Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

A topic models based framework for detecting and forecasting emerging technologies

Shuo Xu[a], Liyuan Hao[a], Guancan Yang[b], Kun Lu[c], Xin An[*,d]

[a] Research Base of Beijing Modern Manufacturing Development, College of Economics and Management, Beijing University of Technology No. 100 PingLeYuan, Chaoyang District, Beijing 100124, P.R. China
[b] School of Information Resource Management, Renmin University of China No. 59 Zhongguancun Street, Haidian District, Beijing 100872, P.R. China
[c] School of Library and Information Studies, The University of Oklahoma 401 W. Brooks St., Norman, Oklahoma 73072, USA
[d] School of Economics & Management, Beijing Forestry University No. 35 Qinghua East Rd., Haidian District, Beijing 100083, P.R. China

ARTICLE INFO

Keywords:
Emerging Technology
Topic Model
Topical N-Grams Model
Name Disambiguation
Credit Allocation

ABSTRACT

The identification of emerging technologies can bring valuable intelligence to enterprises and countries determining research and development (R&D) priorities. Emerging technologies are closely related to emerging topics in terms of several well-documented attributes: relatively fast growth, radical novelty and prominent impact. Our previous work on detecting and forecasting emerging topics is adapted to measure technology emergence, but the dynamic influence model (DIM) is replaced by the topical n-grams (TNG) model in this framework to nominate several emerging technologies in technical terms and to exploit the potential of topic models. Hence, *technologies* are viewed as *term-based themes* in this study. Three indicators are designed to reflect the above attributes: the fast growth indicator, the radical novelty indicator and the prominent impact indicator. The relatively fast growth indicator is calculated from the results of the TNG model and the radical novelty indicator comes from the citation influence model (CIM). As for the prominent impact indicator, the involving authors are used after name disambiguation and credit allocation. The following fields are utilized to develop the models: title, abstract, keywords-author, publication year, byline information, and cited references. We participated in the 2018–2019 Measuring Tech Emergence Contest with the proposed method, and 8 out of 10 submitted ones met the contest organizer's criteria of technology emergence. Criteria included the percentage of high growth terms out of total terms provided, the degree of growth of the terms, and the frequency of those high growth terms across the dataset. Then, a qualitative assessment of overall methodology was conducted by three judges. In the end, we won Second Prize in the contest.

1. Introduction

The emergence and development of emerging technologies provide an opportunity for countries and enterprises to reverse the dilemma of innovation (Kong et al., 2017), achieve cutting-edge technological breakthroughs (Joung and Kim, 2017), and nurture emerging markets (Day et al., 2000). The identification of emerging technologies can bring potential value to enterprises and countries in terms of technology intelligence. There is considerable effort spent towards understanding the emergence of technologies (Burmaoglu et al., 2019; Carley et al., 2018; Joung and Kim, 2017; Lee et al., 2018; Porter et al., 2019; Wang, 2018; Xu et al., 2019a) from science and technology (S&T) information resources, including scientific articles and patents. Such studies can help research foundations and policy-makers focus on research policy and the management of innovation, and individual research agendas.

The last five decades have witnessed significant progress in the domain of emerging technologies detection and forecasting starting with the work of de Solla Price (1965). Several major technologies have been developed, such as citation based approaches (Boyack et al., 2014; Small et al., 2014; Takeda and Kajikawa, 2009), lexical based approaches (Arora et al., 2013; Guo et al., 2011; Weismayer and Pezenka, 2017), and machine learning based approaches (Kyebambe et al., 2017; Lee et al., 2018; Xu et al., 2019a). However, there are no repetitive phenomena or events which can be used as a reference point for improving the accuracy of detection and forecasting (Bolger and Wright, 2017). Therefore, it is basically impossible to directly compare detected and forecasted emerging technologies with realized outcomes.

[*] Corresponding author.
E-mail addresses: xushuo@bjut.edu.cn (S. Xu), haoliyuan@emails.bjut.edu.cn (L. Hao), yanggc@ruc.edu.cn (G. Yang), kunlv@ou.edu (K. Lu), anxin@bjfu.edu.cn (X. An).

https://doi.org/10.1016/j.techfore.2020.120366
Received 23 January 2020; Received in revised form 24 September 2020; Accepted 25 September 2020
Available online 12 October 2020

Though Apreda et al. (2019) re-examined the findings of a technology foresight exercise on the *medical device* industry with realized technological performance, five years later, it is still very difficult to tackle this problem. The 2018–2019 Contest of Measuring Tech Emergence[1] pioneered by Professor Alan L. Porter as a first step in this direction tries to provide a benchmark dataset public available with known emerging technologies.

This contest challenges one to devise a repeatable procedure to identify emerging technologies within a defined S&T domain. Technologies can be *terms*, or *term-based theme*s, but they must appear in the Web of Science (WoS) abstract records. The data resource to be mined is an R&D publication dataset that is provided by the contest organizer, on a designed science or technology domain, drawn from the WoS. The search strategy for the data resource is not open to contest participants. The following is a key criterion: who best predicts technologies that are notably active in the following two years of research?

We participated in this contest with a topic models based framework, and won Second Prize. Here, the methodology and results of our participation are described in more details. Before this, in order to make it easier to understand our framework, a full picture of the indicators is given in the first place. In fact, this work is inspired by the definitions of an emerging technology (Rotolo et al., 2015) and an emerging research topic (Wang, 2018). The former identifies five attributes that feature in the emergence of novel technologies: *radical novelty, relatively fast growth, coherence, prominent impact*, and *uncertainty and ambiguity*, and four attributes are attached to an emerging research topic: *radical novelty, relatively fast growth, coherence*, and *scientific impact*. By comparing carefully these two definitions, two-fold differences can be observed: (1) The socio-economic impact is emphasized in Rotolo et al. (2015), but scientific influence is stressed in Wang (2018); and (2) Wang (2018) considered *uncertainty and ambiguity* to be irrelevant, since scientific influence is more likely to become evident in a short time period after the emergence of a research topic.

It has been shown that scientific publications and patent documents can usually serve as respective proxies of scientific research and technical development (Xu et al., 2019c). In this way, patents can be utilized to measure the socio-economic impact of an interested emerging technology to some extent (Ke, 2020; Veugelers and Wang, 2019). However, only scholarly articles are provided by the contest organizer, search strategy is not open and the contest focuses only on the technologies that are notably active in the subsequent two years of research. In the interim, our previous work on detecting emerging topics (Xu et al., 2019a) indicates that thematic structures from a topic model based framework are sufficiently coherent per se. Therefore, our framework only considers three attributes: relatively fast growth, radical novelty and prominent impact. The following summarizes main contributions of this work:

- A topic models based framework is proposed to detect and forecast emerging technologies by operationalizing the characteristics of relatively fast growth, radical novelty and prominent impact. This differs from our previous work in terms of topic extraction and prominent impact operationalization.
- The Topical N-Grams (TNG) model is used to extract term-based themes. To identify a proper number of themes, the perplexity for the TNG model can be calculated effectively with an iterative formula.
- The involving authors are used to calculate the prominent impact indicator after disambiguating authors' names with a rule-based scoring and clustering method and allocating authorship credit with the sequence-determines-credit schema.

The organization of the rest of this paper is as follows. After related work is briefly reviewed in Section 2, a topic models based framework is put forward in Section 3, and topic extraction and indicator calculation are also described in more details in this section. Section 4 shows experimental results and discussions on a publication dataset provided by 2018–2019 Measuring Tech Emergence Contest organizer, and Section 5 concludes this work.

2. Literature review

Before delving into more specifies, discussion of the literature pertinent to detecting and forecasting emerging technologies is in order. Note that many closely related concepts exist in the literature, such as *emerging research topics, emerging topics, research fronts, emerging trends, emerging research fields*, and *transformative technologies*. This article does not distinguish their connotations and extensions. For more elaborate and detailed surveys we refer the readers to Rotolo et al. (2015), Burmaoglu et al. (2019), Xu et al. (2020), and Lu et al. (2020).

2.1. Methods of detecting and forecasting emerging technologies

The development of emerging technologies detection and forecasting fields can be broadly divided into three different stages (Xu et al., 2020): the emergence stage, exploration stage and development stage. The ground-breaking work of de Solla Price (1965) first defined the concept of research front, which is a sort of growing tip or epidermal layer and emphasized its *novelty* attribute. Several years later, another important work (Small, 1973) laid the methodological foundation, which proposed the co-citation analysis method to identify the emerging technologies for future research. Then, Small and Griffith (1974) merged these two studies and the related research moves to the next stage, namely, the exploration stage.

The exploration stage mainly deals with the research of emerging technologies from the perspective of the citation network analysis, such as co-citation (González-Alcaide et al., 2016), bibliographic coupling (Glänzel and Thijs, 2012; Huang and Chang, 2014), and direct citation (Shibata et al., 2008; Waltman and van Eck, 2012). In more details, this stage has three research streams: (a) the performance comparison among three citation network methods (Boyack and Klavans, 2010; Jarneving, 2007; Shibata et al., 2008); (b) the improved citation-based methodologies, such as citation network based on authors (Ma, 2012; Zhao and Strotmann, 2008) and the hybrid citation-link network method (Small et al., 2014); and (c) the citation-based and lexical-based methodological application (Chen, 2006; Li, 2017; Takeda and Kajikawa, 2009).

A step forward in 2015 triggered a new development in the detection and forecasting of emerging technologies. Rotolo et al. (2015) gave a clear and comprehensive concept of the emerging technologies on the basis of previous works. From 2015 till now, this stage could be classified into two branches. One branch (Burmaoglu et al., 2019; Joung and Kim, 2017; Lee et al., 2018) explored novel definitions and new detection and forecasting approaches. The other branch (Carley et al., 2018; Porter et al., 2019; Xu et al., 2019a) devoted to developing a series of technology emergence indicators.

While approaching *the emergence of technology* from the broad perspective of science, technology and innovation (ST&I), understanding the process of the emergence of technology is essential to technology forecasting research at either the micro or macro level (Zhang et al., 2016). Sommarberg and Mäkinen (2019) utilized a unique survey method using the Visual Analogue Scale (VAS) in seminars to track qualitatively disruptive changes in an industry by expert knowledge. In additon, science overlay maps offer an alternative for observing the ongoing transformation of science and technology (Rafols et al., 2010; Suominen and Toivanen, 2016). In recent years, the forecasting innovation pathway (FIP) approach has shown to be capable of capturing the potential development of emerging technologies (Robinson et al., 2019). In view of forecasting technological emergence, quantitative and

[1] https://vpinstitute.org/academic-portal/tech-emergence-contest/

qualitative mixed methods can serve as a solution to further explore emergence and evolution of a technological domain. Zhou et al. (2019) integrated term clumping, Subject-Action-Object (SAO) technique (Chen et al., 2020), and net effect analysis with FIP to identify the main areas of R&D and track their evolution over time. They then used technology roadmapping to visualize the evolution of the main areas.

