This article was accepted by Scientometrics. Please cite this article as:

Liu, X., and Porter, A.L. (2020), A 3-dimensional analysis for evaluating technology emergence indicators, *Scientometrics* 124 (1), 27-55; DOI: 10.1007/s11192-020-03432-6.

A 3-dimensional Analysis for Evaluating Technology Emergence Indicators

Xiaoyu Liu[1], Alan L. Porter[2,3]

Corresponding Author: Xiaoyu Liu, E-mail: Xiaoyu.liu2019@foxmail.com, ORCID: 0000-0003-2509-8457

1. School of Management and Economics, Beijing Institute of Technology, Beijing, 10086, China
2. School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA
3. Search Technology, Inc, Norcross, GA 30092, USA

Abstract: Technology emergence has become a hot topic in R&D policy and management communities. Various methods of measuring technology emergence have been developed. However, there is little literature discussing how to evaluate the results identified by different methods. This research sharpens a promising Technology Emergence Indicator (TEI) set by assessing alternative formulations on three distinct datasets: Dye-Sensitized Solar Cells, Non-Linear Programming, and Nano-Enabled Drug Delivery. Our TEIs derive from a conceptual foundation including three attributes of emergence: persistence, community, and growth that we systematically address through a 3-dimensional evaluation framework. Comparing TEI behavior through sensitivity analyses shows good robustness for the measures.

The TEI serve to distinguish emerging R&D topics in the field under study. They can further be used to identify highly active players publishing on those topics. Importantly, results show that identified emerging terms and topics persist to a strong degree; thus, they serve to predict highly active R&D foci within the technical domain under study.

Keywords: Technology Emergence Indicators; technology forecasting; emerging technologies; technology emergence assessment; R&D emergence; predictive indicators

JEL Code: O32

MSC Code: 28E15

Introduction

Technological emergence has attracted increasing interest in R&D practice, policy, and management circles. Such interest was accelerated by the U.S. Intelligence Advanced Research Projects Activity (IARPA) Foresight and Understanding from Scientific Exposition (FUSE) Program, funded in 2006 [https://www.iarpa.gov/index.php/research-programs/fuse]. Four sizable research teams worked to develop semi-automated methods to assess emergence using scientific and technical publication and patent information resources. Rotolo et al. (2015) digested extensive literature about the concepts and attributes of technology emergence. Drawing especially on FUSE and Rotolo et al.'s results, we focus on four criteria to guide development of a set of Technology Emergence Indicators (TEIs) for R&D literature datasets: novelty, persistence, community, and growth.

Various methods of identifying technology emergence have been developed; we explore these to a degree in the next section. However, there is little literature assessing the technical emergence identified by different methods, how results compare, how well results reflect the concepts and attributes of technical emergence, the emerging technology identified by various methods is emerging or "emerged," and whether the resulting indictors predict future R&D emphases. The basic idea is to investigate whether the identified emerging technology topics warrant that label. Specific concepts and attributes of technology emergence are discussed in the literature review. This study helps fill the gaps by proposing a 3-dimensional evaluation method to assess TEIs. It focuses on three of the four noted dimensions that address tech emergence criteria: persistence, growth, and community (not directly addressing novelty).

Our collaborating research team at Georgia Tech and Search Technology, with Beijing Institute of Technology, uses bibliometric and text mining techniques to generate a two-tier TEI set by distinguishing: 1) emerging topics, then 2) players (individuals, organizations, or countries) actively researching those topics (Porter et al. 2018; Carley et al. 2019). The present research performs sensitivity analyses on our TEI formulations to test their validity and utility (pointing to prediction of future R&D emphases within a target technical domain).

We focus on three TEI dimensions to: 1) take the growth of the dataset into consideration and simplify the growth factor; 2) apply organization-level, instead of author-level, community criteria; and (3) offer enhanced scope criteria to screen out terms that are common in the field. We apply our emergence identification process to three datasets: Dye-Sensitized Solar Cells (DSSCs), Non-Linear Programming (NLP), and Nano-Enabled Drug Delivery (NEDD). These three datasets range over different growth types from slow (NLP) to fast (DSSCs), and draw from different knowledge bases – manufacturing science and engineering (DSSCs), applied mathematics and computer science (NLP), and biotechnology and medical sciences (NEDD). We analyze a ten-year period to identify technology emergence within these research domains. Then we take the following three years as a test period to assess persistence – do the emerging topics go on to evidence relatively higher research activity in the succeeding years? Through comparative evaluation, we seek to capture the development of technology emergence within a target research field in terms of the number of publications, and the emerging research community engaging those topics.

We aspire to make our TEI generation transparent and easy to use. We also seek to provide an assessment framework facilitating comparison of these and other tech emergence indicators. The paper presents this research via upcoming sections on the literature, our data, methods, indicators, evaluation, discussion and conclusion.

Literature Review

Emerging Technologies

Emerging technologies can exert revolutionary impact on science, the economy, and society. To gain full socio-economic benefits and reduce risks, it is incumbent to measure "tech emergence." Clear and early identification of emerging topics can inform R&D researchers in selecting topics of study. TEIs that effectively anticipate future research emphases (short term – i.e., for two or three years) – can inform science and technology managers and policymakers (e.g., in selecting topics most warranting funding). Such TEIs could also provide bases for technology forecasting to pursue practical applications (i.e., commercializing advanced technology).

"What is technology emergence?" is the first question we need to consider before we get to measurement processes. Rotolo et al. (2015) draw upon a deep review of literatures concerning emerging technology and its measurement. They find extensive discussion of the characteristics of emerging technologies, but no widely-recognized definition with clear boundaries of what is, and what is not, "emerging."

Conceptual foundations of tech emergence run deep and broad. Burmaoglu and colleagues (2018; 2019a; 2019b) have traced back three strong roots that underlie consideration of tech emergence –

- philosophy of science (and technology) (c.f., Goldspink and Kay 2010; Sawyer 2001)
- systems theory (complexity) (c.f., Corning 2002; Crutchfield 2013)
- evolutionary economics (c.f., Martin and Sunley 2012; Foster and Metcalfe 2012; Harper and Endres 2012)

Some concepts brought forth include: how creative processes connect to emergence; roles of various interactions (networks); self-organization; novelty and predictability; system properties differing from the sum of parts (as emergence); R&D emergence facets differing from commercial emergence; roles of entrepreneurship; and emergence vis-à-vis radical innovation.

The following paragraph reviews literature on the attributes of technology emergence. We do not deeply pursue the factors or mechanisms that lead to technology emergence. Instead, our method keys on measurable attributes deriving from R&D literature, or patents, concerning a technical domain under scrutiny. We further recognize that emergence characteristics of research literature differ importantly from those observed in patenting; this paper addresses the former.

Martin (1995) analyzes the science and technology foresight experiences of seven countries, and he believes that prominent impact is a very important attribute of emergence. Srinivasan (2008) notes that "the only certainty with emerging technology is the high degree of uncertainty associated with them." He also points out that fast growth, convergence, dominant designs, and network effects are characteristics of emerging technologies. Halaweh (2013) summarizes six characteristics of emerging technology: uncertainty, network effect, costs, unobvious impact, limitation to creator or inventor country, and no full investigation or research. Rotolo et al.(2015) discuss what an emerging technology is and summarize five vital attributes as radical novelty, relatively fast growth, coherence, prominent impact, as well as uncertainty and ambiguity. The FUSE Program emphasizes criteria for emerging indicators deriving from R&D literature, especially: novelty, persistence, community, and growth.

Additional issues include how to promote the development of technology, and how to control the possible risks associated with development. Porter et al. (2002; 2018) see valuable implications of distinguishing technology emergence on the country level. Rotolo et al. (2015) emphasize institutional supports and regulations. Li et al. (2018) explore the relationship between emergent and disruptive technologies from a bibliometric perspective. How to identify emerging technologies at the early stage is of vital importance.

Identifying Technology Emergence

Two contrasting ways to identify technology emergence are through expertise-based qualitative methods or through quantitative text-analytic methods (c.f., Porter and Cunningham 2005). We focus on the latter, empirical approaches.

Some researchers have made efforts to identify technology emergence from a bibliometric perspective (Arora et al. 2013). Cozzens et al. (2010) distinguish concepts of technology emergence identification and measurement, summarizing three examples, and going on to discuss problems in quantitative monitoring attempts. They believe that bibliometric methods have promise in monitoring emergence, while there are also limitations.

An important distinction concerns "scope" of inquiry. Many analyses of tech emergence focus at a macro level, leading to measures that seek to identify "hot" fields. Others focus within-field, seeking to distinguish "hot" topics therein. Issues and methods diverge between these macro and micro inquiries. This paper focuses at the micro, within-domain level.

On the macro level – i.e., treating all of science – Small, Klavans, and Boyack (Small 1999; Small et al. 2014) identify research fronts, or emerging topics, by using citations and co-citations of scientific data. This method emphasizes two important attributes of technology emergence: novelty and growth. Bettencourt and his colleagues (2008) suggest two criteria: the number of publications and the number of authors, as a way to identify emerging scientific fields.

On the micro level, Corrocher et al. (2003) analyze patent data in the area of information and communication technologies and identify emerging technologies as those with significant growth. Kajikawa and his colleagues (2008) track emerging technologies in energy research by using citation network analysis, and they define an emerging technology based on corresponding technology clusters evidencing rapid development. Most of the research on identifying technological emergence only takes one or two attributes as the criteria -- e.g., significant growth in number of publications, the high impact of publications, a growing research community.

There are some exceptions. Breitzman and Thomas (2015) introduce an Emerging Clusters Model, which identifies emerging technologies as those patent clusters that have a significantly higher impact on subsequent technological developments than do patents outside these clusters. They identify four characteristics to identify emerging technologies: public sector proportion, science index, originality index, and reference index. What we need to notice here is, the originality index measures the diversity of cited technologies: patents get higher scores when their prior art comes from a variety of technologies. In addition, Wang (2018) adjusts four attributes of emerging technology: novelty, fast growth, coherence, and scientific impact to identify emerging research topics.

