Scientometrics (2022) 127:5415–5440
https://doi.org/10.1007/s11192-021-04253-x

Exploring funding patterns with word embedding-enhanced organization–topic networks: a case study on big data

Qianqian Jin[1] · Hongshu Chen[1] · Ximeng Wang[2] · Tingting Ma[3] · Fei Xiong[4]

Received: 5 April 2021 / Accepted: 20 December 2021 / Published online: 5 January 2022

Abstract
Understanding the complex patterns in research funding plays a fundamental role in comprehensively revealing funding preferences and informing ideas for future strategic innovation. This is especially true when the funding policies need to be constantly shifted to accommodate highly complex and ever-changing demands for technological, economic, and social development. To this end, we investigate the associations between funding agencies and the topics they fund in an attempt to understand funding patterns at both an organizational level and a topic level. In this paper, the links between heterogeneous nodes, organizations and topics, are mapped to a two-mode organization–topic network. The collaborative interactions formed by funding organizations and the semantic networks constituted by word embedding-enhanced topics are revealed and analyzed simultaneously. The methodology is demonstrated through a case study on big data research involving 9882 articles from the Web of Science over the period 2010 to 2019. The result shows a comprehensive picture of the topics that governments, academic institutions, and industrial funding organizations prefer to fund, which provide potential decision support for agencies and organizations who are exploring funding patterns, estimating funding trends, and updating their funding strategies.

Keywords Funding patterns · Funded topics · Word embedding · Two-mode · Organization–topic networks

Introduction

Research funding supports high-impact studies, benefits knowledge production, and accelerates technological advancement (Gao et al., 2019; Wang et al., 2012). However, with an increase in the level and cost of research being conducted around the world, choosing

✉ Hongshu Chen
hongshu.chen@bit.edu.cn

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, China

[2] Cyber Finance Department, Postal Savings Bank of China, Beijing, China

[3] School of Logistics, Beijing Wuzi University, Beijing, China

[4] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

research projects and allocating funding rationally is not always easy (Aagaard et al., 2021). As the diverse requirements of technology, economic and social development become ever more complex, funding policies bend and flex to accommodate, reshaping themselves into something new and multifaceted. Exploring and understanding the complex patterns in research funding is beneficial for both academics and policymakers.

Today, research is experiencing unprecedented attention and support from funding organizations (Zeng et al., 2017). Likewise, knowledge is being disseminated, exchanged, merged, and split at a breakneck pace (Brennecke & Rank, 2017). As we know, innovation originates from the recombination of existing knowledge and ideas, and one of the influences on this recombination is research funding (Colatat, 2015). On one hand, funding activities illustrate funding priorities of different agencies; the funded topics, on the other hand, reveal deeper knowledge of dynamics and interactions of these research attentions. In providing financial support for research, funding organizations can become associated with different knowledge elements, and, taking a network view of the funding landscape, they begin to occupy unique positions in the fabric of the research community. In other words, funding patterns are doubly embedded in an agency's funding activities, and co-funded topics constitute a form of coupling between knowledge elements (Stahlman & Heidorn, 2020). Therefore, the ability to recognize the patterns, characteristics, and evolutionary trends of funded topics is of great theoretical and practical significance, for it not only comprehensively reveals funding preferences but can also inform planning ideas for future strategic innovation.

Comparatively little attention has been paid to analyzing funding patterns that involve funding organizations and funded topics simultaneously—although analyses based on funded subjects, fields and keywords have been conducted. Over the past decade, methods of topic extraction and exploration from unstructured text data have been enhanced via the power of modern computing (Chen et al., 2021; De Battisti et al., 2015; Huang & Huang, 2018). From a topic extraction perspective, word embedding has caught researchers' eyes for its promising ability to map words into numeric vectors. With these machine learning techniques, researchers have been able to discover the latent semantics of large-scale text data while considering the context of the content (Mikolov, Sutskever, et al., 2013). Here, vectors replace traditional word representations, which can be manipulated in ways words cannot to support scientific text mining. This evolving branch of data/information science holds great potential for topic extraction and analysis (Chen et al., 2022; Zhang et al., 2018).

With this momentum of interest and aiming to fill the gap of exploring funding patterns at both the organizational and content levels, we have designed a word embedding-enhanced organization–topic network that reveals the pivotal role of different funding agencies in a funded topic network. The network includes government, academic, and industrial funding organizations as well as funded topics, extracted and vectorized from the text of research articles. Then the links between homogeneous nodes (organizations and topics) are mapped to an integrated organization–topic network that comprehensively reveals both the collaborative interactions formed by funding organizations and the semantic networks constituted by word embedding-enhanced topics. A case study on big data research demonstrates the feasibility and effectiveness of the methodology.

The remainder of this paper is organized as follows. Section "Literature review" presents relevant studies on research funding analysis, network perspectives for actor–content analysis and then reviews topic extraction and embedding based textual mining in scientometric research. Section "Methodology" provides a detailed description of the methodology. We then conduct an empirical case study on the field of big data in "Empirical case

study: funding patterns exploration in the field of big data" section. Finally, "Conclusions and future work" section concludes with the contributions, limitations, and research prospects of this study.

Literature review

Research funding analysis in scientometric studies

Research funding gives academics access to advanced equipment, favorable working conditions, and can help to foster collaboration (Grimpe, 2012; Huang & Huang, 2018). However, with an increase in the level and cost of research being conducted around the world, securing adequate funding for one's work is not always easy (Aagaard et al., 2021). Any funding gaps may mean we are slower to integrate knowledge, slower to reorganize and recombine technologies, and slower to progress science as a whole. In turn, this leaves science less able to deal with societal challenges (Munari & Toschi, 2021). Hence, to make rational use of limited funds and promote scientific innovation, we need knowledge of the current funding patterns—i.e., who is funding what research. With such information, we are better able to optimize funding strategies.

Research on funding is one of the most important topics in scientometrics. It spans studies on everything from funding sources to targets and returns. For example, academics like Zhao et al. (2018) have investigated the impact of funded publications using indicators such as an article's usage count or citation rate, while others, like Wang et al. (2018), have evaluated the creativity of scientific output. Several studies have attempted to uncover the collaboration characteristics of funded research, examining collaboration among institutes, countries (Zhou & Tian, 2014), and disciplines (Huang et al., 2016). More recently, the research on funding patterns that has found itself in the limelight has come from a funded themes perspective. For example, Mejia and Kajikawa (2018) classified robotics research into four categories, exploring the source and quantity of funds that each category received. Zhao et al. (2019) focused on Google, tracing how the topics it has funded have evolved. Stahlman and Heidorn (2020) looked at the allocation of research funding among different topics, as well as the scientific outputs of those topics. Impressively, they found that there were no significant relationships between investing in a topic and research outputs.

With the growing interest in research funding analysis, the use of funding acknowledgment data has also received special attention. Some academics claim that the data quality and reliability of individual bibliographic databases is still underexplored (Liu, 2020), while others emphasize the need to analyzing funding results with a critical eye. For example, in Web of Science (WoS), funding acknowledgments have systematically been collected for SCIE since 2009, for SSCI since 2015, and for A&HCI since 2017 (Liu et al., 2020; Tang et al., 2017). Such matters should be noticed in advance, otherwise one might draw misleading conclusions when comparing the funding characteristics in different fields.

Network perspective for actor–content analysis

According to Zeng et al. (2017), scientific research is promoted by the emergence of cutting-edge techniques and research tools, the integration of ideas and theories from multiple disciplines and, moreover, attention and support from funding organizations. Funding

activities indicate the preference of different organizations in supporting research priorities, while funded topics provide insights into dynamics and interactions of these research attentions. From this point of view, the funding agencies are the actors in the network, and the topics they fund are the 'content' of their activities.