With the development of large-scale text-processing techniques, machine learning based methods (Kyebambe et al., 2017; Lee et al., 2018; Xu et al., 2019a; Zhang et al., 2019) are gradually utilized to identify emerging technologies due to their potential power. The framework in this study further exploits the potential of topic models and enriches the machine learning based approaches.

2.2. Topic models aware of word order or syntax information

To nominate emerging technologies, it is vital for a topic model to be able to discover term-based themes. A term-based theme is represented by an ordered list of technical terms. A *technical term* is typically a phrase that functions as a constituent in the syntax of a sentence. However, it is well known that the topic is usually represented by an ordered list of single words in the overwhelming majority of topic models, including the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) and Dynamic Influential Model (DIM) (Gerrish and Blei, 2010). All these single words seem particularly general to inform research prioritization.

We argue that the majority of topic models cannot explicitly take into consideration terms mentioned in the text, since they inherit the bag-of-word (BoW) assumption of the standard LDA model. That is to say, the word order and syntax information are discarded directly by these models. A naïve method is to bind the resulting multiple word tokens with term clumping technique as a pre-processing step. However, this naïve method did not outperform single word modeling, and term clumping performed even poorer than Natural Language Processing (NLP) phrases (Yau et al., 2014).

The HMM-LDA model developed by Griffiths et al. (2005) tries to include syntactic structure information on the basis of the Hidden Markov Model (HMM) (Rabiner, 1989) and the standard LDA model. Though this model can effectively distinguish between content and function word tokens, the components corresponding to content and function word tokens still follow the BoW assumption. The bi-gram statistics and latent topic variable are combined in the Bigram Topic Model (BTM) (Wallach, 2006). A new set of random variables for bi-gram status is introduced in the LDA Collocation (LDA-COL) model by Griffiths et al. (2007) to decide whether to generate a bi-gram or a unigram. As the state-of-the-art model in this direction, the TNG model and its variants (An and Xu, 2019; Wang et al., 2018) are armed with the power to decide whether to form an n-gram term for the consecutive word tokens depending on their nearby context. Hence, the TNG model is utilized in this work.

3. Research framework and methodology

For the purpose of measuring technology emergence, as shown in Fig. 1, our research framework calculates three indicators: relatively fast growth, radical novelty and prominent impact. In this study, technologies are viewed as term-based themes for the following main reasons: (a) to exploit the potential of probabilistic topic models (Blei, 2012), and (b) to nominate several emerging technologies in term of technical terms. Throughout the paper, the term *technology* is used interchangeably with the terms *term-based theme* and *topic*. Here, by term-based theme, we mean that the topic is represented by several phrases rather than single words like in the standard LDA model (Blei et al., 2003) and DIM model (Gerrish and Blei, 2010), since it is very hard to interpret directly the results from the latter in a human-readable manner to inform research prioritization.

The Topical N-Grams (TNG) model (Wang et al., 2007), as the state-of-the-art topic model considering word order information, is utilized here to extract term-based themes from scientific publications after detecting sentence boundaries, tokenizing and lemmatizing each detected sentence, identifying entity mentions and abbreviations, and then filtering stopwords. The relatively fast growth and radical novelty indicators are defined on the basis of the corresponding results from the TNG and CIM models. As for the prominent impact indicator, key researchers in the resulting technologies are identified from the byline information after name disambiguation and credit allocation. In the following subsections, topic extraction and the calculation for three indicators are described at length.

3.1. Topic extraction

Table 1 summarizes the symbols used in this study, where $v = 0$ indicates the start or end marker of the resulting paragraphs or sentences, and $k = 0$ indicates the topic corresponding to $v = 0$. The graphical representation of the TNG model is shown in Fig. 2, where the double-circle node represents the observed variable, the single-circle node represents the latent variable, the arrow means the conditional dependence, and the plate indicates that the internal elements need to be repeated the specified number of times in the bottom right corner of the resulting plate.

For many Bayesian models (An et al., 2014; Wang et al., 2018; Xu et al., 2019c), posterior inference cannot be done exactly in this model. A variety of algorithms have been developed to estimate the parameters in the literature, such as mean-field variational inference (Jordan et al., 1999), Markov chain Monte Carlo (MCMC) (Andrieu et al., 2003), and stochastic variational inference (Hoffman et al., 2013). The collapsed Gibbs sampling algorithm, a special case of MCMC, was originally utilized by Wang et al. (2007) to approximate the posterior of the TNG model.

More specifically, in the collapsed Gibbs sampling procedure, one needs to calculate the posterior distribution, namely, conditional

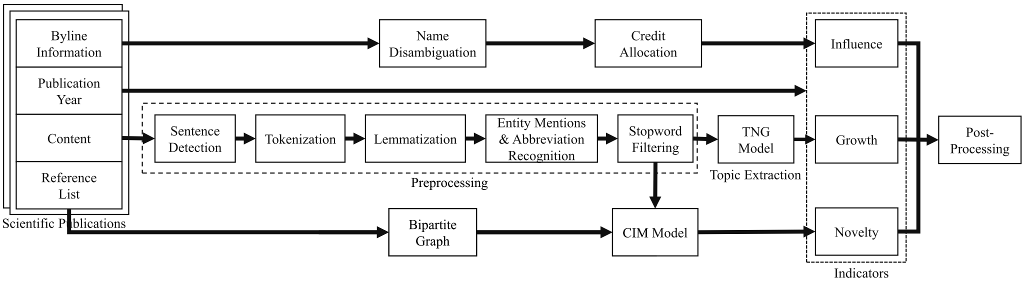


Fig. 1. Research framework for measuring technology emergence.

Table 1
Notations used in the TNG model.

Symbol	Description
K	Number of term-based themes
M	Number of documents
V	Number of unique words
N_m	Number of word tokens in the document m
$\vec{\vartheta}_m$	Multinomial distribution of topics specific to the document m
$\vec{\varphi}_k$	Multinomial distribution of words specific to the topic k
$\vec{\psi}_{k,v}$	Binomial (Bernoulli) distribution of bigram status specific to the topic k and the word v
$\vec{\phi}_{k,v}$	Multinomial distribution of words specific to the topic k and the word v
$z_{m,n}$	Topic associated with the n-th token in the document m
$x_{m,n}$	Bigram status associated with the n-th token in the document m
$w_{m,\ n}$	n-th token in the document m
$\vec{\alpha}, \vec{\beta}, \vec{\gamma}, \vec{\delta}$	Dirichlet/Beta priors (hyperparameter)

distributions of the hidden random variables ($z_{m,n}$ and $x_{m,n}$) given the observations, other hidden variables and hyperparameters, $\Pr(z_{m,n}, x_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}, \vec{\gamma}, \vec{\delta})$, where $\vec{z}_{\neg(m,n)}$ and $\vec{x}_{\neg(m,n)}$ represent respective topic and status assignments for all tokens except $w_{m,\ n}$. After a simple derivation (Wang et al., 2007), the posterior distributions can be formally expressed as follows:

$$\begin{aligned} &\Pr(z_{m,n}, x_{m,n}|\vec{w}, \vec{z}_{\neg(m,n)}, \vec{x}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}, \vec{\gamma}, \vec{\delta}) \\ \propto\ &\left(n_m^{(z_{m,n})} + \alpha_{z_{m,n}} - 1\right) \times \left(n_{z_{m,n-1},w_{m,n-1}}^{(x_{m,n})} + \gamma_{x_{m,n}} - 1\right) \\ &\times \begin{cases} \dfrac{n_{z_{m,n}}^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V} (n_{z_{m,n}}^{(v)} + \beta_v) - 1}, & \text{if } x_{m,n} = 0 \\ \dfrac{n_{z_{m,n},w_{m,n-1}}^{(w_{m,n})} + \delta_{w_{m,n}} - 1}{\sum_{v=1}^{V} (n_{z_{m,n},w_{m,n-1}}^{(v)} + \delta_v) - 1}, & \text{if } x_{m,n} = 1 \end{cases} \end{aligned} \tag{1}$$

where $n_m^{(k)}$ represents the number of words in the document m assigned to the topic k, $n_k^{(v)}$ represents how many times the word v is assigned to the topic k as a unigram, $n_{k,v}^{(v')}$ denotes how many times the word v' is assigned to the topic k as the 2nd term of a bigram given the previous word v, and $n_{k,v}^{(x)}$ denotes how many times the status x (0 or 1) appears given previous word v and previous word's topic k. Using the expectation of Dirichlet/Beta distribution, the model parameters in Table 1 can be readily obtained as follows:

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha_k)} = \frac{n_m^{(k)} + \alpha_k}{N_m + \sum_{k=1}^{K} \alpha_k} \tag{2}$$

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V} (n_k^{(v)} + \beta_v)} \tag{3}$$