Carley et al. (2018) propose emerging indicators to identify frontier R&D topics and players WITHIN a given technological domain under scrutiny. This method focuses on a specific science and technology domain, mines the R&D publications, or patents, in a certain period, and identifies emerging topics. They then distinguish the players most actively publishing or patenting on those topics. They present the resulting emerging topics and players as Technology Emergence Indicators (TEIs). This method embeds four attributes of technological emergence: novelty, persistence, community, and growth. We generate TEIs first on a topic (or term) level (each term would get a EScore); then on a "players" level (aggregating EScores from term to "player")-- authors, organizations, and countries who stand out as leading the R&D on the emerging topics. Both macro and micro TEI endeavors promise useful perspective on R&D emphases, but at divergent levels of inquiry.

Evaluation of Technology Emergence Identification

We have long recognized that emerging technologies can have major socio-economic-environmental impacts (c.f., Porter et al. 1980). Early identification of tech emergence and of potential impacts of particular emerging technologies is vital in any effort to accentuate the positive effects and minimize the negative as development advances. Constructive Technology Assessment pursues such early engagement (c.f., Guston and Sarewitz 2002; Van Merkerk and Smits 2008).

In most cases, R&D leads technological emergence. Meanwhile, emerging technologies also attract attention from firms, who want to look for future business opportunities (Srinivasan 2008). Cutting-edge research could be one input for technological innovation. Firms could optimize the R&D investment by identifying most promising research topics to pursue. In today's technology-based economy, tracking and obtaining emerging technologies could help firms grow and become competitive, especially, in the catch-up process (Kajikawa et al. 2008; Kim et al. 2017).

Toward that end, development of TEIs is an essential step in early identification of potential emerging technologies. The further steps of "technology assessment" to identify, estimate, and evaluate those possible impacts is beyond our scope here (Porter et al. 1980; Roper et al. 2011). Our intent in considering TEI evaluation is to gauge the validity and utility of such metrics for early identification and possible predictive value in anticipating R&D activities.

Another concern in the identification of emerging topics is whether the item identified by various methods is emerging or "emerged"? That is, does it anticipate particularly active attention in the near future?

How best to assess TEIs? To our knowledge, there is little literature that directly addresses this issue. Here, we step back to consider some possible tacks. This paper addresses the challenge of evaluating candidate emerging topics in terms of future growth in activity.

The predictive utility of emerging technology indicators is an important aspect of assessment. Technology forecasting has developed in different forms; for example, roadmapping, competitive technological intelligence, national foresight studies, etc. (Rader and Porter 2008). Many tools are introduced (Porter 2010), such as, TRIZ, scenario management, bibliometric analysis, and data mining (Coates et al. 2001; Jun and Lee 2012; Daim et al. 2006; Saritas and Burmaoglu 2015). Lots of evaluation criteria for technology foresight methods are applied, e.g. data validity, data availability, implementation cost, ease of operation, method adaptability, technology development predictability, etc. (Sohn and Ahn 2003; Esmaelian et al. 2017; Cheng et al. 2008; Guston and Sarewitz 2002). In future research, it would be interesting to explore how TEIs can contribute to such future-oriented technology analyses.

Our TEI Evaluation Approach

This research focuses on how well TEI results (i.e., R&D topics identified as emerging) reflect the connotation of technology emergence. The basic idea here is to investigate whether the identified emerging technology topics warrant that label. Based on this idea, our attempt starts from examining attributes of technology emergence along three dimensions.

Corresponding to "persistence" attributes of emergence, we take the prevalence of emerging topics in a future period (i.e., the following two or three years) as a measure reflecting persistence (as these terms will evidence strong activity in the previous years). This also provides a way to explore the predictive utility of TEI topics or terms.

The second dimension we address is "growth," which attests to the potential for further development of emerging technologies. Porter et al. (2018) use the average number of records containing each of the emerging terms in the future period as a criterion to evaluate predictive utility of the identified emerging terms. We also introduce the notion of relative growth into the TEI assessment. Relative growth evaluates the growth of emerging research taking into account the growth of the tech domain under study.

As a third evaluation dimension, we take the expanse of the community network treating the emerging topics as a criterion. Metcalfe (1995) notes that a network becomes more valuable as it reaches more users, which is known as Metcalfe's law. Halaweh (2013) strongly agrees with the law, and he points out that the network effect is a characteristic of emerging technology, which means that the value of the emerging technology increases by increasing the number of players engaging it. The community could be considered as a network. Network expansion with respect to particular emerging technologies could portend future activity and further development. Given our "micro" level focus (within a target technological domain), operationalizing community with respect to particular identified emergent topics (or terms) is in order.

Novelty, an important attribute of technology emergence, is not a dimension in our evaluation method. Our emergence scoring of candidate topics operationalizes "novelty" as low frequency in a base period relative to the following active period. This approach seems solid. Terms/topics identified as emerging by our metric will, by definition, be precluded from being "new" in the following test time period.

This study explores ways to measure TEIs based on the nature of technology emergence. That is, we consider the four criteria of novelty, persistence, community, and growth, and how to operationalize them – focusing our evaluation on three of those – persistence, community, and growth. Porter et al. (2018) put forth our TEI methodology with results for three case analyses that provide a base for improvement.

Data

Analysis Datasets

Information on science and technology is organized in different channels and forms. Cozzens et al. (2010) discuss the data sources that could be used in identifying technology emergence, including web pages, awards, funding, scientific publications, patents, and business activities. [R1.2.4] They believe the data source should be selected considering coverage, biases, content quality, record structure, and keyword availability. Multiple information resources offer multiple appealing aspects, but also various disincentives for use.

Data from web pages would not necessarily contain specific statement of the emerging techniques, while they may inform to a degree on contextual issues, extending to economic and social effects (so highly pertinent to technological innovation). R&D award compilations or funding data (e.g., acknowledgements in articles) may work as a signal of early funder prioritization and of critical resource investment. However, analyses of funding information resources won't provide a systematic indication of R&D outputs.

Patent data are of special interest. One could expect that some degree of research results will be advanced through patenting to protect intellectual property. Patent analyses may well inform on technological emergence and commercial potential, however also with limitations (e.g., certain technological sectors don't patent heavily). In other studies, we are trying the emergence scoring routine on patent datasets. The present study focuses, instead, on R&D literature analyses, perceiving those as generally "upstream" to patent analyses.

The earliest indication of new high technology capabilities has high possibility to appear in scientific publication. What's more, scientific publication abstract records, compiled in databases such as Web of Science (WOS), offer advantages of wide access; well-characterized content, wide-spectrum coverage; informative text concentrations (titles, abstracts, keywords); and definable, distinguishable research domains.

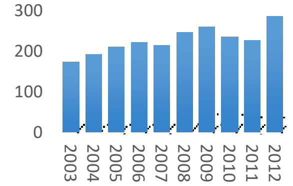
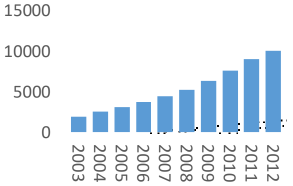
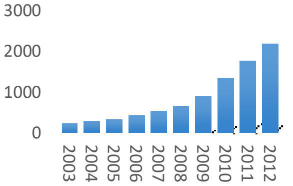
Our testing is applied to three cases. Non-Linear Programming (NLP) is a relatively mature research field related to applied mathematics and computer science. Dye-Sensitized Solar Cells (DSSCs) is a promising and rapidly growing research field whose knowledge base is materials science. Nano-Enabled Drug Delivery (NEDD) is an emerging sub-nanotechnology field that engages medicine sciences, pharmaceutical sciences, and biotechnology. It complements the other two datasets since it covers different research areas and provides a relatively large dataset, which could help test our methods for scalability. These datasets have also been used in previous research by our colleagues, so we gain some opportunity to compare results.

For NEDD and DSSCs, we reapplied the search strategy developed in previous research to download the data from the WOS database (Guo et al. 2012; Zhou et al. 2014). For NLP, we developed the dataset by using a straightforward WOS topic-based search. We take ten years as our validation period to generate emerging terms and indicators, and then take the following three years as the test period to track further development of the emergence we detected.

Table 1 describes the three datasets. The NLP dataset is the smallest, with 2,260 records in the validation period and 965 records in the test period. NEDD is the biggest dataset with 53,957 records, plus 38,559 records in the following period. The DSSCs dataset falls in between in size. Note that in the growth trend figures in Table 1, the three datasets have distinctly different growth patterns.

We separate the 10-year validation period into a base period (the first three years) and an active period (the last seven years). To describe the growth trend for each dataset, we devised a new factor to consider – a growth coefficient, which is the ratio of the total number of records in the active period to the total number of records in the base period. The more rapid the growth, the bigger the growth coefficient.

Table 1. R&D Emergence Indicator Test Datasets (drawn from the Web of Science)

<table>
<tr><th>Dataset</th><th># Records in the Validation Period (2003-2012)</th><th>Growth Trend Plot</th><th>Growth Coefficient</th><th># Records in the Test Period (2013-2015)</th></tr>
<tr><td>Non-Linear Programming (NLP)</td><td>2,260</td><td>
</td><td>2.93</td><td>965</td></tr>
<tr><td>Nano-Enabled Drug Delivery (NEDD)</td><td>53,957</td><td>
</td><td>6.16</td><td>38,557</td></tr>
<tr><td>Dye-Sensitized Solar Cells (DSSCs)</td><td>8,544</td><td>
</td><td>9.39</td><td>8,080</td></tr>
</table>

After retrieving our data from WOS, we clean (consolidate) the data. Term cleaning is a very important step for the preparation of identifying technology emergence in terms of emerging terms (or topics). The clean-up process is performed in VantagePoint (www.theVantagePoint.com). Details on the data cleaning process for terms are provided in the Appendix. For the author affiliation (organization) field and author field, we applied tailored fuzzy matching processes (affiliation.fuz and author.fuz) in *VantagePoint* to help consolidate name variations.