Nowadays, networks are playing an increasingly important role in helping to characterize the features of knowledge integration that are embedded in social connections (Phelps et al., 2012). In this vein, multiplex and heterogeneous networks are attracting particular attention, for they can not only reflect research themes and visualize a domain's knowledge structures, but they can also provide insights into the relations between knowledge elements and actors which affect knowledge integration. For example, Hellsten and Leydesdorff (2020) mapped the interactions between topics and individual disseminators in a socio-semantic network, while Guan et al. (2017) detected focal technology fields in a country–technology network. Further, Brennecke and Rank (2017) constructed a multiplex patent–inventor network to analyze the advice transfers among corporate inventors. Since research funding influences the direction of innovation (Colatat, 2015), understanding the structural properties of knowledge and funding organizations within networks is highly worthwhile.

In short, the past few years have witnessed a growing interest in heterogeneous network analysis. However, relatively few academics have shown concern for the positions funding organizations hold in a knowledge network. Additionally, most studies base their knowledge elements on either the co-occurrence of keywords in scientific articles, or the co-occurrence of international patent classification (IPC) codes in patents (Chang, 2017; Chen et al., 2020). All these methods have shortcomings when it comes to revealing semantic information about the content of the research. New methods are needed if we are to gain insight into this important aspect of research funding.

Topic exaction and analysis in scientific text mining

Topic modeling is the process of extracting latent semantic information from a large collection of textual data. It is based on the supposition that each document is a mixture of topics, and each topic is a mixture of words, and that both follow certain probabilities when generated (Chen et al., 2021). As one of the most frequently used topic models, latent Dirichlet allocation (LDA) provides an effective tool for turning bibliometrics and scientific literature into comprehensible knowledge. (Blei et al., 2003). For example, Lamba and Madhusudhan (2019) used an LDA model to reveal the main research topics and development trends in the disciplines within library and information technology. Song and Suh (2019) collected patents and used an LDA model to probe technological emergence and integration in the field of industrial safety. Impressively, LDA has not only contributed to detecting and tracking topics in a given period or research area, it has also served as a useful tool for: indirectly evaluating the competitiveness of scientific research institutes (Ma et al., 2018); comparing the development strategies of different regions in a certain field (Naumanen et al., 2019); providing an indicator for predicting highly-cited literature (Hu et al., 2020); and many other studies. More recently, Stahlman and Heidorn (2020) used LDA to detect whether investment in a topic has an overall impact on the research output produced. In the process, they establish linkages between metadata and the results of the LDA analysis,

However, although LDA can extract interpretable representations over documents, it fails to capture syntactic and semantic information at a word level (Moody, 2016). Here,

the techniques of word embedding, as typified by the word2vec model (Mikolov, Chen, et al., 2013), perfectly overcome this shortcoming. There are two algorithms for word-2vec model: Skip-Gram and continuous bag-of-words (CBOW). Skip-Gram predicts the context of a given word, CBOW predicts a word given a context. In recent years, word-2vec has garnered substantial interest as a tool for scientific text mining. For example, Lee et al. (2020) combined and learned word2vec with WordNet for semantic relatedness and similarity measurement. Greiner-Petter et al. (2020) applied word2vec to math documents and fill the gap of math language processing and semantic knowledge extraction. Hu et al. (2019) turned keywords into vectors using the indicators of spatial autocorrelation to track topic evolutions geographically.

Yet, to our best knowledge, scientists have seldom integrated LDA and word2vec when extracting topics from scientific literature. Further scientometric studies are needed to showcase the superior insights to be gleaned from this methodology.

Methodology

In this paper, we propose a methodology to map the nodes and their links into an integrated organization–topic network. The network then comprehensively reveals both the collaborative interactions formed by the funding organizations and the semantic networks constituted by word embedding-enhanced topics. The methodology contains four main modules: (1) Data pre-processing; (2) Topic modeling and parameter setting; (3) Topic vectorization with word embedding; and (4) Organization–topic network construction and analysis.

Data pre-processing

We collected articles from the WoS database by searching for any article with an organization in a specific field acknowledging funding support.[1] The metadata and textual data were processed separately. The title and abstract fields were then combined as a corpus, while funding information and other bibliometric fields were stored in a separate file. To support the downstream topic modeling and word embedding tasks, we cleaned the corpus by removing punctuation, single letters and numbers, and eliminated stop words and commonly used academic words (Chen et al., 2021). All words were then lemmatized and consolidated by returning: plural nouns to singular form; inflected forms of verbs to their lemma; and comparative adjectives to their basic form.

To extract funding information from the metadata, we processed the field containing details of the funding organization. Organization names were disambiguated and standardized based on similarity using VantagePoint. To ensure data quality and integrity, we applied both natural language processing techniques and manual checks to further clean and consolidate the data. For example, some of the funding acknowledgments only provided abbreviated names. In other cases, organizations had changed their names over time. Once the organization names were cleaned and consolidated, we tagged them as government, academic, and industrial organizations according to their organizational functions.

[1] The search strategy for funding organization: FO = (a* or b* or c* or d* or e* or f* or g* or h* or i* or j* or k* or l* or m* or n* or o* or p* or q* or r* or s* or t* or u* or v* or w* or x* or y* or z* or 1* or 2* or 3* or 4* or 5* or 6* or 7* or 8* or 9* or 0*).

We followed existing *definitions* to help classify the GA&I organizations using NLP-based classification (Grimpe, 2012; Huang & Huang, 2018; Leydesdorff, 2003), and followed up with manual adjustments where necessary, as follows:

(1) *Government funding agencies* (*G*) this category covers ministry or subsidiary units of the national and local governments with particular areas of responsibility. Agencies such as government-sponsored public bodies and intergovernmental organizations were included in this category as well.
(2) *Academic institutions* (*A*) these comprise universities, educational centers, and scientific institutes. Even though some institutes are subordinate to government departments, they were deemed as academic agencies if they mainly participate in educational activities or scientific research; government funding is based more on direct calls for applications and pays more attention to societal needs (Guerzoni et al., 2014).
(3) *Industrial firms* (*I*) this group comprises private companies, hospitals, and individual groups who provide products or services for profit.

We then tagged the organization names according to a thesaurus of GA&I features. For example, terms such as "Bureau", "Municipal", and "Department" were recognized as government organizations. Organization names containing "University", "Institution", "College", etc., were classified as academic. And the terms "Ltd" (Limited), "Co." (Company), "Inc" (Incorporated), and so forth were used to mark industrial funding organizations. Some organizations fell beyond the scope of GA&I agencies, such as charities and private foundations. We set the funding attribute of these organizations to NULL and excluded these articles as the main focus of this paper is on the funding patterns of GA&I organizations. Considering some of the well-known funding agencies are usually represented via abbreviations in research articles, e.g., "NSFC", "NSF", "NIH", "Google", and so on, we also add these abbreviations to the prepared thesaurus. Figure 1 presents the classification regularities used in this study; some examples are given in the shadowed rectangles.

Topic modeling and parameter setting

Topics are extracted from the articles using an LDA model—one of the most accepted topic modeling techniques. LDA generates a discrete distribution of words to represent topics, and a discrete distribution of topics to explain a corpus. From the perspective of LDA, there is a definite number of entire documents D and unique terms N (Blei et al., 2003). Other observable variables include the number of terms in the dth document in the set D, N_d, and the nth word in the document d, $W_{d,n}$. There are two hyperparameters α and β, which have smoothing effects on the topic distributions and the word distributions, respectively (Heinrich, 2005). To provide a fine-grained decomposition of the document collection, we followed the common setting in the Python Genism toolkit and set the number of iterations of Gibbs sampling to 10,000.