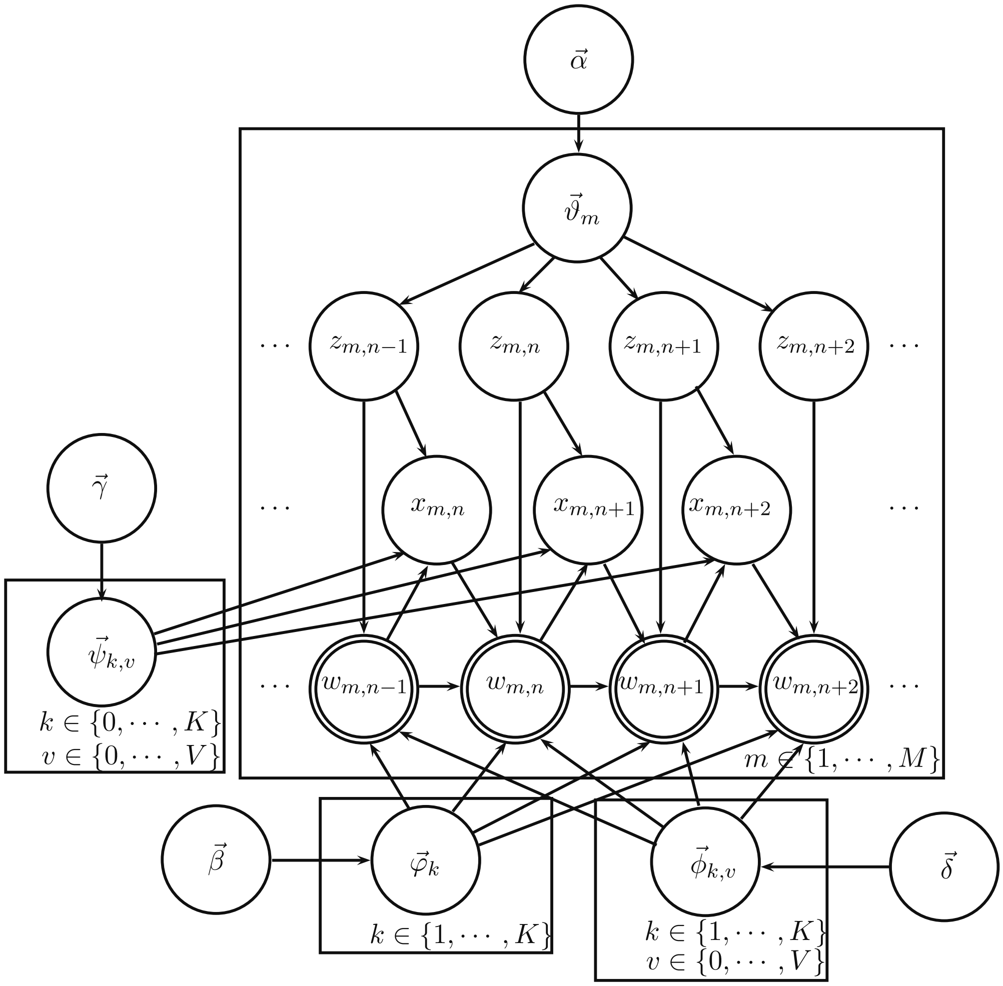


Fig. 2. The graphical model representation of the TNG model.

Table 2
The additional rule set for name disambiguation.

ID	Field	Criterion	Score
1	Email	The email name and main domain name are same.	80
2a	ResearcherID	Same	100
2b	ORCID	Same	100
3	Title & publication venue	Very similar title, but different venues	10

$$\phi_{k,v,v'} = \frac{n_{k,v}^{(v')} + \delta_{v'}}{\sum_{v'=1}^{V} (n_{k,v}^{(v')} + \delta_{v'})} \tag{4}$$

$$\psi_{k,v,x} = \frac{n_{k,v}^{(x)} + \gamma_x}{\sum_{x=0}^{1} (n_{k,v}^{(x)} + \gamma_x)} \tag{5}$$

It is worth mentioning that one can introduce the prior knowledge by setting the bigram status variable $x_{m,n}$. For example, given that the word tokens *synthetic, gene* and *network* can be formed into the term *synthetic gene network*, the bigram status variables $x_{m,n}$ corresponding to the word tokens *gene* and *network* should be set to 1 before posterior inference. That is to say, these bigram status variables are treated as the observed ones. In this work, keywords-author from scientific publications are treated as prior knowledge. The symmetric Dirichlet/Beta priors α, β, γ and δ are set at 0.5, 0.01, 0.1 and 0.01, respectively. The Gibbs sampling is run for 2,000 iterations, including 500 for the burn-in period.

3.2. Indicator calculation

A technology can be considered as emerging if it meets the characteristics of relatively fast growth, coherence, scientific impact, and radical novelty (Rotolo et al., 2015; Wang, 2018). As a matter of fact, similar to Xu et al. (2019a), the term-based themes from the TNG model are also sufficiently coherent per se. Hence, the coherence indicator is not used in this article as a criterion for emerging technologies. The relatively fast growth and radical novelty indicators are operationalized the same as those in Xu et al. (2019a). For completeness, they are re-expressed formally as follows.

Informally, the slope of the popularity reflects the relative growth of the interested topic. Therefore, the relatively fast growth for the topic k at the time t is defined as the slope of the popularity, that is, $\text{Growth}(k, t) = p_{k,t} - p_{k,t-1}$ with the popularity $p_{k,t} = \sum_{m:\text{timestamp}=t} \vartheta_{m,k}$. Additionally, since the CIM model (Dietz et al., 2007; Xu et al., 2019a) can model the topical innovation and topical inheritance via citations between documentations, the radical novelty for the topic k at the time t is formulated as $\text{Novelty}(k, t) = \frac{1}{|\{m : \text{timestamp} = t\}|} \sum_{m:\text{timestamp}=t} [\vartheta_{m,k} \times \lambda_{m,1}]$. Here, $\lambda_{m,1}$ controls the extent from the topic mixture of some cited publication in the CIM model. Thus, $\vartheta_{m,k} \times \lambda_{m,1}$ indicates the innovation of the topic k in the document m. For a topic k to be considered as emerging, the value for Growth(k, t) and Novelty(k, t) in recent years should be higher than their respective average, $\frac{1}{K}\sum_k \text{Growth}(k, t)$ and $\frac{1}{K}\sum_k \text{Novelty}(k, t)$.

The DIM model (Gerrish and Blei, 2010) in Xu et al. (2019a) can directly estimate a meaningful influence measure of each scholarly article, so the scientific impact indicator can be readily stated from the results of this model. Nevertheless, the term-based themes cannot be uncovered by this model, so scientific impact indicator needs to be re-operationalized in this work. Intuitively, an emerging technology will attract more and more research organizations, individuals, and/or countries. This phenomenon is referred to as *research community* in Porter et al. (2019). This study focuses on the key researchers within a particular scientific or technological domain. Typically, a publication is usually written by multiple authors and often covers several topics. Hence, before delving into the calculation of scientific impact indicator from the perspective of the researchers, there are still two problems that need to be resolved: name disambiguation and credit allocation.

3.2.1. Name disambiguation

It is well known that many different researchers share the same name (i.e., the homonym problem), and individual scholars sometimes publish their works under different names (viz., the synonym problem). It is known as the name ambiguity problem. Though many different solutions are put forth in the literature (Caron and van Eck, 2014; Han et al., 2017; Kim, 2018; Torvik and Smalheiser, 2009), the authors in several comprehensive bibliographic databases, such as Scopus and Web of Science (WoS), are not fully unambiguously identified. Moreover, it is not feasible to manually disambiguate so many authors even from a particular scientific or technological field in a limited time. This severely limits the development of bibliometric analysis and tech mining.

To reduce the negative influence of name ambiguity on the scientific impact indicator, a rule-based scoring and clustering method (Caron and van Eck, 2014) is utilized here, but with a different rule set from Caron and van Eck (2014). Apart from the rules in Caron and van Eck (2014), the following three guidelines are used to improve the performance of name disambiguation algorithm in our framework: (a) If two articles share the same email name and main domain name (such as hasty@ucsd.edu and hasty@bioeng.ucsd.edu), it is very likely that these articles were written by the same author; (b) when two publications link to the same ResearcherID or ORCID, the publications should come from the same author; and (c) if the titles of two works are very similar, but their publication venues are different, the resulting authors are likely to be the same person. The last rule comes from our observation on an interesting phenomenon in academia: one usually submits a short and progressive work to a conference due to time limitations, and later submits an extended version to a journal, such as Liu et al. (2009) versus Liu and Wang (2010), and Xu et al. (2017) versus Xu et al. (2018). From the aforementioned rules, one can see that each rule provides a different support degree for disambiguating names. To quantify the support degree, each rule is attached by a different score, as shown in Table 2.

After the rules are formulated, all author names are grouped into blocks (Levin et al., 2012) at first according to the family name and first initial after replacing international letters with their counterparts (such as ä vs. a and č vs. c) and removing all non-alphabetic characters (such as dot and hyphen symbols). For instance, author names ”Csató, Lehel” and ”Bar-Ziv, Roy H.” are assigned to the blocks ”csato_l” and ”barziv_r”, respectively. Then, the pairs of publications are scored according to the curated rules. Finally, the author name blocks are clustered into block size classes 1–6 by means of single-linkage clustering on the basis of the number of publications within a block.

3.2.2. Credit allocation

The increasing collaboration among researchers makes it very difficult to allocate credits to each coauthor when they are ordered in terms of their contribution to a scientific publication. Though many credit allocation schemas have been raised in the literature (Kim and Kim, 2015; Osório, 2018; Xu et al., 2016), there was no consensus about which one is the best until now.

This study prefers the harmonic counting schema (Hagen, 2008), since it has distinct advantages in simultaneously removing both

```
Precondition: $\{\vec{\vartheta}_m\}$ from the TNG model (cf. Table 1)
Precondition: $\{c_{m,i}\}$ from the credit allocation scheme (cf. Subsection 3.2.2)
Precondition: $\tau$ is a threshold preset by user

$\mathrm{score}(k, t, a) \leftarrow 0$ for $\forall k, \forall t, \forall a$   ▹ three-dimensional array, size: $K \times T \times A$; $K, T, A$: number of topics, unique timestamps and authors, respectively
for $m \leftarrow 1$ to $M$ do
    $\vec{a}_m \leftarrow$ author list in the document $m$
    $t_m \leftarrow$ publication year for the document $m$
    for $i \leftarrow 1$ to $A_m$ do                ▹ $A_m$: number of authors in the document $m$
        for $k \leftarrow 1$ to $K$ do
            $\mathrm{score}(k, t_m, a_{m,i}) \leftarrow \mathrm{score}(k, t_m, a_{m,i}) + \vartheta_{m,k} \times c_{m,i}$
        end for
    end for
end for

for $k \leftarrow 1$ to $K$ do
    for $t \leftarrow 1$ to $T$ do
        Sort $\mathrm{score}(k, t, \cdot)$ in ascending order, i.e., $\mathrm{score}(k, t, 1) \geq \mathrm{score}(k, t, 2) \geq \cdots \geq \mathrm{score}(k, t, A)$
        $\mathrm{Influence}(k, t) \leftarrow \arg\min_p \left\{ \frac{\sum_{i=1}^{p} \mathrm{score}(k,t,i)}{\sum_a \mathrm{score}(k,t,a)} \geq \tau \right\}$
    end for
end for

return $\mathrm{Influence}(\cdot, \cdot)$
```

Algorithm 1. Algorithm for calculating scientific impact indicator

inflationary and equalizing bias, and offers a better combination of parsimony and accuracy (Hagen, 2013). For convenience, let A_m denote the number of authors in the document m, and $\vec{a}_m = [a_{m,1}, a_{m,2}, \cdots, a_{m,A_m}]$ be the authors in the document m. The credits award $c_{m,i}$ for the coauthor $a_{m,i}$ can be formally defined as follows (Hagen, 2008):

$$c_{m,i} = \frac{1/i}{\sum_{i=1}^{A_m} 1/i} \tag{6}$$

If a publication has more than one first author and corresponding author, or they do not coincide, these authors are placed in the first positions before applying Eq. (6) and then the average of their scores is assigned to each of them. Thus, the weights of other authors are reduced accordingly.