Random Datasets

Besides the three analysis datasets described, we also developed random WOS datasets. For each year of the validation period, we randomly selected 5,000 records from the WOS database. Specific steps for drawing random sample datasets are described in the Appendix.

We developed a 5000-record random dataset for each year of the validation period. These datasets were used as comparison groups for the cross-dataset Inverse Document Frequency (IDF) test. For random sample datasets, we also cleaned the terms the same way as for the analysis datasets. Specific steps are listed in the Appendix.

Methodology

Tech emergence indicators seek to operationalize components embedded in definitions of technological emergence, such as persistence, novelty, growth, community, and scope. We devised our version of TEIs and gauge this by comparing behaviors across the three cases; we don't address "micro" vs. "macro" performance.

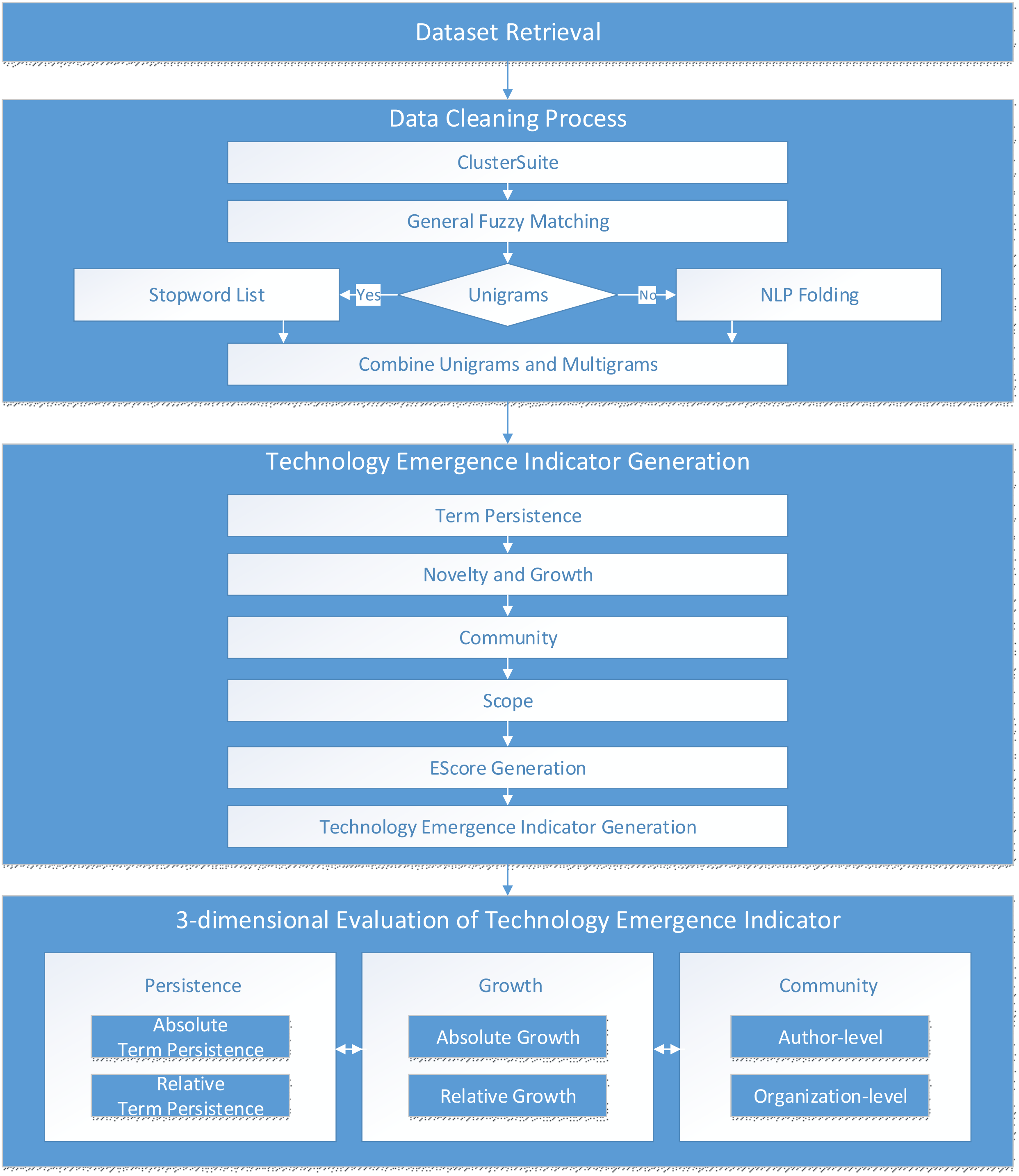


Fig. 1 Technology Emergence Indicator Generation Flowchart

As per the title of this paper, our evaluation of TEIs keys on three of the emergence criteria. As formulated (Carley et al. 2018), calculation of the TEIs sets out thresholds for "persistence" and "community"; we go on to perform sensitivity analyses on those attributes here. "Growth" can be operationalized in several components. Our TEI scoring is based mainly on a composite trend calculation, and that is examined in this study. Fig. 1 offers a schematic to array the data cleaning, generation of emergence indicators (i.e., emerging terms & topics), and evaluation foci of the process.

We only show the sensitivity analysis figures for the NEDD dataset in the manuscript; the corresponding sensitivity tests of the other two datasets are available in the Supplementary Materials. Fig. 2 presents the sensitivity testing of term persistence on two dimensions: number of records and number of year instances, while the vertical axis shows the share of terms meeting the criteria. We mark the default value as the red spot in the figure, which represents the standard of 3 years and 7 records. From Fig. 2, we can see that number of records is a relatively sensitive threshold while number of year instances is not.

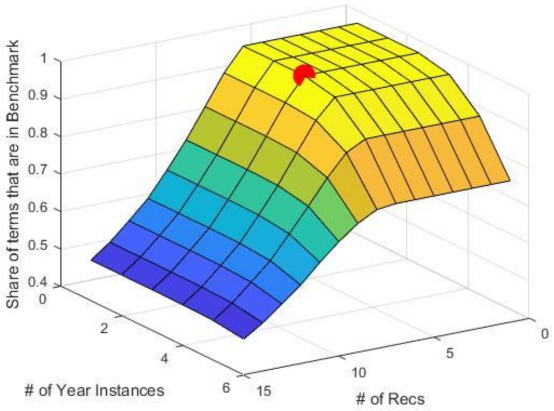


Fig. 2 Sensitivity Test of Term Persistence Criterion in NEDD dataset.

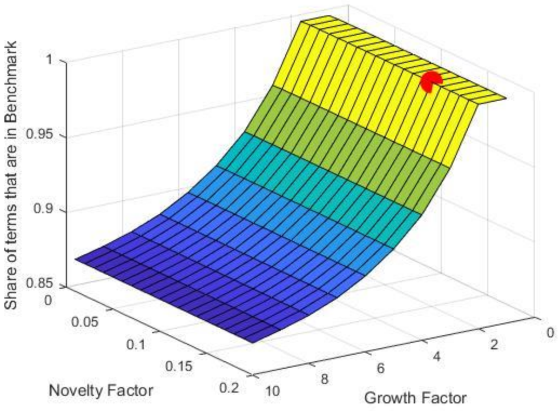


Fig. 3 Sensitivity Test of Novelty and Growth Criteria in NEDD dataset.

For a term to be included as a candidate for emergence calculation, we address "novelty" by checking that it was not common in the base period. Given this formulation seems to perform well, we don't further address novelty in this evaluation endeavor.

In our method of identifying technology emergence as originally presented (Carley et al. 2018; Porter et al. 2018), the "growth factor" was defined as the ratio of the number of records containing at least one instance of the particular term in an active period to the number of records in a base period. For a specific term, if the ratio were more than 2 (the default value), then the term would meet the growth criteria. However, the growth of the study field could differ considerably, affecting the expectation for growth by a given term. As shown in Table 1, NLP is a slow-growing field with a growth coefficient of 2.93, while DSSCs is a fast-growing technological domain with a growth coefficient of 9.39. So, a term appearing four times as often in the active period as in the base period could be considered to be emerging in the NLP dataset, for it exceeds the general NLP growth rate. However, in the DSSCs dataset, this term should not be identified as emerging since it does not meet the average domain growth rate.

Thus, data analysts need to find suitable parameters for each case. However, determining how to adjust the parameter of the growth factor may be a problem in practice. Data analysts may need to adjust the parameters over a wide range to find the proper threshold if using the previous growth criteria. To make our emergence indicator flexible enough for datasets with different growth trends, we associated the growth factor with the growth of the analysis field. Cozzens et al. (2010) noted that "if the number of associated papers or patents is going up faster than the average, the group may be said to be emerging." We defined the growth coefficient as the ratio of the total number of records in the active period to the total number of records in the corresponding base period, which is the average growth for all terms in the dataset under study. It provides the information for the growth trend of the study case. For a specific term, the growth of the term should be faster than the average growth of the total terms, then we would say that the term could qualify as an emerging term. In this evaluation study, once the growth factor of the term is 1.5 times more than the growth coefficient of the technological domain, the term meets the growth criteria. For example, the term meets this growth criterion if its growth factor is more than 1.5*6.16=9.24 in the NEDD dataset. The key advantage of this revision is that the parameters do not need to be readjusted for each different dataset. Fig. 3 shows the visualization results of sensitivity tests of novelty and growth criteria. We can see that both criteria are not sensitive: novelty threshold ranges from 0.01-0.2 and growth threshold ranges from 1.5-9.5, the share of terms that meet both criteria only drop less than 15%. This may be because of the long-tailed distribution of term instances. Specifically, there would be lots of low-frequency terms only appearing in the active period, so when we increase the threshold of the growth criterion, the share of terms does not drop significantly.

