. 32(6):619–639
Egle L, Rechberger H, Zessner M (2015) Overview and description of technologies for recovering phosphorus from municipal wastewater. Resour Conserv Recycl 105:325–346
Eisenhardt KM, Martin JA (2000) Dynamic capabilities: What are they? Strategic Manag J 21(10):1105–1121
Eum W, Maliphol S (2023) Southeast Asian catch-up through the convergence of trade structures. Asian J Technol Innov 31(2):422–446
Fagerberg J, Landström H, Martin BR (2012) Exploring the emerging knowledge base of "the knowledge society. Res Policy 41(7):1121–1131. https://doi.org/10.1016/j.respol.2012.03.007
Feldman MP, Kogler DF, Rigby DL (2015) rKnowledge: The spatial diffusion and adoption of rDNA methods. Regional Stud 49(5):798–817. https://doi.org/10.1080/00343404.2014.980799
Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási AL (2018) Science of science. Science 359(6379). https://doi.org/10.1126/science.aao0185
French B, Saha-Chaudhuri P, Ky B, Cappola TP, Heagerty PJ (2016) Development and evaluation of multi-marker risk scores for clinical prognosis. Stat Methods Med Res 25(1):255–271
Glänzel W, Thijs B (2012) Using "core documents" for detecting and labelling new emerging topics. Scientometrics 91(2):399–416. https://doi.org/10.1007/s11192-011-0591-7
Gläser J, Glänzel W, Scharnhorst A (2017) Same data—different results? Towards a comparative approach to the identification of thematic structures in science. Scientometrics 111:981–998
Glenisson P, Glänzel W, Janssens F, De Moor B (2005) Combining full text and bibliometric information in mapping scientific disciplines. Inf Process Manag 41(6):1548–1572. https://doi.org/10.1016/j.ipm.2005.03.021
Gomes S, Rodrigues G, Martins G, Henriques C, Silva JC (2017) Evaluation of nanofibrous scaffolds obtained from blends of chitosan, gelatin and polycaprolactone for skin tissue engineering. Int J Biol Macromol 102:1174–1185
Griffith R, Redding S, Van Reenen J (2004) Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. Rev Econ Stat 86(4):883–895
Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794
Harijani AM, Mansour S, Karimi B, Lee CG (2017) Multi-period sustainable and integrated recycling network for municipal solid waste–A case study in Tehran. J. Clean. Prod. 151:96–108
Heo PS, Lee DH (2019) Evolution patterns and network structural characteristics of industry convergence. Struct Change Econ Dyn 51:405–426. https://doi.org/10.1016/j.strueco.2019.02.004
Jones BF (2009) The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? Rev. Econ Stud. 76(1):283–317
Jung S, Segev A (2022a) Analyzing the generalizability of the network-based topic emergence identification method. Semantic Web 13(3):423–439
Jung S, Segev A (2022b) Identifying a common pattern within ancestors of emerging topics for pan-domain topic emergence prediction. Knowl Based Syst 258:110020
Kain G, Barbu MC, Richter K, Plank B, Tondi G, Petutschnigg A (2015) Use of tree bark as insulation material. For Products J 65(3-4):S16–S16
Kasperiuniene J, Briediene M, Zydziunaite V (2020) Automatic content analysis of social media short texts: scoping reviewof methods and tools. In Costa. A.P., Reis, L.P., & Moreira, A. (eds.) *Computer Supported Qualitative Research: New Trends on Qualitative Research(WCQR2019)* 4, 89-101
Khan AM, Shawon J, Halim MA (2017) Multiple receptor conformers based molecular docking study of fluorine enhanced ethionamide with mycobacterium enoyl ACP reductase (InhA). J Mol Graph Model 77:386–398
Khan GF, Wood J (2015) Information technology management domain: emerging themes and keyword analysis. Scientometrics 105(2):959–972. https://doi.org/10.1007/s11192-015-1712-5
Kim K, Jung S, Hwang J (2019) Technology convergence capability and firm innovation in the manufacturing sector: an approach based on patent network analysis. RD Manag 49(4):595–606. https://doi.org/10.1111/radm.12350