In addition, as big science becomes even bigger, its scale is increasingly reflected in the byline information of scientific publications. Cronin (2001) coined the term *hyper-authorship* to describe this phenomenon. The extraordinary number of coauthors in these articles results in the failure of many credit assignment schemas when allocating credit awards to individual authors. The sequence-determines-credit (SDC) schema (Tscharntke et al., 2007) suggests that the first author should obtain the full credit, the second author half, the third a third, and so forth, up to the 10th author, after which each remaining author would receive 0.05 credit. Formally, the credit award $c_{m,i}$ for the coauthor $a_{m,i}$ can be delineated as follows:

$$c_{m,i} = \begin{cases} \frac{1}{i}, & i \le 10 \\ 0.05, & \text{otherwise} \end{cases} \tag{7}$$

By comparing Eq. (6) with Eq. (7), it is evident that when $i \le 10$, the harmonic counting schema is the normalized version of the SDC schema. Otherwise, the SDC schema is equivalent to a transformed version of full counting schema (Gauffriau and Larsen, 2005). Hence, the SDC schema is employed to assign the credit award to each coauthor of a scholarly article, but the credit award for each coauthor is normalized.

3.2.3. Scientific impact

Once the author names are disambiguated and credit awards are allocated, scientific impact indicator Influence(k, t) for the topic k at the time t can be operationalized according to Algorithm 1. This algorithm consists of two parts: (a) to compute the contribution of each researcher to the topic k at the time t (Line 1–10); and (b) to determine the scientific impact value by the minimum number of researchers so that their accumulative contributions are greater than or equal to a threshold τ preset by the user (Line 11–15). For a topic k to be considered as emerging, the value for Influence(k, t) in recent years should be higher than its average $\frac{1}{K}\sum_k \text{Influence}(k, t)$.

4. Experimental results and discussion

4.1. Dataset

The fields in the provided dataset include *title, abstract, keywords-author, byline information, publication year, funding acknowledgement, cited references* and many others. The scientific publications span ten years, from 2003 to 2012. The number of scholarly articles is 2,584, and Table 3 reports the number distribution of publications over years. In our framework (cf. Fig. 1), the content information, such as *title* and *abstract*, for each reference is also needed to calculate the radical novelty indicator, but it is unavailable in the provided dataset. Therefore, the cited articles are separately retrieved with the procedure in Xu et al. (2019a), before which the DOIs of cited references are further cleaned with a method in Xu et al. (2019b). In the end, the number of unique references is 71,466.

Table 3
Distribution of the number of publications over year.

Pub. Year	No. of Pub.	Pub. Year	No. of Pub.
2003	132	2008	230
2004	169	2009	286
2005	172	2010	320
2006	191	2011	382
2007	221	2012	481
Σ Pub.	2,584		

The preprocessing steps in this study are very similar to those in Xu et al. (2019a). The sentences in the titles and abstracts are detected with *geniass* (Sætre et al., 2007), and then the splitted sentences are tokenized and lemmatized with *geniatagger* (Tsuruoka et al., 2005). To filter stopwords, an English stopword list from Natural Language Toolkit (NLTK)[2] is customized by expanding several punctuation symbols (such as @, % and so on). All numbers (including integers and floating point numbers) in the citing and cited articles are replaced with a special word *NUMBER*. In order to reduce the interference of unrelated information, copyright information (such as *Published by Elsevier Inc.*) and article status information (such as *received, accepted*, or *first published online* followed by a date) are also removed with human-curated rules based on regular expressions.

In addition, many entities, such as protein, DNA, RNA, cell line, and cell type, are often mentioned in scholarly articles (Chen et al., 2020; Xu et al., 2015). To reduce the size of word vocabulary and improve the performance, these entity mentions are identified with *geniatagger* (Tsuruoka et al., 2005), and then excluded from further analysis. In the meantime, several interesting patterns can be observed for the abbreviations in the text, an example of which is that the first character of each word in the *long form* (i.e., full name) corresponds to one character in the *short form* (i.e., abbreviation), as in *interleukin-12 (IL-12)*. An abbreviation recognition algorithm raised by Schwartz and Hearst (2003) is utilized in this work. In total, 2,348 pairs of the long form and the short form are identified.

After author names are disambiguated, 9,987 unique researchers are obtained, in which about 30% of researchers are attached email, ResearcherId or ORCID information. It is very surprising that a record with UID = ”WOS:000260935200001” in the provided dataset has no authors at all. The distribution of publications with the number of authors illustrated in Fig. 3 make clear that normal multi-authored publications dominates, especially those with 2–7 authors (79.45%). Single-authored and hyper-authored publications (i.e., with more than ten authors) account for 6.08% and 4.49% publications, respectively.

4.2. Number of topics

For the sake of identifying a proper number of topics, the perplexity (Azzonpardi et al., 2003) is calculated for each candidate value from 15 to 50 with a step size 5. As a standard measure for model selection, this measure is defined as the exponential of the negative normalized predictive likelihood under the model $\mathcal{M}$ [cf. Eq. (8)], and a lower value indicates a better modeling performance.

$$\text{Perplexity}(\vec{w}\,|\mathcal{M}) = \exp - \frac{\sum_{m=1}^{M} \log \Pr(\vec{w}_{m,\cdot}|\mathcal{M})}{\sum_{m=1}^{M} N_m} \tag{8}$$

In fact, the calculation of perplexity for the TNG model is more complicated than that of the LDA model due to the conditional dependencies between the latent variables (cf. Fig. 2). Furthermore, Wang et al. (2007) did not stipulate how to effectively calculate the perplexity. Here, the term $\Pr(\vec{w}_{m,\cdot}|\mathcal{M})$ in Eq. (8) can be derived with the

[2] http://www.nltk.org

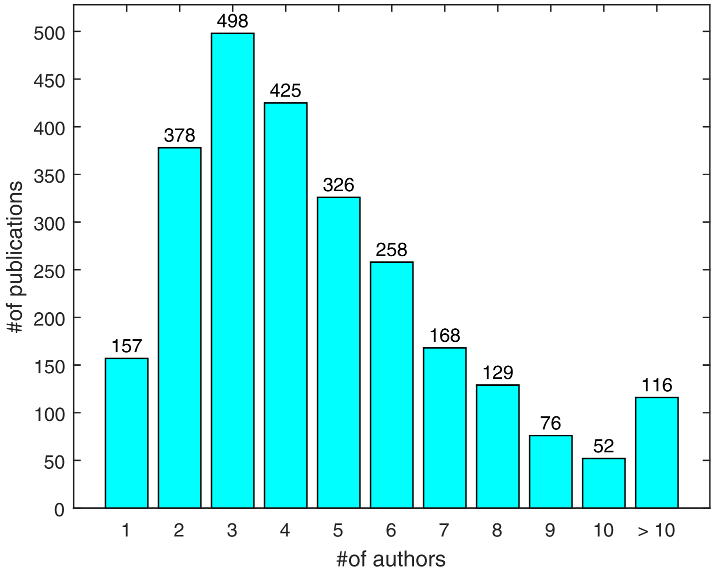


Fig. 3. The distribution of publications with the number of authors.

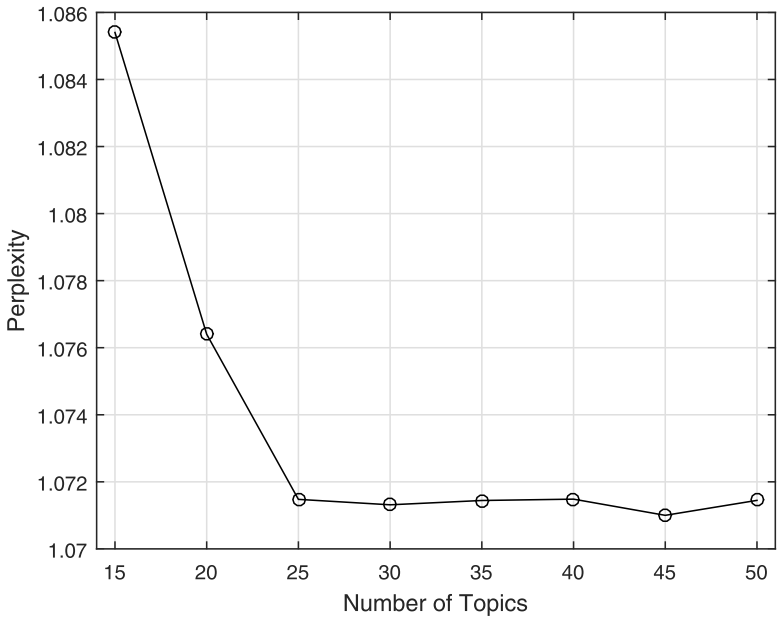


Fig. 4. The perplexity with a different number of topics.

iterative equation ρ(· ; ·) as follows (Please refer to Appendix for more details):

$$\Pr(\vec{\tilde{w}}_{m,\cdot}|\mathcal{M}) = \sum_{k=1}^{K} \rho_k(\vec{\tilde{w}}_{m,\cdot}; N_m)\tilde{\vartheta}_{m,k} \tag{9}$$

with

$$\rho_k(\vec{\tilde{w}}_{m,\cdot}; 1) = \varphi_{k,\tilde{w}_{m,1}}\psi_{0,0,0} \tag{10}$$

$$\begin{aligned}\rho_k(\vec{\tilde{w}}_{m,\cdot}; n) \;=\; & \sum_{k'=1}^{K}\sum_{b=0}^{1}\left[\rho_{k'}(\vec{\tilde{w}}_{m,\cdot}; n-1) \times \tilde{\vartheta}_{m,k'} \times \psi_{k',\tilde{w}_{m,n-1},b}\right. \\ & \left. \times \begin{cases}\varphi_{k,\tilde{w}_{m,n}}, & b = 0 \\ \phi_{k,\tilde{w}_{m,n-1},\tilde{w}_{m,n}}, & b = 1\end{cases}\right]\end{aligned} \tag{11}$$

Fig. 4 depicts the perplexity with a different number of topics. From Fig. 4, one can see that the perplexity of the TNG model converges when the number of topics is 25, so the number of topics K is fixed to 25 in this work. As mentioned in AlSumait et al. (2009), it is very possible for topic models to uncover insignificant themes, or just a collection of irrelevant words. On closer examination of all estimated topics from the TNG model, two topics (id: #13 and #19) are directly discarded in this study. In other words, in the end, 23 term-based themes are further analyzed to check whether the emerging characteristics are met or not.