How to measure community in a precise and timely way is a question we haven't resolved. Previously the "community criterion" was specified for authors (Carley et al. 2018). Specifically, a term met that community criterion if there were more than one author, who doesn't share a same record set, using the term in a publication abstract record. However, that approach builds a complex network for each term. Name disambiguation is also a challenging task -- it is difficult to disentangle authors with the same last names. Many factors complicate this – e.g., common names, use of initials, non-standard use and ordering of first and middle and last names (Carley et al. 2019). This prompted us now to explore use of organizations (author affiliation) instead of authors to discern community status.

As for the organization level, it is relatively easy to deal with name disambiguation for organizations. Second, the number of organizations is almost always less than the number of authors in a dataset, facilitating both automatic and manual data clean-up processes. Third, the number of organizations using the term gives more interesting information on the spread of usage of a term. Multiple organizations using a term is stronger evidence of its recognition beyond those who instigated its use than would be multiple authors. We conduct the sensitivity test of community criterion. Fig. 4 shows the share of terms that meet the community criteria when we adjust the criterion of the number of organizations, varying from 2 to 10. This is a sensitive criterion -- increasing 1 unit causes an 8% decrease of the terms, on average.

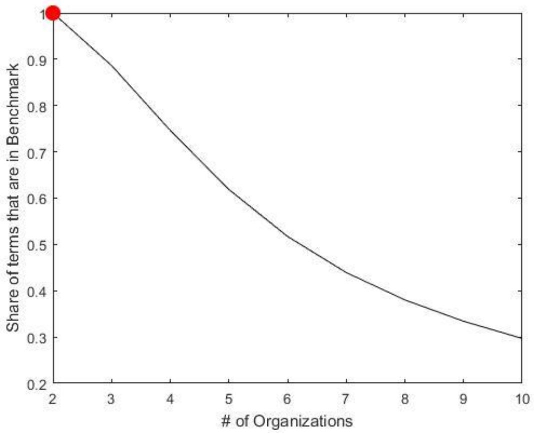


Fig. 4 Sensitivity Test of Community Criterion in NEDD dataset.

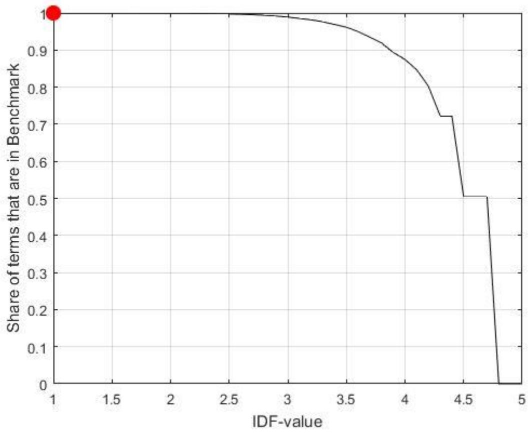


Fig. 5 Sensitivity Test of Within-Dataset IDF Filter (Scope Criterion) in NEDD dataset.

For the identification of emerging terms, some terms are too general to constitute interesting emerging topics (Cozzens et al. 2010). For the scope criteria, we introduce alternative means to screen out very common terms in the dataset. First, we remove the terms with the highest frequency in our dataset by using a within-dataset filter. Here we calculate the Inverse Document Frequency (IDF) value for each term and screen out the terms with extremely low IDF values, which means the term is very common. Fig. 5 shows the distribution of IDF-value of terms in NEDD dataset. The IDF value is not sensitive from 1 to 3. An IDF-value of 1 is recommended as a conservative threshold, and terms that do not meet this criterion would be prevalent terms in the field. For example, "dye sensitized solar cell" has an IDF value of 0.1 in our dataset, which means it appears in 79% of the records. This term is the name of the technology, as well as a component in our search strategy. We don't want to include it as an emerging topic within the DSSCs field; essentially it approximates the field growth rate. So, in this case, a within-dataset filter is used to exclude such terms from the emerging terms set.

Second, we seek to screen out common terms in scientific publications generally. We apply a cross-dataset comparison for an analysis dataset and a suitable random sample dataset. Based on the idea that the more common the term is, the higher the IDF-value is, we screen out the terms with higher IDF values in an analysis dataset than in a random sample dataset. The removed terms could be common scientific phrases or common techniques, such as "take advantage," "shed light," and "significantly different." This step could also be a supplement for the stopword screening because stopword lists typically contain common scientific unigrams and do not work well on scientific phrases. Multigrams, not screened out by a stopword list, could be removed by applying such a cross-dataset IDF comparison.

In calculating Emergence Scores (EScores), we previously gave multigrams higher weight than unigrams because multigrams may contain "more information." However, the weighting difference may cause some biases. First, we doubt whether multigrams uniformly contain "more information." For example, the term "DLS" will have less weight than the term "dynamic light scatter" in the previous method, even though they refer to the same technique. It seems unreasonable to give unigrams and multigrams different weights; some unigrams are abbreviations for a multigram. Second, we question if "more information" actually means "more emerging." Even if the multigram possibly consolidates related phrases, the information it contains is not equal to emergence. Lastly, differential weighting may bias those whose native language is not English. A brief test on an autonomous vehicle dataset says that native English-speaking countries use more unigrams, while non-native English-speaking countries use more multigrams. In this case, an EScore will benefit the non-native-English speaking countries. Thus, this evaluation suggests weighting unigrams and multigrams equally in the calculation of EScores.

After we calculate EScores for candidate terms, we seek to set a suitable threshold to distinguish high emergence score terms (ETs). We consider all the terms that pass all criteria as ET candidates. To set the threshold, we first thought about a percentile selection method. We attempted to choose the ETs with a top 5 percentile EScore among the candidates in a given case, instead of using a fixed value threshold. The top 5 percentile EScore is 2.14 for NEDD, while it is 4.17 for NLP, and 6.62 for DSSCs. A term with an EScore of 3 will be identified as an ET in NEDD, but not in DSSCs or NLP. This phenomenon is thought-provoking. The number of ETs will be decided by the number of candidates. The numbers of candidate ETs are 258 for NLP, 9221 for NEDD (as the much larger dataset, thus, with many more terms), and 737 for DSSCs. Relatively, the number of ETs would be 13, 462, and 37, respectively. Thirteen might be too small for the technology emergence analysis. A fixed threshold may overcome these shortcomings.

First, we capture the share of terms that remains when we increase the EScore threshold from 1.77 to 6. From Fig. 6, we can see that the terms are very sensitive to the EScore threshold. Futhermore, we followed the idea of our previous work by using the predictive utility as a way to set the threshold (Porter et al. 2018). We classified the candidates into four groups according to their EScores. Since these candidate terms meet the persistence, novelty and growth, community, and scope criteria, they could be used for comparing the predictive utility of ETs. Even a term with an EScore of less than 0 (reflecting negative trend slopes) could be a promising term as well. Table 2 shows the predictive utility of the candidate emergent terms in each of the three datasets.

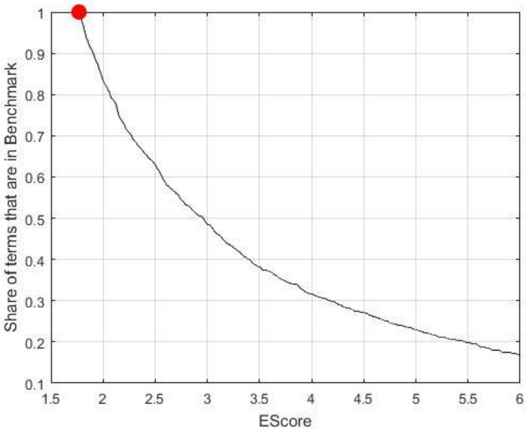


Fig. 6 Sensitivity Test of EScore Threshold in NEDD dataset.

Table 2. Predictive Utility of Candidate Emergence Terms in Three Test Datasets.

Dataset		EScore<0	0<EScore<1	1<EScore<1.77	EScore>1.77
NLP	#	123	58	21	56
Test Period	2013-2015	5.77	5.50	5.57	10.30
Prior Period	2010-2012	4.63	5.52	5.29	9.00
NEDD	#	2802	5092	722	605
Test Period	2013-2015	10.93	14.00	40.76	58.94
Prior Period	2010-2012	7.65	10.21	27.12	43.61
DSSCs	#	174	206	153	204
Test Period	2013-2015	10.37	11.58	14.27	65.22
Prior Period	2010-2012	8.93	8.53	10.63	38.11

For each dataset, the first row shows the number of candidate terms fitting the criteria designated for the respective columns, while the following rows give the average number of records containing each of those candidates in the test period and in the prior period. For example, for the NLP dataset, 123 candidate terms have an EScore of less than 0, and they appear in 4.63 records in 2010–2012 on average, while they appear in 5.77 records in 2013–2015. This shows an expansion of the emerging topics. From Table 2 we see that the average number of records in which terms appear low are in groups with an EScore of lower than 1.77. Terms with an EScore of more than 1.77 usually appear in more records in the prior period (last 3 years of the active period), and also achieve rapid growth in the test period. Rather than use a percentile threshold, we thus chose 1.77 as the default threshold. Users are welcome to adjust it for their specific needs[1]. Corresponding to the threshold of 1.77, the numbers of ETs for NLP, NEDD, and DSSCs are 56, 605, and 204, respectively.[2]

To sum up, the criteria to identify emerging terms and Tech Emergence Indicators are listed in Table 3.

Table 3. Steps of TEI Generation.