Kim K, Jung S, Hwang J, Hong A (2018) A dynamic framework for analyzing technology standardisation using network analysis and game theory. Technol Anal Strat Manag 30(5):540–555. https://doi.org/10.1080/09537325.2017.1340639
Kim MC, Chen C (2015) A scientometric review of emerging trends and new developments in recommendation systems. Scientometrics 104:239–263
Klavans R, Boyack KW (2011) Using global mapping to create more accurate document-level maps of research fields. J Am Soc Inf Sci Technol 62(1):1–18
Kogler DF, Essletzbichler J, Rigby DL (2017) The evolution of specialization in the EU15 knowledge space. J. Econ Geogr 17(2):345–373. https://doi.org/10.1093/jeg/lbw024
Kogler DF, Whittle A, Buarque B (2022) The Science Space of Artificial Intelligence Knowledge Production. In: Kurz HD, Schütz M, Strohmaier R, Zilian SS (eds) The Routledge Handbook of Smart Technologies: An Economic and Social Perspective. Routledge, London, pp 241–268 https://doi.org/10.4324/9780429351921
Kozlow M (2023) "Disruptive" science has declined—even as papers proliferate. Springe Nat 613:225
Kwon S, Liu X, Porter AL, Youtie J (2019) Research addressing emerging technological ideas has greater scientific impact. Res Policy 48(9):103834. https://doi.org/10.1016/j.respol.2019.103834
Larivière V, Haustein S, Börner K (2015) Long-distance interdisciplinarity leads to higher scientific impact. Plos One 10(3):e0122565
Lavalette A, Cointe A, Pommier R, Danis M, Delisée C, Legrand G (2016) Experimental design to determine the manufacturing parameters of a green-glued plywood panel. Eur J Wood Prod 74:543–551
Lee C, Kogler DF, Lee D (2019) Capturing information on technology convergence, international collaboration, and knowledge flow from patent documents: A case of information and communication technology. Inf Process Manag 56:1576–1591
Lee C, Hong S, Kim J (2021) Anticipating multi-technology convergence: a machine learning approach using patent information. Scientometrics 126(3):1867–1896. https://doi.org/10.1007/s11192-020-03842-6
Lee WS, Han EJ, Sohn SY (2015) Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. Technol Forecast Soc Change 100:317–329. https://doi.org/10.1016/j.techfore.2015.07.022

Leydesdorff L (2018) Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance? Scientometrics 116:2113–2121
Leydesdorff L, Rafols I (2011) Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. J Informetr 5(1):87–100. https://doi.org/10.1016/j.joi.2010.09.002
Leydesdorff L, Rafols I, Chen C (2013) Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. J Am Soc Inf Sci Technol 64(12):2573–2586
Leydesdorff L, Wagner CS, Bornmann L (2018) Betweenness and diversity in journal citation networks as measures of interdisciplinarity—A tribute to Eugene Garfield. Scientometrics 114:567–592
Leydesdorff L, Wagner CS, Bornmann L (2019) Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. J Informetr 13(1):255–269
Liu HQ, Li XL (2017) Effect of nursing intervention on liver cancer patients undergoing interventional therapy. Biomed Res 28(12):5285–5288
Liu D, Zhao H, Liu B, Zhang X, Ma Q (2017) Analysis on the expression level of serum MMP-7 in patients with abdominal aortic aneurysm accompanied by hypertension and clinical efficacy of endovascular graft exclusion. Biomed Res (0970-938X), 28(3)
Lowery CM, Cunningham R, Barrie CD, Bralower T, Snedden JW (2017) The northern Gulf of Mexico during OAE2 and the relationship between water depth and black shale development. Paleoceanography 32(12):1316–1335
Lu T (2017) Bayesian nonparametric mixed-effects joint model for longitudinal-competing risks data analysis in presence of multiple data features. Stat Methods Med Res 26(5):2407–2423
Luo S, Lawson AB, He B, Elm JJ, Tilley BC (2016) Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. Stat Methods Med Res 25(2):821–837