4.3. Emerging Technologies

Following Section 3.2, one can easily calculate relatively fast growth, radical novelty, and scientific influence (τ = 0.80, 0.85, 0.90) indicators, as shown in Figures 5–9. The black solid line with cross

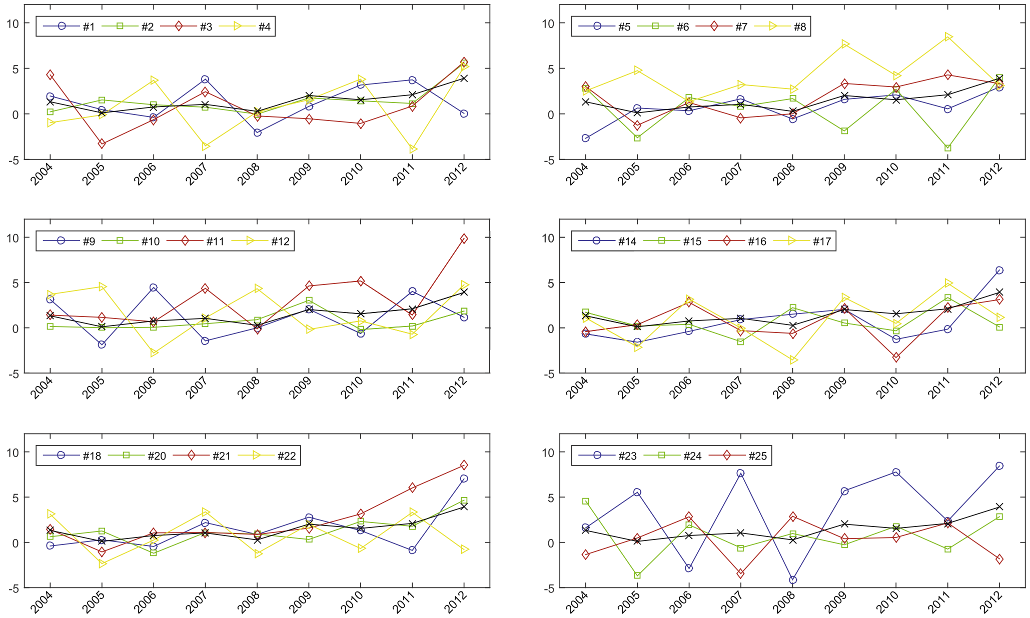


Fig. 5. The trend for the relatively fast growth indicator.

markers in Figures 5–9 corresponds to the average of the resulting indicator. From Fig. 7–9, one can see that the scientific influence indicator is not sensitive to the threshold τ in Algorithm 1. As a matter of convenience, Table 5 summarizes the supports for all topics from each indicator.

To gain some insight of the relationship among the three indicators, the correlation coefficients are calculated in term of Kendall's τ with two-tailed hypothesis (Press et al., 1992), as reported in Table 4. From Table 4, one can see that the prominent influence indicator has a higher positive correlation with the radical novelty indicator than with other pairwise indicators. This point is verified by Table 5 again, since these two indicators show more similar patterns. This is not consistent with the theoretical analysis by Rotolo et al. (2015), which indicates that the emergence of technologies still needs to be further investigated so as to understand the nature of emergence.

From Table 5, one can see that topics #11 and #12 meet all emerging criteria defined in Section 3.2, and topics #2, #4, #5, #7, #8, #9, #14, #21, and #23 conform to two criteria. Nonetheless, the trend for topic #4 changes much too sharply in term of the relatively fast growth indicator. The main reason may be that the TNG model cannot take the dynamics of topics (such as the birth, death, merging and branching of topics) into consideration. For this reason, topic #4 is not

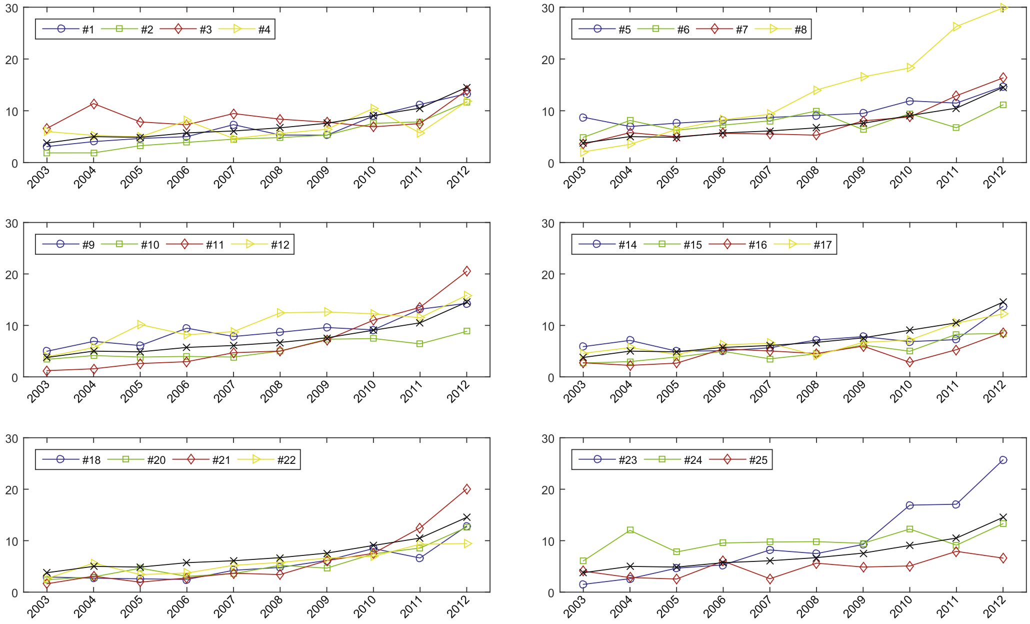


Fig. 6. The trend for the radical novelty indicator.

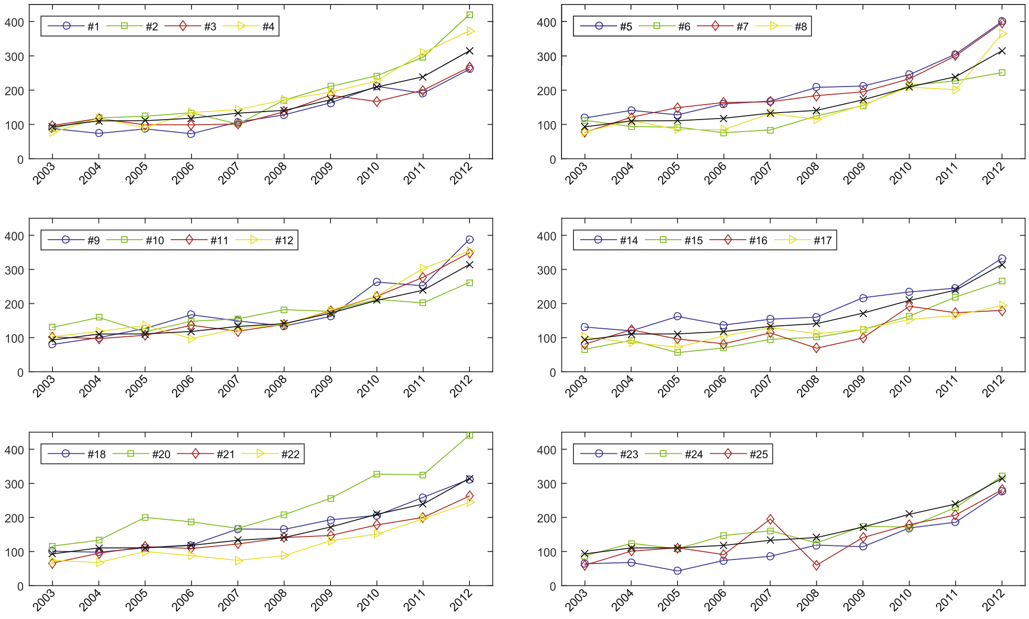


Fig. 7. The trend for the scientific impact indicator ($\tau = 0.80$).

used for subsequent analysis. Fig. 10 shows the top 15 terms that have the highest probability conditioned on each emerging technology, in which the terms with orange color are our submitted technical terms.

The following phenomena can be observed from Fig. 10: First, the term *synthetic biology* appears very frequently, so we speculate that the provided dataset comes from the *synthetic biology* field. Second, the term-based themes can be indeed discovered by the TNG model, but many n-grams are not technical terms, even not terms, such as *shed light, result suggest, synthetic promoter*, and *system biology*. Therefore, to nominate emerging technologies in term of technical terms, one should be careful to choose a proper technical term from the top n-grams list of each topic. The technical term candidates cannot be too general or too specialized. Another rule one should follow is that a fast growth trend can be observed for an interested technical term candidate in recent years.

As noted in Section 1, it is not trivial to evaluate quantitatively the performance of our framework, especially when the *ground truth* is unavailable. Fortunately, the contest organizer has obtained a *ground truth* file through combining title & abstract NLP phrases, and keywords-author & keywords-plus. To determine the final contest winner, entrants were also graded on the percentage of high growth terms out of total terms provided, the degree of growth of their terms, the frequency of those high growth terms across the dataset, and the quality of the terms based on utility, combined with a qualitative assessment by three judges of the model used to make the prediction. In the end, 8 terms (*binding affinity, fluorescence microscopy, metabolic engineering, direct evolution, dna assembly, genetic circuit, gene circuit*, and *gene regulatory network*) from 10 submitted terms met the contest organizer's criteria of technology emergence, and we obtained second place in the 2018–2019 Measuring Tech Emergence Contest. Furthermore, our methodology is a

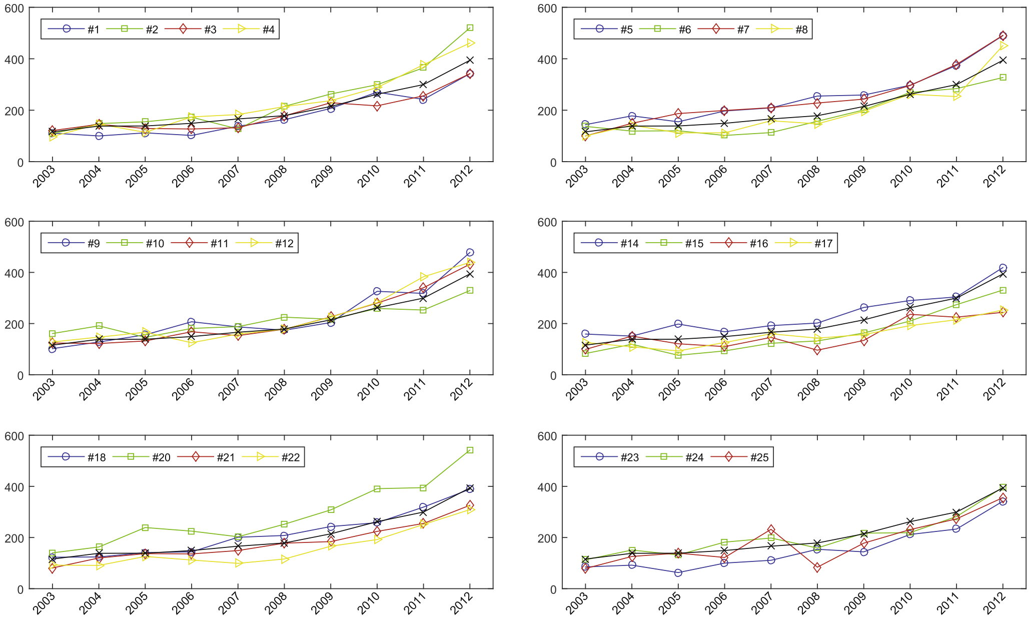


Fig. 8. The trend for the scientific impact indicator ($\tau = 0.85$).