Step		Proposition	Mathematical notation
I	Term Persistence	A term must appear in at least three time periods (years).	$\sum_T x_{it} \geq 3,\ x_{it} = \begin{cases} 1 & \text{if term i appear in time period t} \\ 0 & \text{otherwise} \end{cases}$
		A term must appear in at least seven records in the active period.	$\sum_{t=4}^{10} n_{it} \geq 7,$ n_{it} is the number of records contain term i in time period t
II	Novelty and Growth	The term cannot appear in as many as 15% of the base period records.	$\frac{\sum_{t=1}^{3} n_{it}}{\sum_{t=1}^{3} N_t} \leq 0.15,$ N_t is the total number of records in time period t
		The growth of the terms should be 1.5 times the growth of the total records.	$\frac{\sum_{t=4}^{10} n_{it}}{\sum_{t=1}^{3} n_{it}} \geq 1.5 \times \frac{\sum_{t=4}^{10} N_t}{\sum_{t=1}^{3} N_\mathrm{t}}$
III	Community	Terms should be used by more than one organization.	$\sum_J \mathrm{y}_{ji} \geq 2,\ \mathrm{y}_{ji} = \begin{cases} 1 & \text{if organization j use term i} \\ 0 & \text{otherwise} \end{cases}$
IV	Scope[3]	(Within-dataset filter) Terms should have an IDF-value higher than 1.	$\mathrm{IDF}_i \geq 1,$ IDF_i is the IDF value of term i in analysis dataset.
		(Cross-dataset comparison) Terms should have a lower IDF-value in an analysis dataset than in a random dataset.	$\mathrm{IDF}_i \leq \text{IDF-random}_i,$ IDF-random_i is the IDF value of term i in random dataset.
V	EScore Generation	An EScore for each term is calculated by summing up two times	$\text{Active Trend}_i = \sum_{t=8}^{10} \frac{n_{it}}{\sqrt{N_t}} - \sum_{t=4}^{6} \frac{n_{it}}{\sqrt{N_t}}$

[1] The Emergence script provided in *VantagePoint* enables one to vary many parameters of EScoring. To alter the 1.77 threshold (chosen based on empirical comparisons – see Porter et al. 2018), one just needs to scan the list of terms ordered by EScore in *VantagePoint* or *MS Excel*.

[2] We also tried calculations based on the later two-year test period and last two years of the validation period as the prior period to test the predictive utility. Results are similar.

[3] As of January, 2020, we have adjusted the specialized scope procedure. We abandon the cross-dataset comparison as problematic in requiring a representative population or random sample outside the test dataset. We tighten the within-dataset filter to use IDF >= 1.

		the active trend, recent trend, and slope.	$\text{Recent Trend}_i = (\sum_{t=9}^{10} \frac{n_{it}}{\sqrt{N_t}} - \sum_{t=7}^{8} \frac{n_{it}}{\sqrt{N_t}}) \times 10$ $\text{Slope}_i = (\frac{n_{i10}}{\sqrt{N_{10}}} - \frac{n_{i7}}{\sqrt{N_7}}) \div 3 \times 10$ $\text{Escore}_i = 2 \times \text{Active Trend}_i + \text{Recent Trend}_i + \text{Slope}_i$
VI	TEI Generation	Terms with an EScore of no less than 1.77 are selected as emerging terms.	$\text{Escore}_i \geq 1.77$

Technology Emergence Indicators

We apply the original method and our proposed new approach to three WOS datasets: NLP, NEDD, DSSCs. Table 4 shows the ETs for each dataset. Those are generated by applying the EScore script in *VantagePoint* to the treated abstract and title NLP phrases (see Appendix), and then cutting off ETs at the 1.77 EScore threshold. We now compare the results for each dataset.

Table 4. Number of Emerging Terms in NLP, NEDD, and DSSCs datasets.

Dataset	New Method	Previous Method	Overlapping Terms
NLP	56	49	43
NEDD	605	572	533
DSSCs	204	227	174

By using the default threshold, we get ETs for the NLP field. There were 56 emerging terms identified by our revised method versus 49 emerging terms identified by the previous method. The overlap contained 43 terms. The top 10 emerging terms for those two methods are listed in Table 5, and the full list of ETs are provided in the Supplemental Materials. "Production" gets the highest EScore of 8.31, followed by "Mixed Integer Non-Linear Program MINLP" with 6.75, and "Operating Cost" with 6.34. "Production" and "manufacturer" are two ETs with a high EScore, which suggests a shift of emerging focus from basic research to applied research in the NLP field. We note that terms such as these are not in themselves novel, but their increasing prevalence in abstract records over time offers interesting intelligence concerning research priorities.

In the NEDD dataset, there are 605 emerging terms identified by our new method versus 572 ETs identified by the previous method. They share 533 terms. The top 10 emerging terms for those two methods are listed in Table 5. Through the new method, "cancer cell" gets the highest EScore of 39.36, followed by "cellular uptake" with 36.44, and "potential zeta" with 34.91. In the previous method, the top two terms were "delivery drug" with an EScore of 85.30 and "Nanoparticle" with an EScore of 81.42. Records containing "nanoparticle" make up 10.13% of the total dataset, while records containing "drug delivery" make up 8.76%. These terms are the name of the "nano-enabled drug delivery" technology, as well as being included in the search strategy. They do not provide useful information to capture the Emergent Topics (ETopics) of this field. By applying scope criteria, these terms with extremely low IDF values are screened out. The new formulation is superior in this regard. Among the top 10 ETs, “cancer therapy” and “nanomedicine” give clues about the translation from TEI to medicine and products.

Table 5. Top 10 Emerging Terms in NLP, NEDD, and DSSCs Datasets.

Topic	New Method			Previous Method		
	Rank	Terms	EScore	Rank	Terms	EScore
NLP	1	production	8.31	1	Mixed Integer Non-Linear Program MINLP	6.75
	2	Mixed Integer Non-Linear Program MINLP	6.75	2	non-linear function	6.40
	3	operating cost	6.34	3	operating cost	6.34
	4	proposed algorithm	6.31	4	linear behavior	5.55
	5	linear behavior	5.55	5	model results	5.10
	6	consideration	5.42	6	non-linear response	4.96
	7	model results	5.10	7	particle swarm	4.91
	8	manufacturer	4.95	8	heuristic algorithm	4.77
	9	particle swarm	4.91	9	production	4.73
	10	Genetic Programming	4.88	10	decision variable	4.38
NEDD	1	cancer cell	39.36	1	delivery drug	85.30
	2	cellular uptake	36.44	2	nanoparticle	46.05
	3	potential zeta	34.91	3	cancer cell	39.36
	4	biocompatible	34.66	4	cellular uptake	36.44
	5	NP	29.45	5	potential zeta	34.91
	6	biomedical application	27.48	6	biomedical application	27.48
	7	drug delivery system	26.64	7	drug delivery system	26.64
	8	self assembly	26.41	8	self assembly	26.41
	9	cancer therapy	24.77	9	particle size	24.82
	10	nanomedicine	24.74	10	cancer therapy	24.77
DSSCs	1	photoanode	46.96	1	dye sensitized solar cell DSSC	134.58
	2	counter electrode	35.43	2	DSSC	77.63
	3	Zinc oxide	35.04	3	counter electrode	35.43
	4	power conversion efficiency	32.92	4	Zinc oxide	35.04
	5	titanium tetrachloride	19.89	5	power conversion efficiency	32.92
	6	electron lifetime	18.36	6	photoanode	27.17
	7	organic dye	17.69	7	dye-sensitized solar cell (DSC	21.35
	8	photophysics	15.32	8	titanium tetrachloride	19.89
	9	field emission	13.85	9	electron lifetime	18.36
	10	Pt	13.26	10	organic dye	17.69

After comparing these two datasets, we observe that the NLP dataset is a small dataset, containing 2,260 records in a ten-year period. In contrast, the NEDD dataset is relatively big, containing 53,957 records. The NLP dataset is a relatively mature science domain with a growth coefficient of 2.93. NEDD is an emerging technology domain with a growth coefficient of 6.16, which is more than twice the growth coefficient for NLP. According to the growth of the two research domains, it is easy to understand the difference in size and score for emerging terms within these two fields: there are 56 ETs in NLP, versus 605 in NEDD. The ET with the highest EScore of 8.31 is "production" in NLP, while "cancer cell" gets the highest EScore, 39.36, in NEDD. The ETs in the bigger dataset with rapid growth are more likely to achieve a higher EScore. The average EScore for ETs in NEDD is also higher than that in the NLP dataset.

In the DSSC dataset, there are 204 ETs identified by our new method versus 227 ETs identified by the previous method, of which 174 ETs overlap. The term "photoanode" gets the highest EScore of 46.96, followed by "counter electrode" with 35.43, and "zinc oxide" with 35.04. In the previous method, the terms "Dye Sensitized Solar Cell DSSC," "DSSC," and "Dye-Sensitized Solar Cell (DSC) had three positions in the top 10 ETs with the highest EScores. We consider them as useless terms in analyzing technology emergence within the domain. Through the new method, these terms are screened out, and data analysists could focus on other emerging topics, such as "photophysics" or "photoanode." The growth coefficient for DSSCs is 9.39—the highest among the three datasets. DSSCs also have the highest EScore for ETs, although the size of the DSSC dataset is not as big as that of NEDD.

These comparisons favor the new growth criteria. For different dataset sizes, appealing ETs could be obtained without adjusting the growth parameters, over a wide range -- here, from NLP containing 2,260 records to NEDD with 53,957 records. Our proposed growth criteria avoid the problem of finding a suitable threshold for each tech domain under study, since we incorporate the growth rate of the domain into our process. By applying the scope criteria, we can screen out terms with high frequency of occurrence, and, additionally, screen out the terms that are common in scientific publications generally.

We mention that emerging terms are the basic units of TEIs; thus, we analyze primarily on this level. It is vital to examine the performance and robustness of TEIs first at the level of emerging terms. These are of a highly specific nature, such that many will overlap to a degree and be too specific to interest others than researchers deeply immersed in the domain under study. However, aggregation of those terms into emerging topics can follow, and that can provide interpretable results of much wider interest. Emerging topics can be better assessed in terms of their attributes and implications. And, as noted, relative concentrations of emerging terms/topics in research activity collections can be useful in distinguishing which authors, organizations, or countries are highly active in those frontier elements within the domain being studied. Those lie beyond the scope of this paper; see Porter et al. (2018).