The total number of topics in D are consolidated into K groups. One of the most important concerns when applying an LDA-based approach in scientific text mining is the assumptions made when setting the parameter K, especially when prior knowledge of the target area is limited (Chen et al., 2022). For large-scale data processing, statistical methods usually create a large number of topics (De Battisti et al., 2015). Yet ignoring the latent thematic structure of the corpus and setting K simply for ease of interpretation will reduce the reliability of the result. Addressing these concerns,

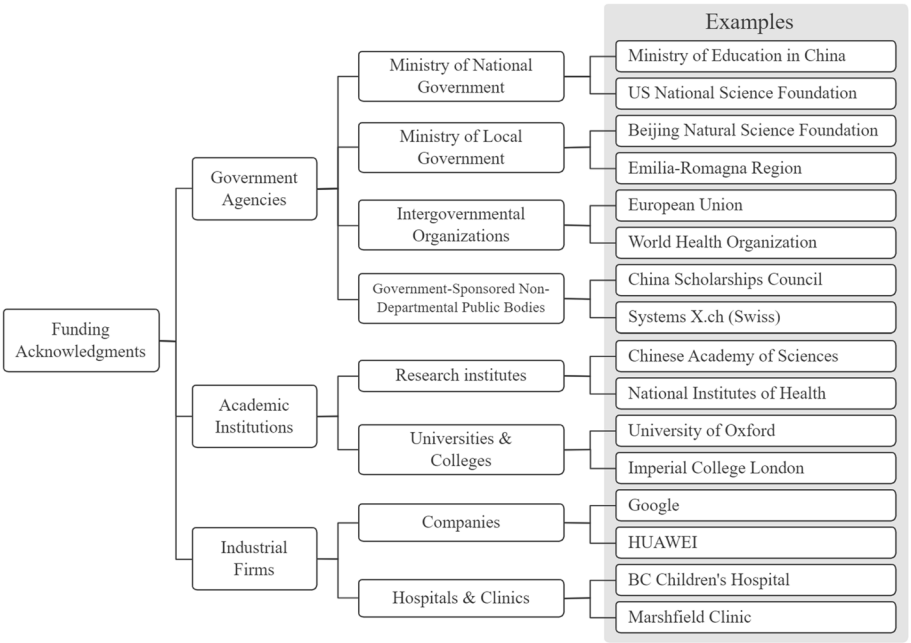

Fig. 1 Examples of GA&I organization classification

one must strike a balance between statistical evaluation and human interpretation. For this reason, the number of topics K should be determined based on a perplexity score, which is algebraically equivalent to the reciprocal geometric mean of the likelihood (Huang et al., 2018), as shown in Eq. (1). A lower perplexity score is preferred, which illustrates a lower misrepresentation of the words in the dataset. Formally, this is calculated as

$$\text{Perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p\left(w_d\right)}{\sum_{d=1}^{M} N_d}\right\}, \tag{1}$$

where N_d indicates the document length of d in D, which has a total of M documents, and $\sum \log\left(p\left(w_d\right)\right)$ represents the likelihood of the corpus given the trained model with a specific value of K. The larger the number of topics, the harder they are to interpret.

The extracted K topics are meaningful decompositions of the corpus, which reveals the knowledge structure of the target funded domain via both word distributions (topic–word matrix) and topic distributions (topic–document matrix). Moreover, based on the topic distribution matrix generated by the LDA model, topics can be further featured with bibliometric information provided by the metadata. For example, the participation of GA&I organizations in each topic per year can be calculated according to the publication time and funding organization of the corresponding articles (Chen et al., 2017). In turn, this can reveal structural changes in the funding sources of a given field.

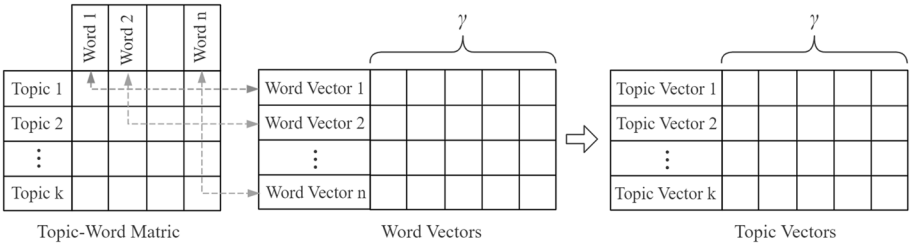

Fig. 2 The calculation process of topic vectors

Topic vectorization with word embedding

In recent years, word embedding has garnered substantial interest as a tool to map words into numeric vectors, which can be used to replace traditional word representations in scientific text mining (Zhang et al., 2018). Topic modelling can effectively reveal global relations between documents and words, while word embedding shows its capacity to capture token-level syntactic and semantic information from contexts (Moody, 2016). Our methodology calls for the extracted topics to be vectorized using word2vec, one of the most efficient word embedding techniques (Mikolov, Chen, et al., 2013). This procedure maps global and contextual semantics into fixed-length numeric vectors.

There are two specific models for word2vec: one is Skip-Gram model and the other is CBOW model (Le & Mikolov, 2014). According to the independent benchmarking conducted by Levy et al. (2015), these two models show no fundamental performance difference in practice. We selected the Skip-Gram algorithm to train word2vec model (Levy et al., 2015) and used the Genism toolkit to assist with generating the word vectors. The input to the model is a sequence of words, $P = \{w_{i-k}, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_{i+q}\}$, in which w_i stands for a target word and q is the context size of the target word. The window size is denoted as $S = 2q + 1$. Based on the Skip-Gram model proposed by Mikolov, Sutskever, et al. (2013) and the notation given by Zhang et al. (2018), the main objective is summarized as maximizing the average log probability $L(D)$, in which the probability $\Pr(w_{i+c}|w_i)$ is formulated with a softmax function:

$$L(D) = \frac{1}{\varphi}\sum_{i=1}^{\varphi}\sum_{-q\le c\le q, c\neq 0}\log\Pr(w_{i+c}|w_i), \tag{2}$$

$$\Pr(w_{i+c}|w_i) = \frac{\exp(x_{i+c}\cdot x_i)}{\sum_{w\in W}\exp(x\cdot x_i)}. \tag{3}$$

Here,φ is the size of corpus, indicating the number of unique terms, and x_{i+c} is the vector representation of a context word for the target word w_i. The dimension of fixed-length word vectors was set to γ. Then a negative sampling technique is used to train $L(D)$—we used the Genism toolkit. For each unique term, the model should return a γ-dimensional vector.

The top n words that contribute most to a topic are then used to generate a topic–word matrix. As shown in Fig. 2, the topic–word matrix and word vectors are integrated to calculate a weighted average for all contributing word vectors for all topics. As illustrated in Eq. 4, $w_k^n(n = 1, 2, \ldots)$ denotes the words appearing in topic k, with its probability represented as $p(w_k^n)$. The result gives K topic vectors with γ dimensions with which to interpret the contextual and relational semantics of the target corpus.

	Topic 1 Topic 2 … Topic k	Tagged Org 1 Tagged Org 2 … Tagged Org m
Topic 1 Topic 2 … Topic k	Correlation Matrix of Topics (Semantic Network)	Organization-Topic Matrix (2-mode Network)
Tagged Org 1 Tagged Org 2 … Tagged Org m	Organization-Topic Matrix (2-mode Network)	Co-occurrence Matrix of Tagged Organizations (Collaboration Network)

Fig. 3 The two-mode matrix for organization–topic networks construction

$$\text{vector}\left(T_k\right) = \sum\nolimits_n p\left(w_k^n\right) * \text{vector}(w_k^n). \tag{4}$$

Organization–topic networks construction and analysis

After topic vectorization, each article is associated with tagged GA&I funding organizations and a distribution of topic vectors. From here, a two-mode organization–topic network is used to reveal both the collaborative interactions formed by the funding organizations and the semantic networks constituted by coupling among the topics these organizations supported. As shown in Fig. 3, the organization–topic network is made up of three types of matrices: a topic correlation matrix, a co-occurrence matrix of tagged organizations, and an organization–topic matrix. The final organization–topic network is a heterogeneous combination of topic and organization nodes and links that explains not only the funding preferences of the organizations but also the semantic relations between topics.

The detailed process of network construction is illustrated in Fig. 4. Organization categories are set up and all the articles are allocated to their corresponding funding agencies. Then a homogeneous collaboration network is constructed based on the co-occurrence matrix of the funding organizations. This network reflects situations where two agencies have both provided financial support to research the same topic. However, this network is also based on word embedding-enhanced topic extraction and vectorization that has been applied to map topics as fixed-length numeric vectors. Calculating the cosine similarities of every two topic vectors forms a correlation matrix of topics. Meanwhile, the organization–topic matrix is generated by linking tagged organizations and funded topics with topic–document matrix generated via LDA. Finally, these three types of matrices are merged into an organization–topic network and visualized using Pajek (Nooy et al., 2011) and VOSviewer (van Eck & Waltman, 2010).