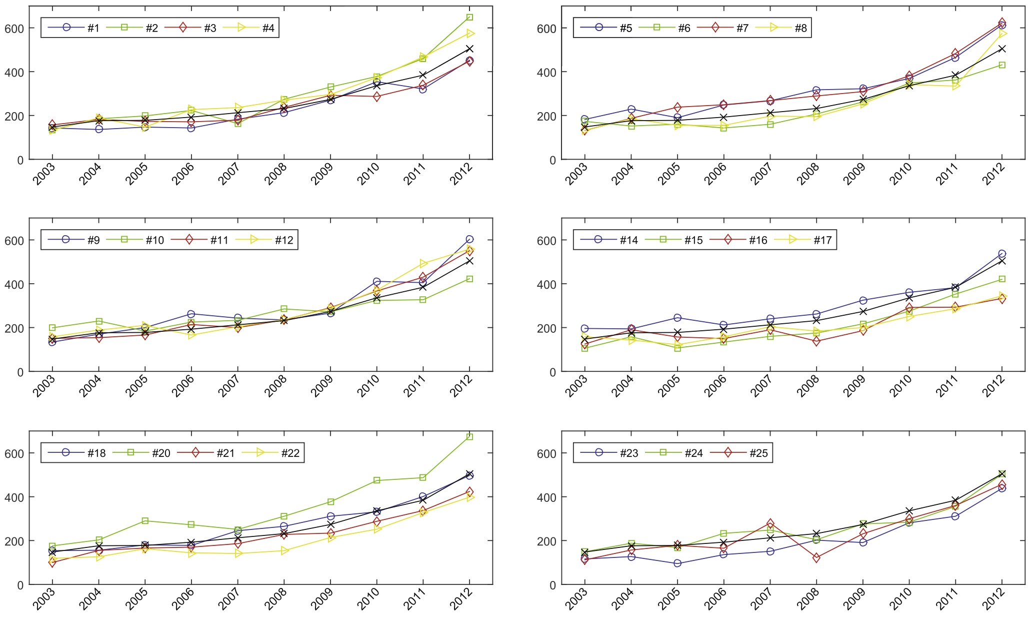


Fig. 9. The trend for the scientific impact indicator ($\tau = 0.90$).

Table 4
The correlation coefficients among the three indicators in term of Kendall's τ with two-tailed hypothesis.

	Growth	Novelty	Influence
Growth	1.0	0.4686	0.2309
Novelty	0.4686	1.0	0.6220
Influence	0.2309	0.6220	1.0

Table 5
The supports for all term-based themes from each indicator.

Topic	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Growth		•	•	•						
Novelty					•		•	•	•	
Influence		•		•	•		•	•	•	
Topic	#11	#12	#14	#15	#16	#17	#18	#20	#21	#22
Growth	•	•	•				•	•	•	
Novelty	•	•							•	
Influence	•	•	•							
Topic	#23	#24	#25							
Growth	•									
Novelty	•									
Influence										

repeatable procedure. The corresponding source codes can be made available from the first or corresponding author upon request for academic use.

5. Conclusions

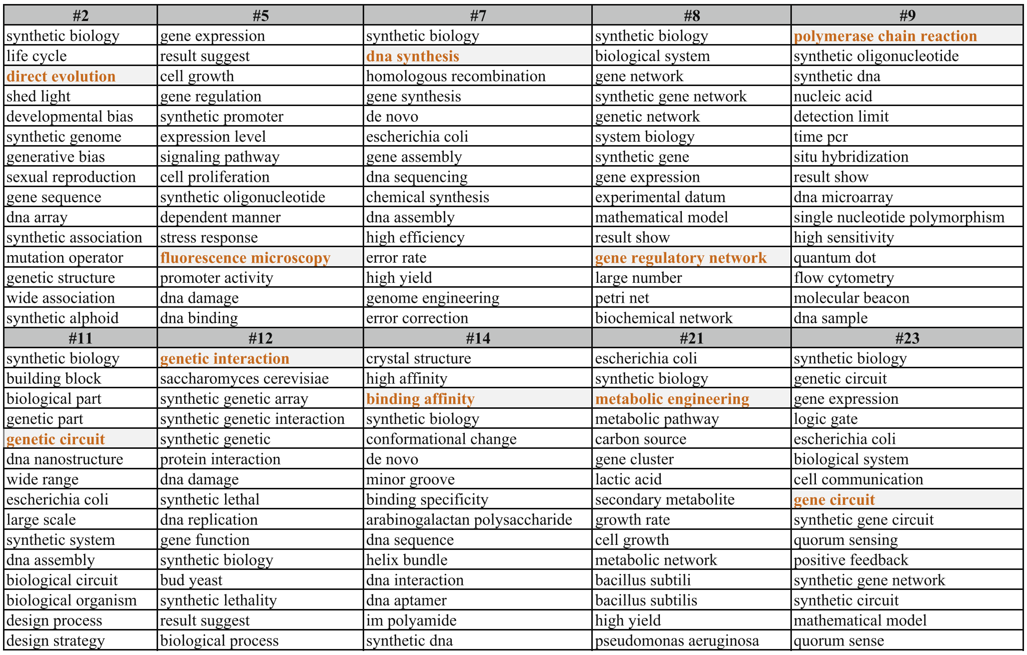
Studies on detecting and forecasting emerging technologies can help research foundations and policy-makers focus on research policy and management of innovation as well as individual research agendas. After more than 50 years of development, several major technologies have been developed. The challenge is that there are no benchmark datasets publicly available to improve further the accuracy of detection and forecasting. The 2018–2019 Measuring Tech Emergence Contest provides a more objective validation approach, which stimulates rethinking technology emergence to design novel and viable indicators.

It is well known that emerging technologies are closely related to emerging topics in terms of relatively fast growth, radical novelty and prominent impact. In this study, technologies are viewed as term-based themes. Therefore, our framework on detecting and forecasting emerging topics (Xu et al., 2019a) is generalized to measure technology emergence. In more details, our framework is flexible enough to accommodate emerging research topics and emerging technologies. For purpose of detecting and forecasting emerging technologies, the DIM model is replaced by the TNG model to discover the term-based themes and the prominent impact indicator is designed on the basis of the involving researchers.

Despite its usefulness, this study is subject to limitations. The term-based themes can indeed be discovered by the TNG model, but many n-grams are not technical terms, even not terms. Hence, to nominate emerging technologies in technical terms, one should be careful to choose a proper technical term from the top n-grams list of each topic. In addition, apart from word order and syntax information, it is better for the topic models to consider the theme life cycle, which is the birth, death, merging and branching of term-based themes. Thus, dramatic changes may be avoided in some indicators. The TNG model will be extended to consider the dynamics of topics in the near future. This study mainly focuses on three indicators: relatively fast growth, scientific impact, and radical novelty. Other characteristics of the emergence of technologies (such as uncertainty and ambiguity) still need to be considered to reinforce the framework. Finally, our case study has been limited to the field provided by the contest organizer. A scientific verification of our methodology still needs to be further investigated in our next work.

CRediT authorship contribution statement

Shuo Xu: Conceptualization, Methodology, Writing - original draft. **Liyuan Hao:** Data curation, Validation, Writing - review & editing. **Guancan Yang:** Methodology, Validation, Writing - review & editing. **Kun Lu:** Conceptualization, Validation, Writing - review & editing. **Xin An:** Supervision, Methodology, Software, Writing - review & editing.



#2	#5	#7	#8	#9
synthetic biology	gene expression	synthetic biology	synthetic biology	**polymerase chain reaction**
life cycle	result suggest	**dna synthesis**	biological system	synthetic oligonucleotide
direct evolution	cell growth	homologous recombination	gene network	synthetic dna
shed light	gene regulation	gene synthesis	synthetic gene network	nucleic acid
developmental bias	synthetic promoter	de novo	genetic network	detection limit
synthetic genome	expression level	escherichia coli	system biology	time pcr
generative bias	signaling pathway	gene assembly	synthetic gene	situ hybridization
sexual reproduction	cell proliferation	dna sequencing	gene expression	result show
gene sequence	synthetic oligonucleotide	chemical synthesis	experimental datum	dna microarray
dna array	dependent manner	dna assembly	mathematical model	single nucleotide polymorphism
synthetic association	stress response	high efficiency	result show	high sensitivity
mutation operator	**fluorescence microscopy**	error rate	**gene regulatory network**	quantum dot
genetic structure	promoter activity	high yield	large number	flow cytometry
wide association	dna damage	genome engineering	petri net	molecular beacon
synthetic alphoid	dna binding	error correction	biochemical network	dna sample
#11	**#12**	**#14**	**#21**	**#23**
synthetic biology	**genetic interaction**	crystal structure	escherichia coli	synthetic biology
building block	saccharomyces cerevisiae	high affinity	synthetic biology	genetic circuit
biological part	synthetic genetic array	**binding affinity**	**metabolic engineering**	gene expression
genetic part	synthetic genetic interaction	synthetic biology	metabolic pathway	logic gate
genetic circuit	synthetic genetic	conformational change	carbon source	escherichia coli
dna nanostructure	protein interaction	de novo	gene cluster	biological system
wide range	dna damage	minor groove	lactic acid	cell communication
escherichia coli	synthetic lethal	binding specificity	secondary metabolite	**gene circuit**
large scale	dna replication	arabinogalactan polysaccharide	growth rate	synthetic gene circuit
synthetic system	gene function	dna sequence	cell growth	quorum sensing
dna assembly	synthetic biology	helix bundle	metabolic network	positive feedback
biological circuit	bud yeast	dna interaction	bacillus subtili	synthetic gene network
biological organism	synthetic lethality	dna aptamer	bacillus subtilis	synthetic circuit
design process	result suggest	im polyamide	high yield	mathematical model
design strategy	biological process	synthetic dna	pseudomonas aeruginosa	quorum sense

Fig. 10. The emerging technologies with the resulting technical terms, in which the highlighted terms are submitted to the contest.