3-Dimensional Evaluation of Technology Emergence Indicators

The prominence of emergent terms/topics, and the second order indicators of players most engaged with those, is a vital attribute, but it cannot be evaluated immediately. Here, we try to measure the predictive utility of the emerging topics, as evidence of likely science and technology importance. "Technology (or science) growth can take place in many pertinent dimensions: within the technology space overall or of various components; into other technology spaces; within the R&D community" (Porter et al. 2018). As stated, this article keys on three dimensions: persistence, growth, and community. A 3-dimensional method is proposed to evaluate the predictive utility of TEIs.

Persistence

Persistence is an important attribute in technology emergence, relating to technology forecasting as well. For the persistence criteria, we check whether the emerging topics persist in the future and how many ETs achieve growth in publication activity. We take the last three years in the validation period, 2010–2012, as the prior period for this purpose, and the three years after the end of the validation period as the test period, 2013–2015. For the NEDD and NLP datasets, all ETs appear in the following test period, and only one term in the DSSCs dataset does not. This shows that ETs identified by our method have good temporal persistence, and the average share persisting is 99.8%.

As for term persistence, we measured it by absolute term persistence and relative term persistence. Absolute term persistence refers to the number of ETs appearing in more records in the test period than in the prior period. NEDD and DSSCs have better performances than NLP; more than 92% and 82% of the ETs appear in more records in the test period than in the prior period in the NEDD and DSSCs datasets, respectively, while just more than half of the ETs grow in NLP. This is not as high as in the other datasets, but it is understandable since NLP research is a relatively mature field. Relative term persistence refers to the number of ETs appearing in a higher share of records in the test period than in the prior period. Accordingly, the NEDD and DSSC datasets achieve better results -- 57.2% and 53.4% of the ETs show relative increase, while 39.3% of ETs achieve relative increase in the NLP dataset. We also checked the persistence of ETs generated by the previous method, and the results are generally similar to the new calculations.

Table 6. Persistence Dimension for Evaluating TEI.

Dataset	Absolute Term Persistence		Relative Term Persistence	
	#ETs	**Share**	**#ETs**	**Share**
NLP	32	57.14%	22	39.29%
NEDD	560	92.56%	346	57.19%
DSSCs	169	82.84%	109	53.43%

Note: Absolute term persistence is measured by the number of ETs appearing in more records in the test period than in the prior period. Relative term persistence is measured by the number of ETs appearing in a higher share of records in the test period than in the prior period.

Growth

Growth (in the form of three research activity trends) constitutes our technology emergence scoring, as well as being a meaningful aspect in evaluating emerging terms/topics. Here we take two kinds of growth into consideration. One is absolute growth, measured by the average number of records containing ETs. The basic idea is that researchers are going to use ETs in their future research, which will cause growth in the number of publication abstract records in which ETs appear. The other criterion is relative growth, which is measured by the share of records containing ETs in the total datasets. Terms with high relative growth value mean they account for a greater proportion than other terms; furthermore, they are more promising as ETs of real interest.

Table 7 provides information on the growth dimension for predictive utility. We use the last three years in the validation period and the three-year test period in the comparison. We calculated their absolute growth and relative growth for each dataset. For example, in the NLP dataset, 9 records contained 1 ET on average in 2010–2012; 10.3 records contained 1 ET on average in 2013–2015; and the "Growth%" was

(10.3-9)/9=14.5%. We see that all three datasets had significant absolute growth, especially in the case of NEDD and DSSCs. In the DSSCs dataset, the average number of records in which an ET appears in the prior period was 38.1, rising to 65.2 in the test period, giving a growth rate of 71.1%. The NEDD publication rate grew substantially, 64.2%. For all the datasets, the new method performed slightly better than the previous method in terms of absolute growth.

Table 7. Growth Dimension for Evaluating TEI

Dataset	Absolute growth			Relative growth		
	2010–2012	2013–2015	Growth%	2010-2012	2013-2015	Growth
NLP	9	10.3	14.48%	39.95%	36.79%	-3.16%
NEDD	116.78	191.69	64.15%	77.49%	81.99%	4.50%
DSSCs	38.11	65.22	71.13%	68.66%	72.95%	4.29%

Note: Absolute growth is calculated by the average number of records in which a given ET appears in a certain time period. "Growth%" is calculated by the relative growth rate of absolute growth. Relative growth is measured by the share of number of records in which ET appears as a ratio over the total number of records in a certain time period. "Growth" of relative growth reflects the relative growth share in the prior period subtracted from the relative growth share in the test period and.

Relative growth measures the growth of emerging research in the total dataset normalized for the size and growth of the total dataset. For the NEDD dataset, all the records containing ETs identified by the new method made up 77.5% of the total dataset in 2010–2012 and 82% in 2013–2015, which shows an expansion of 82%-77.5% = 4.5% in research activity. The last column of Table 7, "Growth" of "Relative growth," is a simple subtraction. The emerging research for DSSCs sees an expansion of 4.3% in the total DSSCs scientific area. The mature NLP dataset is the only one for which the share of abstract records mentioning ETs shrinks. For all three datasets, the new method performs slightly better than the previous method.

Community

Growth can be measured along different dimensions. Besides what we have discussed, the extent of authors or organizations conducting research on particular topics could speak to their potential as emerging topics. In the community dimension, we could take author-level expansion and organization-level expansion into consideration, then compare the expansion of the ET group with the non-ET group. Table 8 lists the number of authors in a 10-year validation period, 2003–2012, and a 3-year test period, 2013–2015, and specifically, the number of newcomers whose first emerging topic research in these domains appears in the test period (i.e., they do not conduct related research in the prior validation period). For example, in the NLP dataset, there are 1,796 authors and 1,211 authors participating in emerging research during the validation period and the test period, respectively, and 1,126 out of 1,211 authors are newcomers. Compared to the validation period, the size of the author group conducting emerging research expands to 1126/1796=62.7% in three years. As for the non-ET group, there is also an expansion of 42.3% in terms of the number of authors; note that it is less than the expansion of the ET group.

For the comparison of number of authors, the ET-group has a larger expansion than the non-ET group in all three datasets, which means that emerging research, so identified, is more attractive than other research topics within the technical domain. Comparing the three datasets, we found that DSSCs attract the highest share of newcomers to both the ET group and the non-ET group. This may relate to its evidencing the highest growth rate among the three datasets.

Table 8. Community Dimension for Evaluating TEI on an Author Level

Dataset	# Authors in ET Group				# Authors in Non-ET Group			
	2003–2012	2013–2015			2003–2012	2013–2015		
	Total	Total	Newcomers	Expansion	Total	Total	Newcomers	Expansion
NLP	1796	1211	1126	62.69%	4312	1964	1825	42.32%
NEDD	106744	100612	71224	66.72%	67977	35905	27674	40.71%
DSSCs	12213	15108	10944	89.61%	8176	7441	5599	68.48%

Note: "Expansions" are calculated by the ratio of number of newcomers in the test period to the number of authors in the validation period.

This article also examines community expansion on an organizational level. Table 9 gives the details of the number of organizations for the two time periods and three datasets. Consistent with the author-level community expansion, the ET group also "wins" on the organization-level: the expansion share for the ET group is higher than the expansion share for the non-ET group for all three datasets. Also, the DSSCs dataset attracts the highest share of new organizations (Table 9).

Table 9. Community Dimension for Evaluating TEI on the Organization Level

Dataset	# Organizations in ET Group				# Organizations in Non-ET Group			
	2003–2012	2013–2015			2003–2012	2013–2015		
	Total	Total	Newcomer	Expansion	Total	Total	Newcomer	Expansion
NLP	736	479	296	40.22%	1551	885	526	33.91%
NEDD	9530	9634	5385	56.51%	7050	4801	2364	33.53%
DSSCs	1768	2232	1230	69.57%	1373	1381	748	54.48%

Note: "Expansions" are calculated by the ratio of number of newcomers in the test period to the number of organizations in the validation period.

Discussion

There is little literature discussing the assessment of the various methods for identifying technology emergence. The present study attempts to assess Technology Emergence Indicators (TEIs) deriving from term/topic prevalence patterns in research literature by answering these questions: How well do the results reflect the concepts and attributes of technical emergence? Are the emerging topics prominent? Is the emergence research growing? Does the emerging topics research community expand?

To address these assessments, we analyze three datasets – NLP, NEDD, and DSSCs. We select these to provide diversity in research fields, dataset size, and developmental stage. Carley et al. (2017)found that persistence and predictive utility of emergent topics could be influenced by both field (technical domain) and dataset size (scale). We believe that the present results offer reasonable generalizability based on the dataset diversities. However, we caution that results speak to technical domain foci, not, for instance, search sets focused on a scholarly discipline or a disease or a non-technical domain (e.g., commercialization opportunities or WOS publications of a given university).

This research revisits the attributes of technology emergence, providing an updated methodology of measuring TEIs proposed by Carley et al. (2018) and Porter et al. (2018). We separately consider ways to capture the emergence attributes of growth, persistence, and community. In so doing, we alter the computations in several ways.

For the growth criterion, we change TEI (emergence scoring) calculations to associate the growth factor with the growth of the field (e.g., NLP, NEDD, DSSCs). No matter if the technical domain being analyzed is an emerging field with rapid growth or a field with a relatively stable number of publications per year, the new method employs the same threshold in calculating TEIs. This offers considerable advantage in computational simplicity and generality of results.

We replace an author-level community criterion with an organization-level community criterion. This is conceptually appealing and far simpler to compute because disambiguation is much easier and development of thesauri to consolidate organization name variations is quite feasible.

A major concern with the prior TEI (emergence scoring) process was that the ETs, as presented, included very general technical domain terms -- e.g., variations on the domain name (like Dye-Sensitized) and search terms. The new scope criterion reduces the inclusion of such general terms, making it much easier to present interesting emerging specializations within the technical domain under study. Table 5 compares these ET sets. This process was enabled by introducing two-tier Inverse Document Filter (IDF) filtering: (1) within-dataset IDF to screen out those very frequent terms in the domain; and (2) cross-dataset IDF to help screen out terms common in scientific publication generally. We recognize some likelihood of loss from (2) in that certain common scientific terms (e.g., "clinical") that show accelerating appearance in the technical domain could contain provide useful intelligence. But on balance, incorporation of the IDF screening produces more compelling emergence indicators.[4]

Through a series of sensitivity analyses summarized herein, we investigated the effects of such methodological changes on the TEIs. Changes showed as generally moderate, providing evidence that the general TEI formulation is robust to adjustments of the novelty, community, and persistence factor thresholds and to growth component calculations. For example, most terms identified as ETs in the previous method are included in the new methodology too. We also are reassured in that the new methodology evidences changes that seem beneficial – i.e., moving the indicators in appropriate directions.