Lyutov A, Uygun Y, Hütt MT (2021) Machine learning misclassification of academic publications reveals non-trivial interdependencies of scientific disciplines. Scientometrics 126(2):1173–1186. https://doi.org/10.1007/s11192-020-03789-8
MacKay DJ (2003) *Information theory, inference and learning algorithms*. Cambridge University Press
Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci USA 101(SUPPL. 1):5287–5290. https://doi.org/10.1073/pnas.0307626100
McInnes L, Healy J, Melville J (2016) UMAP: Uniform manifold approximation and projection for dimension reduction. http://arxiv.org/abs/1802.03426
Mejia C, Kajikawa Y (2020) Emerging topics in energy storage based on a large-scale analysis of academic articles and patents. Appl Energy 263:114625. https://doi.org/10.1016/j.apenergy.2020.114625
Newman D, Bonilla EV, Buntine W(2011) Improving topic coherence with regularized topic models. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems* 2011:1–9
Palma-Rojas S, Caldeira-Pires A, Nogueira JM (2017) Environmental and economic hybrid life cycle assessment of bagasse-derived ethanol produced in Brazil. Int J Life Cycle Assess 22:317–327
Petersen AM, Ahmed ME, Pavlidis I (2021) Grand challenges and emergent modes of convergence science. Human Soc Sci Commun 8(1):1–15
Qian Y, Härdle WK, Chen C (2017) Industry Interdependency Dynamics in a Network Context. SFB 649 Discussion Paper 2017-012, Humboldt University of Berlin. https://doi.org/10.2139/ssrn.2961703
Qi Y, Hao S, Zhang J, Zhao C, Lian Y (2017) Effects of comprehensive nursing on the pain and joint functional recovery of patients with hip replacements. Biomed Res India 28:12
Rafols I, Meyer M (2010) Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics 82(2):263–287. https://doi.org/10.1007/s11192-009-0041-y
Rafols I, Porter AL, Leydesdorff L (2010) Science overlay maps: A new tool for research policy and library management. J Am Soc Inf Sci Technol 61(9):1871–1887
Rapach DE, Strauss JK, Tu J, Zhou G (2015) Industry interdependencies and cross-industry return predictability. Working paper 12-2015. Singapore Management University, Lee Kong Chian School of Business
Rey-Martí A, Ribeiro-Soriano D, Palacios-Marqués D (2016) A bibliometric analysis of social entrepreneurship. J Bus Res 69(5):1651–1655. https://doi.org/10.1016/j.jbusres.2015.10.033
Rizeei HM, Saharkhiz MA, Pradhan B, Ahmad N (2016) Soil erosion prediction based on land cover dynamics at the Semenyih watershed in Malaysia using LTM and USLE models. Geocarto Int 31(10):1158–1177
Rotolo D, Hicks D, Martin BR (2015) What is an emerging technology? Res Policy 44(10):1827–1843
Saadati F, Rahmani M, Ghahramani F, Piri F, Shayani-Jam H, Yaftian MR (2017) Synthesis of a novel ion-imprinted polyaniline/hyper-cross-linked polystyrene nanocomposite for selective removal of lead (II) ions from aqueous solutions. Desalination Water Treat 82:210–218
Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523. https://doi.org/10.1163/187631286X00251
Samsir S, Saragih RS, Subagio S, Aditiya R, Watrianthos R (2023) BERTopic modeling of natural language processing abstracts: Thematic structure and trajectory. J Media Inform Budidarma 7(3):1514–1520
Schumpeter JA (1942) Capitalism, socialism and democracy. Harper and Row, New York
Schumpeter JA (1934) *The Theory of Economic Development*. Harvard Univeristy Press
Shamim A, Abbasi SW, Azam SS (2015) Structural and dynamical aspects of Streptococcus gordonii FabH through molecular docking and MD simulations. J Mol Graph Model 60:180–196
Shin H, Kim K, Kogler DF (2022) Scientific collaboration, research funding, and novelty in scientific knowledge. PLoS ONE 17(7):e0271678. https://doi.org/10.1371/journal.pone.0271678
Sjögårde P (2022) Improving overlay maps of science: Combining overview and detail. Quant Sci Stud 3(4):1097–1118
Small H, Boyack KW, Klavans R (2014) Identifying emerging topics in science and technology. Res Policy 43(8):1450–1467