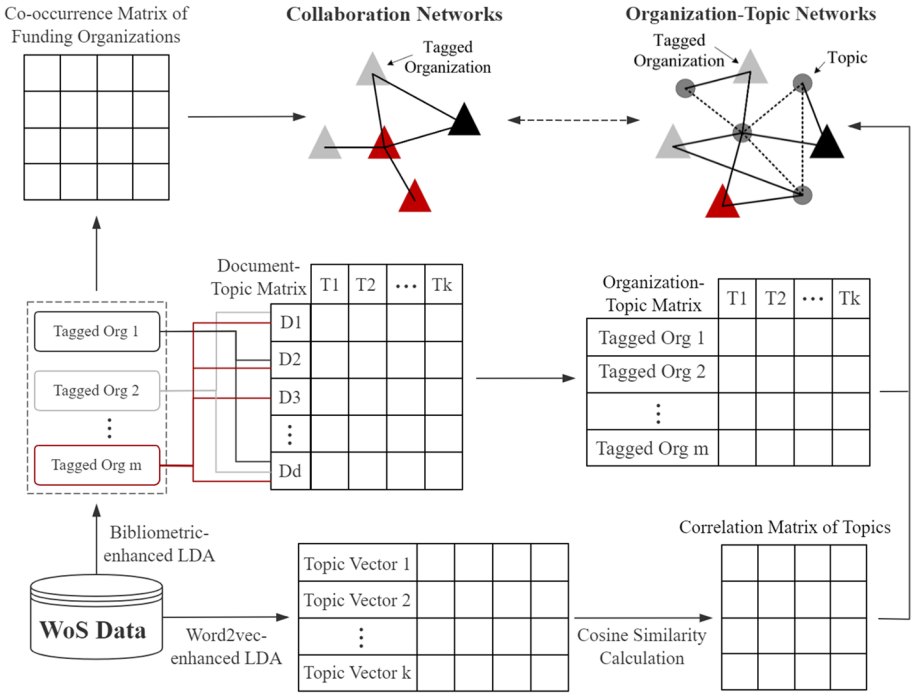

Fig. 4 The construction process of heterogeneous organization–topic networks

Empirical case study: funding patterns exploration in the field of big data

Data

Our case study concerns the field of big data. Ever since data-driven technologies and applications have begun to exert an influence on modern business, government, science, and medicine, funding organizations have attached great significance to big data research, providing considerable financial support for many undertakings in this arena. Studies on big data took the limelight during the 2010s, leading to a dramatically increasing number of scientific outputs (Zhang et al., 2019). To demonstrate the feasibility and effectiveness of our methodology, we extracted for our sample a collection of articles from the WoS core collection database, which covers the SCIE (Citation Index Expanded), SSCI (Social Science Citation Index), and A&HCI (Arts & Humanities Citation Index). Although funding acknowledgments have only been systematically collected for SSCI and A&HCI for the last 5 years or so (Liu et al., 2020), we retrieved any papers where funding information was included as per the data collection procedure outlined in "Data pre-processing" section. Note however, that the prevalence of topics funded in the fields of social science, arts, and humanities may be underestimated due to biased coverage in the indexes.

In total, 9882 funded articles were retrieved that contain *big data* in their titles, abstracts, or keywords, and were published between 2010 and 2019 inclusive. Figure 5 depicts the number of articles funded per year, as well as the proportion of funding coming from each

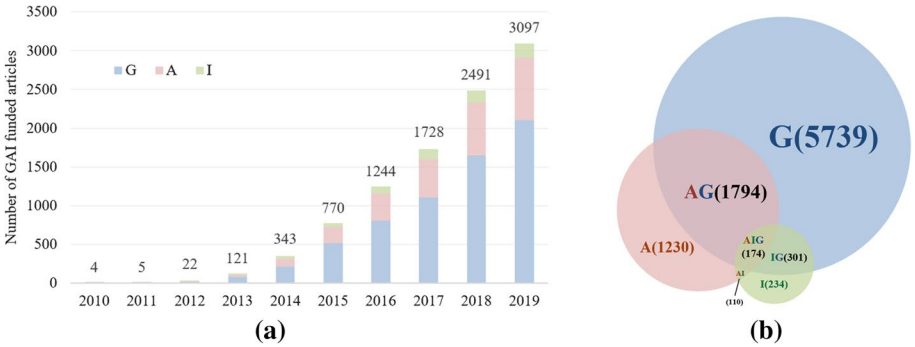

Fig. 5 **a** Number of articles funded in the big data field between 2010 and 2019, and **b** proportion of articles funded by GA&I organizations

source. As we can see, the number of funded articles has increased significantly, especially during the past 5 years. Government organizations are the prime body to provide financial support in the field of big data. 8008 Articles acknowledged government funding (81.51%), with papers funded by academic agencies accounting for 33.67% of all funded articles. Only 819 articles received industry support, among which 71.43% were also jointly funded by the other two types of organizations.

Interactive patterns of GA&I organizations in collaboration networks

'Co-fund' networks are different from collaboration networks. Collaboration networks are spontaneous; they are constructed when authors or institutions collaborate with each other consciously. However, co-fund networks are constructed when two funding organizations fund the same research project. These two funding organizations are only connected by research projects. However, what it does show is that each funding organization is paying attention to the same issues.

Figure 6 contains the top 80 most frequently appearing agencies. Each node denotes an organization, while the node size indicates the number of corresponding funded articles. The colors represent the organizational sector—government in blue, academia in red, and industry in green. The thickness of the links as well as the distances between the nodes reflects the number of articles co-funded by the two organizations.

As we can see from Fig. 6, government agencies took up a prominent position in big data research, followed by academic institutions. Only 6 industrial organizations are listed in the top 80, most of which are technology-driven enterprises. Furthermore, the network reveals that government agencies tend to collaborate with all types of organizations, while there is comparatively much less collaborative funding from academic–industry partnerships.

The main funding organizations in Fig. 6 are listed in Table 1, grouped by country. As shown, the governments of these countries have agencies that specialize in allocating funding to big data research, including the National Natural Science Foundation of China (NSFC), National Science Foundation of the USA (NSF), UK Research and Innovation (UKRI), National Research Foundation of Korea (NRF), Australian Research Council (ARC), and so forth. These government agencies served as the main role in supporting big data research and development.

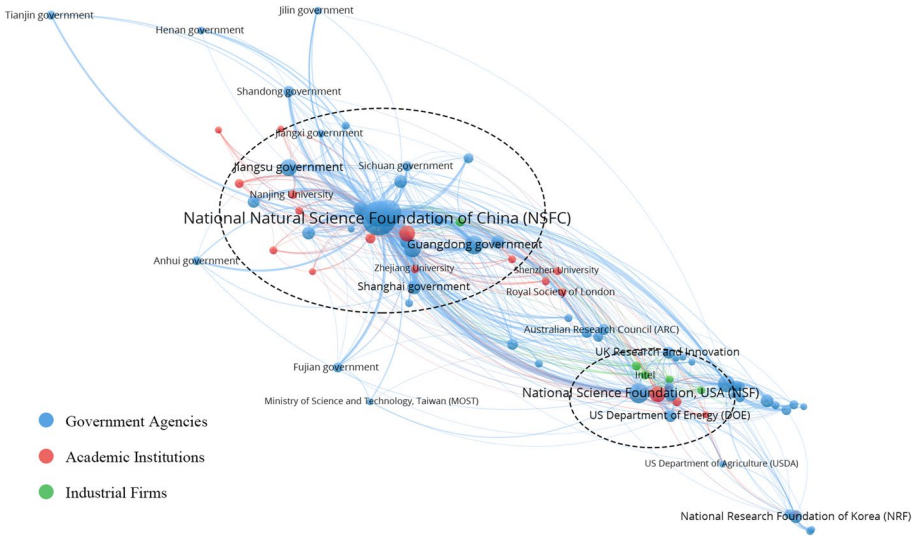

Fig. 6 The co-occurrence network of GA&I funding organizations

As shown in Fig. 6, there are two main organizational clusters, as circled in black, which have NSFC and US NSF as core elements respectively. These two clusters reflect two different collaborative modes of financial support in big data research. The first cluster with the NSFC represents a collaboration mode where national government agencies play a leading role in funding big data research, and local governments play a complementary role. Meanwhile, financial support from academia and industry serves as a supplement. The second cluster with the NSF shows more interactions between GA&I organizations. Although government agencies contribute an indispensable part of scientific funding to big data research, the academic community also provides financial support for a considerable number of papers. Particularly, industry is more involved in this type of funding collaborative mode.