Community measures are also of importance in understanding emergence phenomena. Compared to non-emerging research topics, emerging topics attract more researchers and organizations in the next (future) period (of two or three years). Of particular interest, ETs show as more attractive to new researchers entering the technical domains under study. Further study is warranted – for instance, we note the richest draw of newcomers in DSSCs (Table 8), and that is an especially cohesive research community [e.g., an imposing volume of highly cited (internally to the field) publications].

Predictive utility of the TEIs remains generally similar to that using the previous methodological formulation (i.e., indicating robustness), and quite impressive. The last column of Table 2 shows that, on

[4] As of January, 2020, we have adjusted the specialized scope procedure. We abandon the cross-dataset comparison as problematic in requiring a representative population or random sample outside the test dataset. We tighten the within-dataset filter to use IDF >= 1.

average, Emerging Terms are more prominent in publications in the following time period than in the most recent years of the validation period on which data they were calculated. In other words, based on these results, we could predict that ETs will continue to be actively researched in the future few years. Thus, a researcher considering promising topics to pursue might well want to consider these.

We also note limitations re: prediction – these indications are "on average," and future ET activity is tempered by overall activity of the technical domain (e.g., in Table 7, last column, the relatively slow-growth NLP field shows non-growth for ETs). Also, we feel that basing TEIs on multiple criteria, including novelty, persistence, and community is sound, but we recognize that were one to just consider growth rates of term usage, trend extrapolation would likely correspond highly in terms of spotting accelerating research topics. However, the TEI methodology importantly helps set apart the emerging terms and topics from huge, messy sets of all, or even rapid growth per se, topical terms. *It thus distinguishes emerging topics from the overall domain research activity.* That enables examination of these as promising foci for one's research and also enables policy analyses focusing on emerging topics.

Conclusions, Future Research, and Applications

Conclusions

This research contributes to the literature on technology emergence by experimenting with a scoring routine to discern emerging terms/topics in research literatures, and evaluating their properties.

The basic idea is to investigate whether the identified emerging terms/topics evidence emergence. Based on this idea, our attempt starts from three attributes of technology emergence: persistence, growth, and community; so, we perform a 3-dimentional analysis for evaluating "TEIs" (i.e., the emerging terms/topics). This research builds from conceptualizations of emergence to explore emerging indicators and their attributes.

Second, this research offers evidence that this TEI formulation performs effectively and robustly by revisiting the attributes of technology emergence and updating the methodology of measuring TEIs proposed by Carley et al. (2018) and Porter et al. (2018). We conduct a series of sensitivity analyses to investigate the effects of such methodological changes on the TEIs. The results show that TEIs reflect well on the concepts and attributes of technical emergence, and TEIs are proper indicators for identifying emergence, based on attributes of research publication data. The sensitivity analyses also provide points of reference to analysts who might want to further adjust the emergence scoring parameters.

Third, we provide evidence to support our claim that technological emergence identified by our TEI method is not an "emerged" technology, but an emerging technology. There are lots of concerns about whether technology emergence indicators, as identified by various methods, indicate past prominence, or are prolog to future R&D emphases. As discussed in our Literature Review, many identification methods take rapid growth as the signal of technology emergence, which leads to the question: are they already emerged? We believe "emerging" is a flow that happens in a time period rather than at a point in time. The result of our evaluation shows that the publications and research communities addressing the identified emergence ideas keep growing in the following time period. This observation lends confidence to our TEIs as offering predictive value to anticipate terms/topics likely to be actively pursued in the next 2-3 years.

We also need to keep in mind that technological emergence is a complex socio-technical phenomenon. Beyond research publication or patent data and their analyses, other aspects matter. In terms of prospects of future research attention, one would want to consider funding priorities, for one prominent example. As one treats translation from basic research to development to commercial (or medical or military) application, a host of factors matter – e.g., costs, collaboration and competition, and intellectual property regimes. And as mentioned in the introductory sections, one ought to consider potential benefits and costs of innovations based on the R&D in question. One would well want to consider consultative processes focusing on Emerging Topics, as per Constructive Technology Assessment (noted earlier).

Future Research

We see a number of directions for future research on TEIs. Near at hand, this paper focuses on Emerging Terms (ETs). Further enhancements of the process (Fig. 1) are on our agenda. For one, we are working to streamline an acronym identifier to remedy the issues caused by mixing of acronyms with their full terms.

We are working on consolidating ETs into Emergent Topics, at various degrees of clustering. We aspire to understand the behavior of those topics as well. An arm of this research briefly mentioned is to take the ETs as a first tier and go on to study a second tier – the "players" who evidence most activity in publishing (or patenting) on the ETs. There is room to analyze individual researchers or inventors, their organizations, and countries in terms of patterns of engagement with tech emergence (Porter et al. 2018). Our 3-dimensional evaluation approach could well help tune TEIs at the player level.

Our focus here has been concentrated on the TEI set being advanced by the Georgia Tech Program in Science, Technology & Innovation Policy (STIP) and Search Technology, along with collaborators from Beijing Institute of Technology and the Manchester Institute of Innovation Research (MIoIR). We want to engage other approaches as well. For one, we orchestrated a contest to advance alternative methods to measure tech emergence (http://www.vpinstitute.org/academic-portal/tech-emergence-contest/) that widens treatment of R&D data, as presented at the 2019 Global Tech Mining Conference (http://www.vpinstitute.org/global-techmining-conference/2018-global-techmining-conference/).

Further, we recognize other approaches. The TEI set under scrutiny here is "term based"; others focus on "record based" processes – i.e., to spotlight cutting edge papers. Via our approach, we can also discern emergent papers by tallying the richness of their ET content. And as mentioned, weighing expert opinion on interestingly Emerging Topics counterbalances the empirical approaches of focus here. Some combination of empirical and expert opinion components seems attractive. With our TEIs, modest steps in this direction include engaging domain experts in formulating search queries (i.e., scoping), reviewing intermediate results, and interpreting TEIs in terms of what is interesting, surprising, and/or promising.

Broadening beyond a singular technical domain, we want to explore how well the TEI methods perform on data from channels beyond research. We want to learn to what extent application of these methods to distinguish emerging terms/topics in patent abstract record compilations yield similar results? Further steps would be to analyze search sets from commercial databases like ABI/INFORM, popular media compilations like LexisNexis, or web compilations based on Google to detect emerging topics. Combining study on one topic drawing from multiple data resources holds promise – e.g., combine R&D publication plus patent data, along with agency funding data and commercial activity information. Does the tech emergence we purport to detect translate beyond research activity?

A challenging extension would be to analyze one or several issue areas. That is, rather than searching for activity on a single tech domain (e.g., DSSCs), search for emerging energy technologies. Or to go further possibly, and tackle particular global challenges (e.g., global warming) to distinguish emerging topics in R&D or other information resource sets.

Applications

To sum up, this empirical study strengthens the foundations of a set of Technological Emergence Indicators (TEIs). It demonstrates a degree of generalizability for these indicators by performing and comparing computations for three technological domains. A series of sensitivity analyses demonstrate robustness of the TEIs to changes in the filters applied to emergence criteria of novelty, community, and persistence.

The research supports application of the TEIs at two tiers – the basic level that discerns emerging terms (ETs) and topics (ETopics), and the "player" level that tabulates which players (individuals, organizations, countries) most actively engage those cutting edge topics. Basic level analyses of ETs can inform researchers' identification of prime opportunities. For instance, a grad student hunting for a dissertation topic within a target domain (e.g., NEDD) might pursue opportunities at her/his university with faculty actively researching particular ETs. Pointing out ETopics offers payoffs for R&D program managers in choosing which sub-domains to support.

At the player level, delineation of "who" is actively researching the emerging terms/topics in a domain provides a complementary indicator to extant indicators, such as overall domain publication intensity and citations garnered. TEIs could augment various science indicators. For instance, in other TEI analyses we have tracked Iran's ascending research publication on nanotechnology, and then generated ETopic player analyses to discern which organizations in Iran are particularly active on, say, NEDD "hot topics" (Wang et al., 2019).

TEIs can contribute to various analyses of emerging technologies. They can help develop technology roadmaps showing the evolution of topical emphases within a domain, like NEDD (c.f., Huang et al., 2015). Moreover, as supported by the present study, TEIs offer predictive utility – they show real value in discerning topics within a domain of interest that are more likely to be actively researched over the next two or three years. Such predictive power has immense potential to serve needs of researchers, developers, and funders who want to forecast directions that a particular technology is likely to follow.

Acknowledgements

This work was undertaken with support from the US National Science Foundation (Award #1759960 – "Indicators of Technological Emergence") to Search Technology, Inc. and Georgia Tech. The findings and observations contained in this work are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We wish to express our appreciation to Jan Youtie for sharing pearls of wisdom with us and to Seokbeom Kwon who provided the sensitivity tests of TEI.

References

Arora, S. K., Porter, A. L., Youtie, J., & Shapira, P. (2013). Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs. *Scientometrics, 95*(1), 351-370.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chavez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics, 75*(3), 495-518.

Breitzman, A., & Thomas, P. (2015). The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy, 44*(1), 195-205.

Burmaoglu, S., Porter, A. L., & Souminen, A. What is technology emergence? A micro level definition for improving tech mining practice. In *Portland International Conference on Management of Engineering and Technology (PICMET), Honolulu, 2018*

Burmaoglu, S., Sartenaer, O., & Porter, A. (2019a). Conceptual definition of technology emergence: A long journey from philosophy of science to science policy. *Technology in society*, 101-126, doi:https://doi.org/10.1016/j.techsoc.2019.04.002.