Acknowledgements

This research received financial support from the Fundamental Research Funds for the Central Universities under grant number 2018RW26, from the Social Science Foundation of Beijing Municipality under grant number 17GLB074, and from the National Natural Science Foundation of China under grant number 72074014 and 72004012. Many thanks to Professors Zhengyin Hu and Haiyun Xu from Chengdu Library and Information Center, Chinese Academy of Sciences for their valuable consultation and review of this manuscript. Our gratitude also goes to the anonymous reviewers for their valuable comments.

Appendix A

The predicative likelihood of a word vector can in principle be calculated by integrating out all parameters from the joint distribution of the word observations in a document. For the TNG model, the likelihood of a text document of the test corpus $\Pr(\vec{\tilde{w}}_{m,\cdot}|\mathcal{M})$ can be directly expressed as a function of the multinomial and Binomial parameters:

$$\begin{aligned}\Pr(\vec{\tilde{w}}_{m,\cdot}|\mathcal{M}) &= \sum_{k=1}^{K}\Pr\left(\vec{\tilde{w}}_{m,\cdot}|z_{m,N_m}=k,\ \mathcal{M}\right)\Pr\left(z_{m,N_m}=k|\mathcal{M}\right)\\ &= \sum_{k=1}^{K}\mathrm{P}\left(\vec{\tilde{w}}_{m,\cdot}|z_{m,N_m}=k,\ \mathcal{M}\right)\tilde{\vartheta}_{m,k}\end{aligned} \tag{A.1}$$

The term $\Pr(\vec{\tilde{w}}_{m,\cdot}|z_{m,N_m}=k,\ \mathcal{M})$ in Eq. (A.1) can be decomposed according to the assumption of the TNG model:

$$\begin{aligned}&\Pr(\vec{\tilde{w}}_{m,\cdot}|z_{m,N_m}=k,\ \mathcal{M})\\ =&\sum_{k'=1}^{K}\sum_{b=0}^{1}\left[\Pr\left(\vec{\tilde{w}}_{m,1:N_m-1}|z_{m,N_m-1}=k',\ \mathcal{M}\right)\times\tilde{\vartheta}_{m,k'}\times\psi_{k',\tilde{w}_{m,N_m-1},b}\right.\\ &\left.\times\begin{cases}\varphi_{k,\tilde{w}_{m,N_m}}, & b=0\\ \phi_{k,\tilde{w}_{m,N_m-1},\tilde{w}_{m,N_m}}, & b=1\end{cases}\right]\end{aligned} \tag{A.2}$$

It is easy to see that Eq. (A.2) is a recursive equation which can be simplified into the ρ function in Eq. (10)-(11). From the ρ function, the likelihood [Eq. (9)] can be readily obtained.

References

AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C., 2009. Topic significance ranking of LDA generative models. In: Goebel, R., Siekmann, J., Wahlster, W. (Eds.), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 67–82. https://doi.org/10.1007/978-3-642-04180-8_22.

An, X., Xu, S., 2019. Topical analysis of scientific and technical reports based on topical n-grams model. Agric. Lib. Inf. 31 (6), 21–30. https://doi.org/10.13998/j.cnki.issn1002-1248.2019.06.19-0550.

An, X., Xu, S., Wen, Y., Hu, M., 2014. A shared interest discovery model for co-author relationship in SNS. Int. J. Distribut. Sensor Netw. 2014, 1–9. https://doi.org/10.1155/2014/820715.

Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I., 2003. An introduction to MCMC for machine learning. Mach. Learn. 50 (1-2), 5–43. https://doi.org/10.1023/A:1020281327116.

Apreda, R., Bonaccorsi, A., dell'Orletta, F., Fantoni, G., 2019. Expert forecast and realized outcomes in technology foresight. Technol. Forecast. Soc. Change 141, 277–288. https://doi.org/10.1016/j.techfore.2018.12.006.

Arora, S.K., Youtie, J., Shapira, P., Gao, L., Ma, T., 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms. Scientometrics 95 (3), 1189–1207. https://doi.org/10.1007/s11192-013-0950-7.

Azzonpardi, L., Girolami, M., van Risjbergen, K., 2003. Investigating the relationship between language model perplexity and IR precision-recall measures. Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, USA, pp. 369–370. https://doi.org/10.1145/860435.860505.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Bolger, F., Wright, G., 2017. Use of expert knowledge to anticipate the future: Issues, analysis and directions. Int. J. Forecast. 33 (1), 230–243. https://doi.org/10.1016/j.ijforecast.2016.11.001.

Boyack, K.W., Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? J. Am. Soc. Inf. Sci. 61 (12), 2389–2404. https://doi.org/10.1002/asi.21419.

Boyack, K.W., Klavans, R., Small, H., Ungar, L., 2014. Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. J. Eng. Technol. Manag. 32, 147–159. https://doi.org/10.1016/j.jengtecman.2013.07.001.

Burmaoglu, S., Sartenaer, O., Porter, A., Li, M., 2019. Analysing the theoretical roots of technology emergence: an evolutionary perspective. Scientometrics 119 (1), 97–118. https://doi.org/10.1007/s11192-019-03033-y.

Carley, S.F., Newman, N.C., Porter, A.L., Garner, J.G., 2018. An indicator of technical emergence. Scientometrics 115 (1), 35–49. https://doi.org/10.1007/s11192-018-2654-5.

Caron, E., van Eck, N.-J., 2014. Large scale author name disambiguation using rule-based scoring and clustering. Proceedings of the 19th International Conference on Science and Technology Indicators. pp. 79–86.

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. https://doi.org/10.1002/asi.20317.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., Yang, G., 2020. A deep learning based method for extracting semantic information from patent documents. Scientometrics 125 (1), 289–312. https://doi.org/10.1007/s11192-020-03634-y.

Cronin, B., 2001. Hyperauthorship: a postmodern perversion or evidence of a structural shift in scholarly communication practices? J. Associat. Inf. Sci. Technol. 52 (7), 558–569. https://doi.org/10.1002/asi.1097.

Day, G.S., Schoemaker, P.J.H., Gunther, R.E., 2000. Wharton on Managing Emerging Technologies. Wiley & Sons, Inc, New York, NY, USA.

Dietz, L., Bickel, S., Scheffer, T., 2007. Unsupervised prediction of citation influences. Proceedings of the 24th International Conference on Machine Learning. ACM, pp. 233–240. https://doi.org/10.1145/1273496.1273526.

Gauffriau, M., Larsen, P.O., 2005. Counting methods are decisive for rankings based on publication and citation studies. Scientometrics 64 (1), 85–93. https://doi.org/10.1007/s11192-005-0239-6.

Gerrish, S.M., Blei, D.M., 2010. A language-based approach to measuring scholarly impact. Proceedings of the 27th International Conference on Machine Learning. Omnipress, pp. 375–382.

Glänzel, W., Thijs, B., 2012. Using 'core documents' for detecting and labelling new emerging topics. Scientometrics 91 (1), 399–416. https://doi.org/10.1007/s11192-011-0591-7.

González-Alcaide, G., Llorente, P., Ramos, J.M., 2016. Bibliometric indicators to identify emerging research fields: Publications on mass gatherings. Scientometrics 109 (2), 1283–1298. https://doi.org/10.1007/s11192-016-2083-2.

Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B., 2005. Integrating topics and syntax. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA, pp. 537–544.

Griffiths, T.L., Steyvers, M., Tenenbaum, J.B., 2007. Topics in semantic representation. Psychol. Rev. 114 (2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211.

Guo, H., Weingart, S., Börner, K., 2011. Mixed-indicators model for identifying emerging research areas. Scientometrics 89 (1), 421–435. https://doi.org/10.1007/s11192-011-0433-7.

Hagen, N.T., 2008. Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. PLoS ONE 3 (12), e4021. https://doi.org/10.1371/journal.pone.0004021.

Hagen, N.T., 2013. Harmonic coauthor credit: A parsimonious quantification of the byline hierarchy. J. Informetric. 7 (4), 784–791. https://doi.org/10.1016/j.joi.2013.06.005.

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., Xu, S., 2017. Semantic fingerprints-based author name disambiguation in Chinese documents. Scientometrics 111 (3), 1879–1896. https://doi.org/10.1007/s11192-017-2338-6.

Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J., 2013. Stochastic variational inference. J. Mach. Learn. Res. 14 (May), 1303–1347.

Huang, M.-H., Chang, C.-P., 2014. Detecting research fronts in OLED field using bibliographic coupling with sliding window. Scientometrics 98 (3), 1721–1744. https://doi.org/10.1007/s11192-013-1126-1.

Jarneving, B., 2007. Bibliographic coupling and its application to research-front and other core documents. J. Inf. 1 (4), 287–307. https://doi.org/10.1016/j.joi.2007.07.004.

Jordan, M., Grhahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Mach. Learn. 37 (2), 183–233. https://doi.org/10.1023/A:1007665907178.

Joung, J., Kim, K., 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technol. Forecast. Soc. Change 281–292. https://doi.org/10.1016/j.techfore.2016.08.020.

Ke, Q., 2020. Technological impact of biomedical research: the role of basicness and novelty. Res. Policy 49 (7), 104071. https://doi.org/10.1016/j.respol.2020.104071.

Kim, J., 2018. Evaluating author name disambiguation for digital libraries: a case of DBLP. Scientometrics 116 (3), 1867–1886. https://doi.org/10.1007/s11192-018-2824-5.

Kim, J., Kim, J., 2015. Rethinking the comparison of authorship credit allocation schemes. J. Informetric. 9 (3), 667–673. https://doi.org/10.1016/j.joi.2015.07.005.

Kong, D., Zhou, Y., Liu, Y., Xue, L., 2017. Using the data mining method to assess the innovation gap: a case of industrial robotics in a catching-up country. Technol. Forecast. Soc. Change 119, 80–97. https://doi.org/10.1016/j.techfore.2017.02.035.

Kyebambe, M.N., Cheng, G., Huang, Y., He, C., Zhang, Z., 2017. Forecasting emerging technologies: a supervised learning approach through patent analysis. Technol. Forecast. Soc. Change 125, 236–244. https://doi.org/10.1016/j.techfore.2017.08.002.