Distribution of funded topics vs. the proportion of GA&I funding

The results in this section were largely drawn from the topic extraction analysis. We first explore the distribution of funded topics, and then illustrate how GA&I organizations change their investments into certain topics over the years. According to the methodology proposed in "Data pre-processing" and "Topic modeling and parameter setting" sections, the titles and abstracts were cleaned and consolidated for topic extraction and analysis. 15,971 Unique words remained after data pre-processing. We selected the value of K based on perplexity calculation and detected 30 topics from 9225 articles. As shown in Table 2, these topics imply a series of important tools, techniques, and applications in the big data research field. The rank in the first column indicates the topic's popularity.

As illustrated in Table 1, with the explosive growth of multisource, heterogeneous, and unstructured data, *Algorithms* (Topic 1) of data mining and analysis have taken center stage during the past decade, providing classical theories and new inspirations for scientists. *Machine Learning* (Topic 8) has spurred broad interest in big data research, unlocking the critical bottleneck of traditional algorithms. In addition, *Cloud*

Table 1 Main GA&I funding organizations in big data research

Country	Government agencies	Academic institutions	Industrial firms
China	National Natural Science Foundation of China Ministry of Science and Technology Ministry of Education National Social Science Foundation of China Local government (Guangdong, Jiangsu, Beijing)	Chinese Academy of Sciences Tsinghua University Nanjing University Central South University	HUAWEI
USA	National Science Foundation Department of Defense Department of Energy	National Institutes of Health National Aeronautics and Space Administration Stanford University	Microsoft Intel Google Amazon IBM
EC	European Commission		
UK	UK Research and Innovation	Royal Society of London	
Korea	National Research Foundation of Korea Ministry of Education	Kyungsung University	
Australia	Australian Research Council		
Spain	Spanish Ministry of Economy and Competitiveness		
Canada	Natural Sciences and Engineering Research Council Social Sciences and Humanities Research Council		
Germany	German Research Foundation		
Japan	Japan Science and Technology Agency Ministry of Education, Culture, Sports, Science and Technology		
Brazil	Brazilian National Council for Research and Development Coordination for the Improvement of Higher Education Personnel		
Portugal	Portuguese Foundation for Science and Technology		

Table 2 Topics extracted from articles in the big data field

Rank	Topic name	Detail content
1	Algorithms	Algorithm, problem, set, classification, performance
2	Computational Social Science	Social, science, digital, medium, public
3	Cloud Computing	Cloud, computing, service, application, cloud computing
4	Decision Management	Decision, management, making, industry, business
5	MapReduce	Processing, performance, MapReduce, memory, Hadoop
6	Real-time Processing	Time, real, real time, processing, query
7	Knowledge Mining	Knowledge, mining, text, language, technique
8	Machine Learning	Learning, machine, machine learning, deep, neural
9	Smart Cities	Traffic, smart, city, flow, sensing
10	Clustering	Clustering, social, graph, network, algorithm
11	Disease Risk	Patient, risk, clinical, predictive, treatment
12	Healthcare	Health, care, disease, patient, healthcare
13	Urban Spaces	Spatial, area, land, urban, temporal
14	Chemical Biology	Drug, brain, material, chemical, science
15	Time Series Analysis	Time, series, product, time series, online
16	Genome Sequences	Tool, genome, sequence, datasets, sequencing
17	Wireless Communication	Network, node, sensor, wireless, communication
18	Matrixes	Matrix, signal, fault, sparse, sampling
19	Visualization	Visualization, simulation, scientific, interactive, tool
20	Human Mobility Patterns	Mobile, activity, urban, human, pattern
21	Cancer and Gene Expression	Gene, cancer, cell, protein, disease
22	Energy Consumption	Power, consumption, emission, carbon, building
23	Behavior Detection	Detection, behavior, temporal, time, prediction
24	Databases	Database, distribution, statistical, inference, spectrum
25	Climate Change	Climate, global, specie, change, earth
26	Privacy and Security	Privacy, security, attack, protection, encryption
27	Group Experiments	Age, level, group, driving, composition
28	Recommendation	Recommendation, preference, tourism, particle, stock
29	Search Engines	Search, engine, image, recognition, emotion
30	Education	Imaging, student, education, teaching, magnetic

Computing (Topic 3) is a hotspot in big data research, together with topics such as *Real-time Processing* (Topic 6) and *Databases* (Topic 24), raising a disruptive revolution to deal with massive and complex data. As productions of cloud computing technology, Hadoop and *MapReduce* (Topic 5) make it convenient and flexible for data storage, processing, and management (Qadir et al., 2020). All these topics have garnered substantial attention from scientists and funding agencies.

Knowledge Mining (Topic 7) has been another focus of funding organizations. In the era of big data, it integrates various techniques and tools, rendering it possible for scientists to extract useful information, reveal hidden insights, and understand patterns and mechanisms of target research objects (Cheng et al., 2018). Some basic analytics, including *Time Series Analysis* (Topic 15) and *Visualization* (Topic 19), have been applied in scientific studies long before the era of big data. However, the boom of big data brings new targets, new opportunities, and new challenges in the emergence

of deep learning, greatly improving the quality and efficiency of traditional knowledge mining.

Funding organizations also highlight the role of big data in dealing with real-world problems. Topics of *Computational Social Science* (Topic 2) and bioinformatics, including *Genome Sequence* (Topic 16), *Cancer and Gene Expression* (Topic 21) have attracted enormous attention. Moreover, big data also contributes to *Decision Management* (Topic 4) for both enterprises and government (Acharya et al., 2018). It promotes the development of *Urban Spaces* (Topic 13) and the construction of *Smart Cities* (Topic 9) (Okeke & Ukonze, 2019). It encourages environmentally-friendly *Energy Consumption* (Topic 22), patterns to mitigate *Climate Change* (Topic 25) (Zhang et al., 2020), and helps to improve *Education* (Topic 30) and *Healthcare* (Topic 12) standards.

After profiling funded topics, we further explore the structural changes of their funding sources from a dynamic perspective. Here, topics that have garnered different attention will show unequal funding levels (Stahlman & Heidorn, 2020). As shown in Fig. 7, we selected six topics with rising, falling, or stable popularity as examples to illustrate how GA&I organizations have changed their investments into particular topics. By topic investment, we mean the preference, or say focus, of different funding organizations in supporting research priorities. It reflects the number of funding grants that a topic has received, but not the amount of money awarded from the funding, which is not information made available in the article acknowledgements. In this figure, lines indicate the change of topics' popularity of the selected topics, where the x-axis represents time, which is set as 2013 to 2019 in this diagram, owing to the low output of big data publications before 2013 (only 5, 6, and 41 articles published in 2010–2012 respectively). The y-axis represents the proportion of topics per article per year, which is acquired from the result of LDA, indicating the popularity of these topics from a funding perspective (Chen et al., 2021). As for pie charts, they demonstrate the participation of GA&I organizations in each topic, which are quantified according to the number of funds that topics receive.[2] In this figure, GA&I organizations are highlighted in orange, yellow, and green respectively.