Song CH, Han JW, Jeong B, Yoon J (2017) Mapping the patent landscape in the field of personalized medicine. J Pharm Innov 12(3):238–248. https://doi.org/10.1007/s12247-017-9283-z
Sugimoto CR, Weingart S (2015) The kaleidoscope of disciplinarity. J Documentation 71(4):775–794. https://doi.org/10.1108/JD-06-2014-0082
Suominen A, Toivanen H (2016) Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. J Assoc Inf Sci Technol 67(10):2464–2476. https://doi.org/10.1002/asi
Velden T, Boyack KW, Gläser J, Koopman R, Scharnhorst A, Wang S (2017) Comparison of topic extraction approaches and their results. Scientometrics 111(2):1169–1221. https://doi.org/10.1007/s11192-017-2306-1
Wang Y, Bashar MA, Chandramohan M, Nayak R (2023) Exploring topic models to discern cyber threats on Twitter: A case study on Log4Shell. Intell Syst Appl 20:200280
Wang Z, Chen J, Chen J, Chen H (2023) Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 0123456789. https://doi.org/10.1007/s11192-023-04776-5
West JD, Jensen MC, Dandrea RJ, Gordon GJ, Bergstrom CT (2013) Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. J Am Soc Inf Sci Technol 64(4):787–801
White K (2019) Publications Output: U.S. Trends and International Comparisons. In Nsb-2020-6. https://ncses.nsf.gov/pubs/nsb20206/
Winnink JJ, Tijssen RJW, van Raan AFJ (2019) Searching for new breakthroughs in science: How effective are computerised detection algorithms? Technol Forecast Soc Change 146:673–686. https://doi.org/10.1016/j.techfore.2018.05.018
Wu W, Zhang S, Wang S (2017) A novel lattice Boltzmann model for the solid–liquid phase change with the convection heat transfer in the porous media. Int J Heat Mass Transf 104:675–687
Xu J, Bu Y, Ding Y, Yang S, Zhang H, Yu C, Sun L (2018) Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. Scientometrics 117:973–995
Xu J, Ding Y, Bu Y, Deng S, Yu C, Zou Y, Madden A (2019) Interdisciplinary scholarly communication: an exploratory study for the field of joint attention. Scientometrics 119:1597–1619
Yau CK, Porter A, Newman N, Suominen A (2014) Clustering scientific documents with topic modeling. Scientometrics 100(3):767–786. https://doi.org/10.1007/s11192-014-1321-8
Zahedi Z, van Eck NJ (2018) Exploring topics of interest of Mendeley users. J Altmetrics 1(1):1–12. https://doi.org/10.29024/joa.7
Zhang J, Zhang G, Zhou Q, Ou L (2016) Thermodynamics, kinetics and isotherm studies on the removal of methylene blue from aqueous solution by calcium alginate. J Water Reuse Desalination 6(2):301–309
Zhao YM, Wang J, Wu ZG, Yang JM, Li W, Shen LX (2016) Extraction, purification and anti-proliferative activities of polysaccharides from Lentinus edodes. Int J Biol Macromol 93:136–144

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1G1A1006464) and by Handong Global University Research Grants (No. 202300710001). Dieter F. Kogler would like to acknowledge funding from the European Research Council (https://erc.europa.eu/) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715631, ERC TechEvo). Further, the authors, Keungoui Kim & Dieter F. Kogler, would also like to acknowledge funding from the Science Foundation Ireland (SFI; https://www.sfi.ie/) under the SFI Science Policy Research Programme (grant agreement No 17/SPR/5324, SciTechSpace). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

The authors confirm contribution to the paper as follows: Conception or design of the work: Keungoui Kim, Sira Maliphol; Acquisition, analysis, or interpretation of data: Keungoui Kim, Dieter F. Kogler, Sira Maliphol; Creation of new software used in the work: Keungoui Kim; Drafted the work or substantively revisions: Keungoui Kim, Dieter F. Kogler, Sira Maliphol; Corresponding Author: Sira Maliphol, correspondence to: sira.maliphol@sunykorea.ac.kr

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Sira Maliphol.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.