Burmaoglu, S., Sartenaer, O., Porter, A., & Li, M. (2019b). Analysing the theoretical roots of technology emergence: an evolutionary perspective. *Scientometrics, 119*(1), 97-118.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2017). A measure of staying power: Is the persistence of emergent concepts more significantly influenced by technical domain or scale? *Scientometrics, 111*(3), 2077-2087.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics, 115*(1), 35-49.

Carley, S. F., Porter, A. L., & Youtie, J. L. (2019). A Multi-match Approach to the Author Uncertainty Problem. *Journal of Data and Information Science, 4*(2), 1-18.

Cheng, A. C., Chen, C. J., & Chen, C. Y. (2008). A fuzzy multiple criteria comparison of technology forecasting methods for predicting the new materials development. *Technological Forecasting and Social Change, 75*(1), 131-141, doi:https://doi.org/10.1016/j.techfore.2006.08.002.

Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H. A., Pistorius, C., et al. (2001). On the future of technological forecasting. *Technological Forecasting and Social Change, 67*(1), 1-17.

Corning, P. A. (2002). The re-emergence of “emergence”: A venerable concept in search of a theory. *Complexity, 7*(6), 18-30.

Corrocher, N., Malerba, F., & Montobbio, F. (2003). *The emergence of new technologies in the ICT field: main actors, geographical distribution and knowledge sources.* Department of Economics, University of Insubria.

Cozzens, S., Gatchair, S., Kang, J., Kim, K.S., Lee, H. J., Ordóñez, G., et al. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management, 22*(3), 361-376.

Crutchfield, J. P. (2013). Is anything ever new? Considering emergence. In M. A. Bedau & P. Humphreys (Eds.), *Emergence: Contemporary readings in philosophy and science*. MIT Press Scholarship Online: The MIT Press.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8), 981-1012.

Esmaelian, M., Tavana, M., Di Caprio, D., & Ansari, R. (2017). A multiple correspondence analysis model for evaluating technology foresight methods. *Technological Forecasting and Social Change, 125*, 188-205, doi:https://doi.org/10.1016/j.techfore.2017.07.022.

Foster, J., & Metcalfe, J. S. (2012). Economic emergence: An evolutionary economic perspective. *Journal of Economic Behavior & Organization, 82*(2-3), 420-432.

Goldspink, C., & Kay, R. (2010). Emergence in organizations: The reflexive turn. *Emergence: Complexity and Organization, 12*(3), 47-63.

Guo, Y., Xu, C., Huang, L., & Porter, A. (2012). Empirically informing a technology delivery system model for an emerging technology: illustrated for dye-sensitized solar cells. *R&D Management, 42*(2), 133-149.

Guston, D. H., & Sarewitz, D. (2002). Real-time technology assessment. *Technology in society, 24*(1-2), 93-109.

Halaweh, M. (2013). Emerging technology: What is it. *Journal of technology management & innovation, 8*(3), 108-115.

Harper, D. A., & Endres, A. M. (2012). The anatomy of emergence, with a focus upon capital formation. *Journal of Economic Behavior & Organization*, 82(2-3), 352-367. 166.

Huang, Y., Ma, J., Porter, A.L., Kwon, S., and Zhu, D. (2015). Analyzing collaboration networks and developmental patterns of nano-enhanced drug delivery (NEDD) for brain cancer, *Beilstein Journal of Nanotechnology 6* (Special issue on Nanoinformatics), 1666-1676. http://www.beilstein-journals.org/bjnano/content/6/1/169.

Jun, S., & Lee, S. J. (2012). Emerging technology forecasting using new patent information analysis. *International Journal of Software Engineering and Its Applications, 6*(3), 107-116.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change, 75*(6), 771-782.

Kim, D. H., Lee, H., & Kwak, J. (2017). Standards as a driving force that influences emerging technological trajectories in the converging world of the Internet and things: An investigation of the M2M/IoT patent network. *Research Policy*, 46(7), 1234-1254, doi:https://doi.org/10.1016/j.respol.2017.05.008.

Li, M., Porter, A. L., & Suominen, A. (2018). Insights into relationships between disruptive technology/innovation and emerging technology: A bibliometric perspective. *Technological Forecasting and Social Change*, 129, 285-296.

Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management, 7*(2), 139-168.

Martin, R., & Sunley, P. (2012). Forms of emergence and the evolution of economic landscapes. *Journal of Economic Behavior & Organization, 82*(2-3), 338-351. doi:https://doi.org/10.1016/j.jebo.2011.08.005

Metcalfe, B. (1995). Metcalfe's law: A network becomes more valuable as it reaches more users. *Infoworld, 17*(40), 53-53.

Porter, A. L. (2010). Technology foresight: types and methods. *International Journal of Foresight and Innovation Policy, 6*(1-3), 36-45.

Porter, A.L., & Cunningham, S.W. (2005). *Tech Mining: Exploiting New Technologies for Competitive Advantage*, Wiley, New York [Chinese edition, Tsinghua University Press, 2012].

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2018). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change*. 146, 628-643 doi: https://doi.org/10.1016/j.techfore.2018.04.016

Porter, A. L., Roessner, J. D., Jin, X.Y., & Newman, N. C. (2002). Measuring national 'emerging technology'capabilities. *Science and Public Policy, 29*(3), 189-200.

Porter, A. L., Rossini, F.A., Carpenter, S.R. and Roper, A.T. (1980). *A Guidebook for Technology Assessment and Impact Analysis*. New York: North Holland.

Rader, M., & Porter, A. (2008). Fitting future-oriented technology analysis methods to study types. In *Future-Oriented Technology Analysis* (pp. 25-40): Springer.

Roper, A.T., Cunningham, S.W., Porter, A.L., Mason, T.W., Rossini, F.A., & Banks, J. (2011), *Forecasting and Management of Technology*, 2d edition, New York: John Wiley.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827-1843.

Saritas, O., & Burmaoglu, S. (2015). The evolution of the use of Foresight methods: a scientometric analysis of global FTA research output. *Scientometrics, 105*(1), 497-508, doi:10.1007/s11192-015-1671-x.

Sawyer, R. K. (2001). Emergence in sociology: Contemporary philosophy of mind and some implications for sociological theory. *American journal of sociology, 107*(3), 551-585.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American society for Information Science, 50*(9), 799-813.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.

Sohn, S. Y., & Ahn, B. J. (2003). Multigeneration diffusion model for economic assessment of new technology. *Technological Forecasting and Social Change, 70*(3), 251-264, doi:https://doi.org/10.1016/S0040-1625(02)00200-7.

Srinivasan, R. (2008). Sources, characteristics and effects of emerging technologies: Research opportunities in innovation. *Industrial Marketing Management, 37*(6), 633-640.

Van Merkerk, R. O., & Smits, R. E. (2008). Tailoring CTA for emerging technologies. *Technological Forecasting and Social Change, 75*(3), 312-333.

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology, 69*(2), 290-304.

Wang, Z., Porter, A.L., Kwon, S., Youtie, J., Shapira, P., Carley, S.F., and Liu, X. (2019). Updating a search strategy to track emerging nanotechnologies, *Journal of Nanoparticle Research,* 21: article #199; doi.org/10.1007/s11051-019-4627-x.

Zhou, X., Porter, A. L., Robinson, D. K., Shim, M. S., & Guo, Y. (2014). Nano-enabled drug delivery: A research profile. *Nanomedicine: Nanotechnology, Biology and Medicine, 10*(5), e889-e896.

Appendix

Detailed steps for the **term cleaning process** are listed below.

1. Use *VantagePoint's* natural language processing to extract noun phrases of abstracts and titles. [We have experimented with use of WOS Keywords-Plus and Keywords-Authors. We don't use those here so as to make our evaluation processes more generalizable to other data resources lacking such fields.]
2. Merge the two fields (abstract noun phrases and title noun phrases) together and remove the terms with instances fewer than 2.
3. Apply Cluster Suite (a *VantagePoint* script available at www.VPInstitute.org, under Resources) to
    a. eliminate single characters,
    b. remove keywords beginning with non-alpha numeric characters,
    c. remove XML tags,
    d. consolidate chemical compounds and their abbreviations,
    e. consolidate multiple keywords common to scientific and academic publications into one header.
4. Run a general fuzzy routine in *VantagePoint* to consolidate name variations in the list.
5. Divide the terms into unigrams and multigrams:
    a. For unigrams, run a WOS stopwords list to remove common academic words.
    b. For multigrams, use *VantagePoint's* Folding NLP macro to consolidate the phrases.

6. Combine the unigrams and multigrams, and finally get the consolidated terms field as input to identify emerging terms and calculate their emergence scores.

Here is our process to generate annual, **random sample datasets** for WOS:

1. Split the total records into several search queries with under 100,000 records each. In recent years, there have been more than 2 million records in the WOS database for each year. Since the WOS could show no more than 100,000 records in one search query, we needed to split the records into several search queries. We first split the data by the capital letters of organizations. If the number of records for one capital letter was also more than 100,000, then we would separate the search by the Web of Science Categories (WCs). This way, we could split the total records for each year into 39 search queries, which constitute the whole record set.
2. Calculate the number of records we should download for each query. The data were split into 39 search queries, and the number of selections for each query was calculated according to the share of the number of records for each query out of the total number of records. For example, there are 90,079 records that are assigned to organizations beginning with the letter "k", making 3.52% of the total 2,559,592 records in 2014. Thus, we would download 3.52%*5000=176 records for this query.
3. Generate a specific amount of random numbers and download the corresponding records. For example, for the search query "OG=K* AND PY=2014", we generate 176 random numbers between 1 and 90079, and download the relative records.

Supplemental Materials

Extensive supplemental materials are available at: http://..... These materials provide the sensitivity test of TEIs in NLP dataset and DSSCs dataset, and the full lists of emerging terms identified by both the new method and the previous method for all datasets.