As shown in Fig. 7, topics with rising popularity are *decision management* and *machine learning*. Government plays a dominating role in supporting these topics and has paid more attention to this area in recent years than before. Especially for the latter one, the proportion of government funds is around 78% in 2019, up from 58% in 2015. Similar trends can also be observed from *computational social science* and *cloud computing*. Nearly 82% and 74% of funds are acquired from government in 2019. Despite that the popularity of these two topics has witnessed a progressive decrease, they still occupy a considerable share among all funded themes. In general, government organizations prefer to support mainstream topics and have promoted a rapid growth of several topics. As we can observe in Fig. 7, the popularity of *healthcare* and *disease risk* has not changed much over the years. Academia and industry pay more attention to these two topics with nearly half of the articles related to these topics being funded by academic institutions and industrial firms. As a matter of fact, topics that attract academic organizations most always show stable popularity. For most topics, although the number of articles funded by industry is increasing over years, the proportion is decreasing. This is mainly owing to the fact of growth rate of industry funding is lower than that of government and academia.

[2] Complete and detailed funding proportion for all topics are provided in "Appendix".

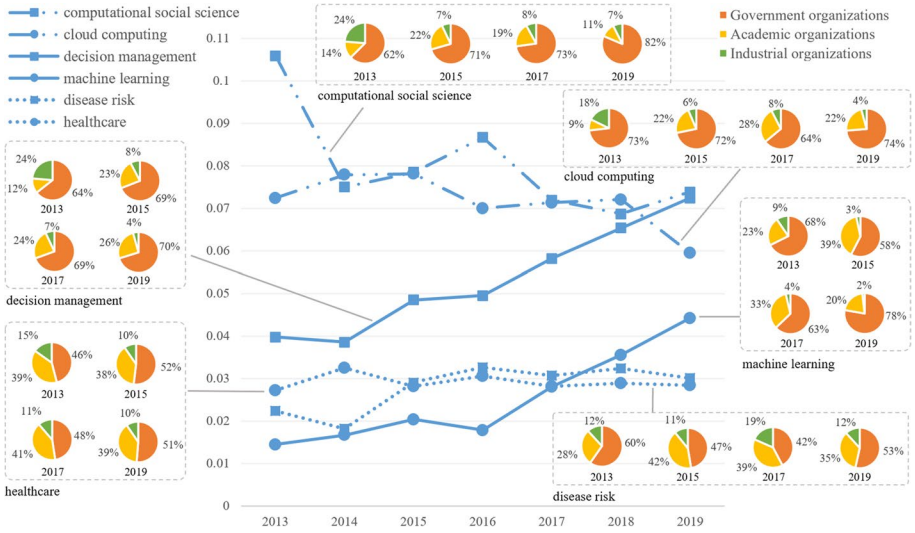

Fig. 7 GA&I organizations' "topics investments" over years

Funding patterns analysis in organization–topic networks

Following our methodology of investigating interactions between funded topics and organizations, we pushed our research further by exploring funding patterns via heterogeneous organization–topic networks. The two-mode network for big data field highlights the pivotal role of GA&I organizations in a knowledge network formed by topic vectors. It also reveals the semantic relations between topics, implying their potential for combination and integration. The network is shown in Fig. 8, where nodes with different colors denote organizations and topics respectively, while the weight of edges represents funding intensity. For each topic, it is more likely to get funding from organizations with a closer distance, and neighboring topics enjoy more semantic connections to some extent.

As we can observe in Fig. 8, GA&I organizations all emphasize financial support for research topics *including algorithms, decision management, MapReduce, computational social science, real-time processing, and cloud computing*. Interesting patterns were revealed in that most of these topics have strong semantic relations with each other. There are some topics, such as *privacy and security*, that barely receive financial support from funding communities. However, privacy and security is one of the most important issues with the big data boom (Liu et al., 2017). This warrants further discussion, research and funding support.

According to the heterogeneous organization–topic network, funding preferences of GA&I organizations share similarities but also present differences. As we can see, government agencies are willing to support the *smart cities* topic. Since urbanization has become a general trend, government policymakers have fully appreciated the necessity to promote smart city construction and substantial development. Figure 8 shows that the *smart cities* topic is semantically linked to topics such as *energy consumption* and *wireless communication*. Some novel insights might be acquired if knowledge of these topics can be integrated. Therefore, these topics could be considered simultaneously when funding strategies are updated. Another topic that government funding organizations

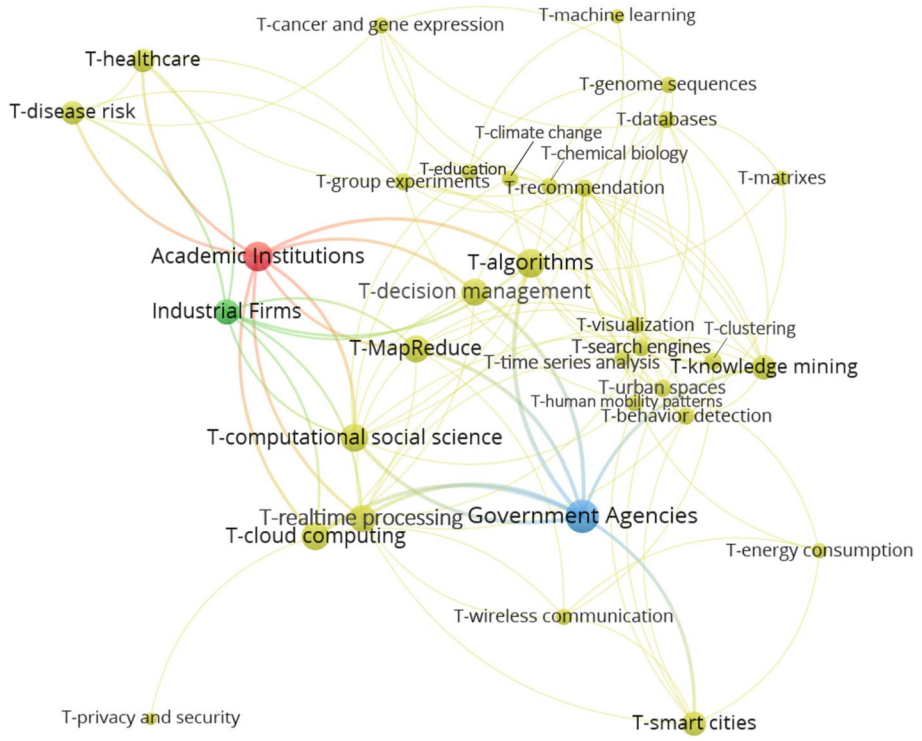

Fig. 8 Funding preference of GA&I organizations in the organization–topic network

are relatively concern about is *knowledge mining*, which has provided new insights for information extraction, knowledge discovery, and trend prediction. As shown in Fig. 8, knowledge mining is connected to several topics. Some of them are methods or tools, including *visualization*, *clustering*, and *databases*. Others may involve applications of knowledge mining, for example *recommendation*, *behavior detection*, *chemical biology*, and so on.

Academic institutions and industrial firms show similar funding preferences. In addition to the focal topics in the field of big data, they also focus on topics of *healthcare* and *disease risk*. These topics are semantically linked to *cancer and gene expression*, *group experiments*, and so forth. The theories and approaches of these topics could be integrated, which might be helpful for knowledge discovery, technological innovation, and achievement transformation. We notice that some academic communities are devoted to problems in the life sciences and biomedicine field, aiming to promote science and benefit humanity. Additionally, pharmaceutical companies and medical facilities are an important part of industry. These companies are paying increasing efforts on supporting scientific research and gaining more profit through innovation.

Furthermore, we push our analysis to explore individual-level organizations' positions in the heterogeneous network. For each type of GA&I, we select five organizations with the highest funding frequency and constructed the heterogeneous organization–topic network. As shown in Fig. 9, topics are colored in yellow, while GA&I organizations are colored in blue, red, and green, respectively.