Lee, C., Kwon, O., Kim, M., Kwon, D., 2018. Early identification of emerging technologies: a machine learning approach using multiple patent indicators. Technol. Forecast. Soc. Change 127, 291–303. https://doi.org/10.1016/j.techfore.2017.10.002.

Levin, M., Krawczyk, S., Bethard, S., Jurafsky, D., 2012. Citation-based bootstrapping for large-scale author disambiguation. J. Assoc. Inf. Sci. Technol. 63 (5), 1030–1047. https://doi.org/10.1002/asi.22621.

Li, M., 2017. An exploration to visualise the emerging trends of technology foresight based on an improved technique of co-word analysis and relevant literature data of WOS. Technol. Anal. Strategic Manag. 29 (6), 655–671. https://doi.org/10.1080/09537325.2016.1220518.

Liu, X., Wang, S., 2010. Development of an in vivo computer for the SAT problem. Math. Comput. Modell. 52 (11–12), 2043–2047. https://doi.org/10.1016/j.mcm.2010.06.006.

Liu, X., Wang, S., Qiang, X., 2009. Development of an in vivo computer for 3-SAT problem. Proceedings of the 4th International Conference on Bio-Inspired Computing. pp. 328–331. https://doi.org/10.1109/BICTA.2009.5338084.

Lu, X., Yang, G., Xu, S., Zhang, Y., 2020. Review of research progress on emerging technologies identification based on quantitative and evolutionary perspectives. J. China Soc. Sci. Technic. Inf. 39 (6), 651–661. https://doi.org/10.3772/j.issn.1000-0135.2020.06.009.

Ma, R., 2012. Author bibliographic coupling analysis: a test based on a Chinese academic database. J. Informetric. 6 (4), 532–542. https://doi.org/10.1016/j.joi.2012.04.006.

Osório, A., 2018. On the impossibility of a perfect counting method to allocate the credits of multi-authored publications. Scientometrics 116 (3), 2161–2173. https://doi.org/10.1007/s11192-018-2815-6.

Porter, A.L., Garner, J., Carley, S.F., Newman, N.C., 2019. Emergence scoring to identify R&D topics and key players. Technol. Forecast. Soc. Change 146, 628–643. https://doi.org/10.1016/j.techfore.2018.04.016.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: the Art of Scientific Computing, 2nd edition. Cambridge University Press, New York, USA.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286. https://doi.org/10.1109/5.18626.

Rafols, I., Porter, A.L., Leydesdorff, L., 2010. Science overlay maps: a new tool for research policy and library management. J. Assoc. Inf. Sci. Technol. 61 (9), 1871–1887. https://doi.org/10.1002/asi.21368.

Robinson, D.K.R., Lagnau, A., Boon, W.P.C., 2019. Innovation pathways in additive manufacturing: methods for tracing emerging and branching paths from rapid prototyping to alternative applications. Technol. Forecast. Soc. Change 146, 733–750. https://doi.org/10.1016/j.techfore.2018.07.012.

Rotolo, D., Hicks, D., Martin, B.R., 2015. What is an emerging technology? Res. Policy 44 (10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006.

Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., Ohta, T., 2007. AKANE system: protein-protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask. Proceedings of the 2nd BioCreative Challenge Evaluation Workshop. pp. 209–212.

Schwartz, A.S., Hearst, M.A., 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. Proceedings of the Pacific Symposium on Biocomputing. pp. 451–462. https://doi.org/10.1142/9789812776303_0042.

Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K., 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28 (11), 758–775. https://doi.org/10.1016/j.technovation.2008.03.009.

Small, H., 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. 24 (4), 265–269. https://doi.org/10.1002/asi.4630240406.

Small, H., Boyack, K.W., Klavans, R., 2014. Identifying emerging topics in science and technology. Res. Policy 43 (8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005.

Small, H., Griffith, B.C., 1974. The structure of scientific literatures I: identifying and graphing specialties. Sci. Stud. 4 (1), 17–40.

de Solla Price, D.J., 1965. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. Science 149 (3683), 510–515. https://doi.org/10.1126/science.149.3683.510.

Sommarberg, M., Mäkinen, S., 2019. A method for anticipating the disruptive nature of digitalization in the machine-building industry. Technol. Forecast. Soc. Change 146, 808–819. https://doi.org/10.1016/j.techfore.2018.07.044.

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.23596.

Takeda, Y., Kajikawa, Y., 2009. Optics: A bibliometric approach to detect emerging research domains and intellectual bases. Scientometrics 78 (3), 543–558. https://doi.org/10.1007/s11192-007-2012-5.

Torvik, V.I., Smalheiser, N.R., 2009. Author name disambiguation in MEDLINE. ACM Trans. Knowl. Discov. Data 3 (3), 11:1–11:29. https://doi.org/10.1145/1552303.1552304.

Tscharntke, T., Hochberg, M.E., Rand, T.A., Resh, V.H., Krauss, J., 2007. Author sequence and credit for contributions in multiauthored publicaitons. PLoS Biol. 5 (1), e18. https://doi.org/10.1371/journal.pbio.0050018.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J., 2005. Developing a robust part-of-speech tagger for biomedical text. In: Bozanis, P., Houstis, E.N. (Eds.), Proceedings of the 10th Panhellenic Conference on Informatics, pp. 382–392. https://doi.org/10.1007/11573036_36.

Veugelers, R., Wang, J., 2019. Scientific novelty and technological impact. Res. Policy 48 (6), 1362–1372. https://doi.org/10.1016/j.respol.2019.01.019.

Wallach, H.M., 2006. Topic modeling: Beyond bag-of-words. Proceedings of the 23rd International Conference on Machine Learning. pp. 977–984. https://doi.org/10.1145/1143844.1143967.

Waltman, L., van Eck, N.J., 2012. A new methodology for constructing a publication-level classification system of science. J. Assoc. Inf. Sci. Technol. 63 (12), 2378–2392. https://doi.org/10.1002/asi.22748.

Wang, Q., 2018. A bibliometric model for identifying emerging researh topics. J. Assoc. Inf. Sci. Technol. 69 (2), 290–304. https://doi.org/10.1002/asi.23930.

Wang, X., McCallum, A., Wei, X., 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. Proceedings of the 7th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, pp. 697–702. https://doi.org/10.1109/ICDM.2007.86.

Wang, Z., Xu, S., Zhu, L., 2018. Semantic relation extraction aware of n-gram features from unstructured biomedical text. J. Biomed. Inf. 86, 59–70. https://doi.org/10.1016/j.jbi.2018.08.011.

Weismayer, C., Pezenka, I., 2017. Identifying emerging research fields: a longitudinal latent semantic keyword analysis. Scientometrics 113 (3), 1757–1785. https://doi.org/10.1007/s11192-017-2555-z.

Xu, J., Ding, Y., Song, M., Chambers, T., 2016. Author credit-assignment schemas: a comparison and analysis. J. Assoc. Inf. Sci. Technol. 67 (8), 1973–1989. https://doi.org/10.1002/asi.23495.

Xu, S., An, X., Zhu, L., Zhang, Y., Zhang, H., 2015. A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. J. Cheminformatic. 7 (Suppl 1), S11. https://doi.org/10.1186/1758-2946-7-S1-S11.

Xu, S., Hao, L., An, X., Pang, H., Li, T., 2020. Review on emerging research topics with key-route main path analysis. Scientometrics 122 (1), 607–624. https://doi.org/10.1007/s11192-019-03288-5.

Xu, S., Hao, L., An, X., Yang, G., Wang, F., 2019. Emerging research topics detection with multiple machine learning models. J. Informetric. 13 (4). https://doi.org/10.1016/j.joi.2019.100983.

Xu, S., Hao, L., An, X., Zhai, D., Pang, H., 2019. Types of DOI errors of cited references in Web of Science with a cleaning method. Scientometrics 120 (3), 1427–1437. https://doi.org/10.1007/s11192-019-03162-4.

Xu, S., Liu, J., Wang, Z., 2017. Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. Proceedings of the 16th International Conference on Scientometrics & Informetrics. pp. 1007–1012.

Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., Pang, H., 2018. Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. Scientometrics 117 (1), 61–84. https://doi.org/10.1007/s11192-018-2841-4.

Xu, S., Zhai, D., Wang, F., An, X., Pang, H., Sun, Y., 2019. A novel method for topic linkages between scientific publications and patents. J. Assoc. Inf. Sci. Technol. 70 (9), 1026–1042. https://doi.org/10.1002/asi.24175.

Yau, C.-K., 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.

Zhang, Y., Huang, Y., Porter, A.L., Zhang, G., Lu, J., 2019. Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. Technol. Forecast. Soc. Change 146, 795–807. https://doi.org/10.1016/j.techfore.2018.06.007.

Zhang, Y., Robinson, D.K.R., Porter, A.L., Zhu, D., Zhang, G., Lu, J., 2016. Technology roadmapping for competitive technical intelligence. Technologic. Forecast. Soc. Change 110, 175–186. https://doi.org/10.1016/j.techfore.2015.11.029.

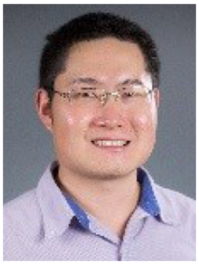
Zhao, D., Strotmann, A., 2008. Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. J. Assoc. Inf. Sci. Technol. 59 (13), 2070–2086. https://doi.org/10.1002/asi.20910.

Zhou, X., Huang, L., Porter, A., Vicente-Gomila, J.M., 2019. Tracing the system transformations and innovation pathways of an emerging technology: solid lipid nanoparticles. Technol. Forecast. Soc. Change 146, 785–794. https://doi.org/10.1016/j.techfore.2018.04.026.

Shuo Xu works as Professor in the Management Science and Engineering of College of Economics and Management at Beijing University of Technology. He received Ph.D. degree from Agricultural University of China in 2008. His research interests include scientific fronts detection, technology foresight, knowledge management, data mining and big data.

Liyuan Hao is a postgraduate student in the School of Economics and Management, Beijing University of Technology and majors in information management and information system.

Guancan Yang works as Assistant Professor in the Information Resource Management School at the Renmin University. He received Ph.D. degree from Wuhan University in 2013. His research interests include technology competitive intelligence, and technology evaluation and forecasting.



Kun Lu is an Associate Professor in the School of Library and Information Studies at the University of Oklahoma. He received his Ph.D. degree from the University of Wisconsin-Milwaukee in 2012. His research interests including information retrieval, text mining and organization of information.

Xin An works as Associated Professor of School of Economics and Management at Beijing Forestry University. She obtained her M.S. from China Agriculture University, and Ph.D. from University of International Business and Economics. Her current research interest includes technology foresight, data mining, and statistics.