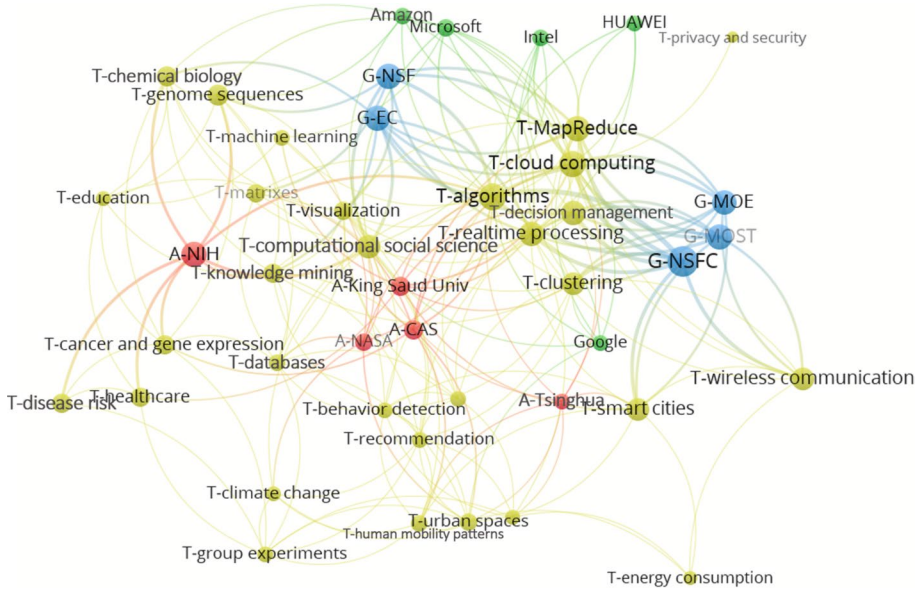

Fig. 9 Funding preference of the major organizations in the organization–topic network. Abbreviations in this figure are explained as follows: *NSF* US National Science Foundation, *EC* European Commission, *MOE* Ministry of Education, China, *MOST* Ministry of Science and Technology, China, *NSFC* National Science Foundation of China, *NIH* National Institutes of Health, *CAS* Chinese Academy of Sciences, *NASA* National Aeronautics and Space Administration

As shown in Fig. 9, the NSFC, MOE (Ministry of Education, China), and MOST (Ministry of Science and Technology, China), which are subordinate to the Chinese government, show similar topic investment preferences, while the NSF (US National Science Foundation) and the EC (European Commission) always have the same funding focus. These five government organizations are all willing to provide financial support for *algorithms*, *real-time processing*, *MapReduce*, and *cloud computing* research, most of which are important techniques in the field of big data. Yet these agencies also have preferences towards some other topics. The NSFC, MOE, and MOST attach greater importance to accelerating the efficiency of *decision management*, promoting the progress of *wireless communication* technology, and emphasize the construction of the *smart cities*. While NSF and EC particularly focus on the development of *computational social science*. And it can be noticed that this topic is connected to *knowledge mining* and *visualization*.

As for academic agencies, some organizations pay particular attention to specific topics owing to their goals and mission. As a global leader in space exploration, NASA (National Aeronautics and Space Administration) mainly funds research related to *urban space* and *climate change*, which show relations with *time series analysis* and *human mobility patterns*. The NIH (National Institute of Health), one of the world's foremost medical research centers, provides most of its funds for topics such as *healthcare*, *disease risk*, *cancer and gene expression*, *chemical biology*, and *genome sequence*. Knowledge of these topics is integrated, making it efficient for NIH to deal with health and medical issues. Besides, since the CAS (Chinese Academy of Sciences), the King

Saud University, and the Tsinghua University are comprehensive universities or research institutes, they tend to fund topics with more varieties, but not be confined to a specific field.

Within the scope of industrial organizations, Amazon, Microsoft, Google, Intel, and HUAWEI have provided considerable support for big data research, especially highlighting the development of popular techniques and algorithms. They are also interested in supporting *decision management* in big data environment, which plays an increasingly important role in modern business operations. As mentioned in the first half of this section, some topics show relatively strong semantic relations with others. Such relationships can be considered when organizations develop their funding strategies.

Conclusions and future work

Funding efficiency and funding directions have always been research foci that rely on comprehensive analyses of funding patterns. In recent years, understanding the funding patterns of government, academic, industry organizations has attracted the attention of researchers and policymakers alike. For example, the Chinese government has begun to emphasize the superiority of industry funding (Li et al., 2017), while the US government is taking steps to increase its investments into critical technology fields. It is also pushing for a coordinated response across academia, industry, and government (National Academies of Sciences, Engineering and Medicine, 2020).

From an actor–content perspective, funding activities illustrate the preference of different organizations in encouraging research priorities. The topics funded, on the other hand, reveal deeper knowledge of dynamics and interactions of these research focuses. To this end, we devised a methodology based on a word embedding-enhanced organization–topic network to reveal the pivotal role of different funding agencies in supporting research topics. This framework can be used to explore the interactive patterns of GA&I organizations, the topical structures of funded research, and the funding focus of GA&I organizations. The proposed methodology is effective in summarizing funding characteristics and revealing the role that key organizations are playing in supporting innovation. These findings can assist these agencies to upgrade their funding strategies and benchmark with other key players. For academics, this research is useful since an increasing number of researchers are actively seeking access to financial support. The insights yielded can be used for topic analysis and trend evaluation and may point academics to where the opportunities for funding are.

Although this paper provides heuristic research on exploring funding patterns with word embedding-enhanced organization–topic networks, it has several limitations that need to be explored in future research. (1) The data pre-processing module for the funding organization field is semi-NLP-based and time-consuming, which mainly because inconsistent funding acknowledgments leads to low data quality. Funding organizations may change their names over time, and a considerable number of agencies are better known by their abbreviations, at least to their primary audiences. Some are not written in English or have a native name and an English name. Facing this challenge, a well-prepared list of funding organizations would be very helpful for the NLP model and could save a great deal of manual checking time. (2) Our empirical analysis is limited to the field of big data, which is an emerging, data-driven areas. Therefore, it is hard to verify the patterns we discovered in our empirical study. However, the research design in this paper could be replicated

and funding patterns in other contexts could be further studied to corroborate our findings. (3) In this paper, we did not dig deeply into the dynamic mechanisms of how interactions in funding agency collaborations drive knowledge upgrades and exchanges in the semantic network. Further studies that systematically measure the interactions of topic networks and funding organization collaboration networks as a complex system will be addressed in our future research. Furthermore, the observations in this article should be regarded as an exploration. Econometric models could be undertaken to support, refute, or contextualize our findings.

Appendix

See Table 3.

Table 3 Breakdown of GA&I funding topics have received over the years

Topic	Participation of GA&I Algorithms	Participation of GA&I Computational social science	Participation of GA&I Cloud computing	Participation of GA&I Decision management	Participation of GA&I MapReduce
2010	1.00, 0.00, 0.00	0.97, 0.03, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00
2011	1.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 1.00, 0.00	1.00, 0.00, 0.00
2012	0.54, 0.31, 0.14	1.00, 0.00, 0.00	0.65, 0.33, 0.02	0.53, 0.46, 0.01	0.61, 0.24, 0.15
2013	0.74, 0.20, 0.05	0.62, 0.14, 0.24	0.73, 0.09, 0.18	0.64, 0.12, 0.24	0.71, 0.15, 0.13
2014	0.64, 0.26, 0.10	0.73, 0.17, 0.10	0.70, 0.25, 0.05	0.67, 0.26, 0.07	0.66, 0.23, 0.11
2015	0.73, 0.22, 0.05	0.71, 0.22, 0.07	0.72, 0.22, 0.06	0.69, 0.23, 0.08	0.67, 0.25, 0.07
2016	0.67, 0.26, 0.07	0.79, 0.14, 0.07	0.70, 0.25, 0.06	0.70, 0.24, 0.06	0.69, 0.20, 0.11
2017	0.68, 0.25, 0.06	0.73, 0.19, 0.08	0.64, 0.28, 0.08	0.69, 0.24, 0.07	0.74, 0.18, 0.08
2018	0.70, 0.25, 0.05	0.75, 0.18, 0.07	0.69, 0.25, 0.06	0.69, 0.25, 0.06	0.71, 0.22, 0.07
2019	0.71, 0.23, 0.06	0.81, 0.11, 0.07	0.74, 0.22, 0.04	0.70, 0.25, 0.04	0.72, 0.21, 0.07
Topic	Real-time processing	Knowledge mining	Machine learning	Smart cities	Clustering
2010	1.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 1.00, 0.00	0.00, 0.00, 0.00
2011	0.72, 0.28, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00
2012	0.83, 0.15, 0.02	0.64, 0.30, 0.07	0.23, 0.77, 0.00	0.80, 0.20, 0.00	0.67, 0.05, 0.28
2013	0.67, 0.21, 0.12	0.67, 0.32, 0.01	0.68, 0.23, 0.10	0.71, 0.11, 0.18	0.71, 0.14, 0.16
2014	0.65, 0.23, 0.12	0.66, 0.24, 0.10	0.49, 0.50, 0.01	0.82, 0.14, 0.04	0.81, 0.13, 0.06
2015	0.69, 0.21, 0.09	0.68, 0.28, 0.04	0.58, 0.39, 0.03	0.74, 0.19, 0.06	0.78, 0.15, 0.07
2016	0.65, 0.29, 0.06	0.65, 0.28, 0.07	0.55, 0.42, 0.03	0.74, 0.19, 0.07	0.78, 0.17, 0.05
2017	0.67, 0.26, 0.08	0.63, 0.31, 0.06	0.63, 0.34, 0.04	0.69, 0.24, 0.07	0.71, 0.26, 0.04
2018	0.68, 0.25, 0.07	0.66, 0.28, 0.06	0.73, 0.24, 0.03	0.69, 0.26, 0.05	0.66, 0.30, 0.04
2019	0.69, 0.24, 0.07	0.66, 0.28, 0.06	0.78, 0.20, 0.02	0.74, 0.21, 0.05	0.61, 0.36, 0.03
Topic	Disease risk	Healthcare	Urban spaces	Chemical biology	Time series analysis
2010	0.00, 0.00, 0.00	0.04, 0.96, 0.00	0.00, 0.00, 0.00	0.63, 0.37, 0.00	1.00, 0.00, 0.00
2011	0.00, 1.00, 0.00	0.00, 1.00, 0.00	0.69, 0.31, 0.00	0.00, 1.00, 0.00	1.00, 0.00, 0.00

Table 3 (continued)

Topic	Disease risk	Healthcare	Urban spaces	Chemical biology	Time series analysis
2012	1.00, 0.00, 0.00	0.52, 0.45, 0.03	0.80, 0.01, 0.19	0.85, 0.15, 0.00	0.63, 0.24, 0.13
2013	0.60, 0.28, 0.12	0.46, 0.39, 0.15	0.68, 0.13, 0.19	0.59, 0.33, 0.09	0.80, 0.14, 0.05
2014	0.55, 0.35, 0.11	0.41, 0.39, 0.19	0.77, 0.16, 0.07	0.53, 0.38, 0.09	0.58, 0.31, 0.11
2015	0.47, 0.42, 0.11	0.52, 0.39, 0.10	0.67, 0.23, 0.10	0.59, 0.34, 0.07	0.67, 0.26, 0.07
2016	0.44, 0.41, 0.15	0.47, 0.40, 0.13	0.71, 0.21, 0.08	0.55, 0.39, 0.07	0.67, 0.29, 0.05
2017	0.42, 0.39, 0.18	0.48, 0.41, 0.11	0.73, 0.21, 0.06	0.56, 0.33, 0.11	0.64, 0.28, 0.07
2018	0.50, 0.36, 0.15	0.51, 0.38, 0.11	0.72, 0.23, 0.04	0.62, 0.30, 0.08	0.67, 0.29, 0.04
2019	0.53, 0.35, 0.12	0.51, 0.39, 0.10	0.76, 0.19, 0.04	0.61, 0.31, 0.08	0.68, 0.28, 0.03
Topic	Genome sequences	Wireless communication	Matrixes	Visualization	Human mobility patterns
2010	0.51, 0.49, 0.00	0.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00
2011	1.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00
2012	0.48, 0.09, 0.43	1.00, 0.00, 0.00	0.62, 0.27, 0.10	0.53, 0.24, 0.23	0.60, 0.11, 0.29
2013	0.64, 0.30, 0.06	0.76, 0.21, 0.03	0.87, 0.11, 0.02	0.67, 0.21, 0.12	0.78, 0.05, 0.17
2014	0.51, 0.39, 0.10	0.68, 0.22, 0.10	0.72, 0.23, 0.05	0.60, 0.32, 0.08	0.64, 0.20, 0.16
2015	0.53, 0.39, 0.08	0.70, 0.22, 0.09	0.68, 0.25, 0.07	0.72, 0.24, 0.05	0.74, 0.20, 0.06
2016	0.53, 0.39, 0.08	0.73, 0.21, 0.06	0.69, 0.22, 0.08	0.64, 0.29, 0.06	0.61, 0.27, 0.11
2017	0.55, 0.34, 0.11	0.68, 0.25, 0.07	0.73, 0.22, 0.05	0.67, 0.26, 0.06	0.68, 0.28, 0.05
2018	0.62, 0.32, 0.06	0.68, 0.26, 0.05	0.72, 0.22, 0.06	0.67, 0.26, 0.06	0.66, 0.29, 0.05
2019	0.61, 0.31, 0.09	0.70, 0.25, 0.05	0.73, 0.21, 0.06	0.67, 0.27, 0.07	0.68, 0.28, 0.04
Topic	Cancer and gene expression	Energy consumption	Behavior detection	Databases	Climate change
2010	0.62, 0.38, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00	0.00, 0.00, 0.00	0.92, 0.08, 0.00
2011	1.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00
2012	0.10, 0.90, 0.00	0.85, 0.15, 0.00	0.63, 0.28, 0.09	0.88, 0.00, 0.12	0.50, 0.29, 0.22
2013	0.36, 0.63, 0.02	0.72, 0.04, 0.24	0.66, 0.28, 0.06	0.71, 0.23, 0.06	0.71, 0.17, 0.12

Table 3 (continued)

Topic	Cancer and gene expression	Energy consumption	Behavior detection	Databases	Climate change
2014	0.34, 0.62, 0.04	0.61, 0.30, 0.08	0.66, 0.29, 0.05	0.64, 0.32, 0.04	0.60, 0.33, 0.07
2015	0.35, 0.59, 0.06	0.63, 0.27, 0.10	0.75, 0.21, 0.04	0.74, 0.22, 0.04	0.60, 0.33, 0.07
2016	0.31, 0.61, 0.08	0.73, 0.19, 0.08	0.64, 0.29, 0.07	0.68, 0.25, 0.07	0.66, 0.29, 0.05
2017	0.37, 0.53, 0.09	0.69, 0.26, 0.06	0.61, 0.29, 0.10	0.67, 0.26, 0.07	0.64, 0.30, 0.06
2018	0.40, 0.52, 0.07	0.74, 0.21, 0.04	0.65, 0.29, 0.06	0.66, 0.28, 0.06	0.63, 0.28, 0.09
2019	0.37, 0.58, 0.05	0.71, 0.23, 0.06	0.71, 0.25, 0.04	0.69, 0.26, 0.05	0.68, 0.27, 0.05
Topic	Privacy and security	Group experiments	Recommendation	Search engines	Education
2010	0.00, 0.00, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00
2011	0.00, 1.00, 0.00	1.00, 0.00, 0.00	0.00, 0.00, 0.00	1.00, 0.00, 0.00	1.00, 0.00, 0.00
2012	0.59, 0.26, 0.14	0.21, 0.00, 0.79	0.58, 0.35, 0.07	0.53, 0.12, 0.35	0.65, 0.30, 0.04
2013	0.51, 0.15, 0.34	0.49, 0.50, 0.01	0.54, 0.05, 0.41	0.75, 0.21, 0.04	0.62, 0.18, 0.20
2014	0.67, 0.23, 0.10	0.59, 0.31, 0.11	0.68, 0.29, 0.02	0.63, 0.30, 0.08	0.46, 0.47, 0.07
2015	0.72, 0.24, 0.04	0.50, 0.46, 0.05	0.65, 0.27, 0.08	0.66, 0.28, 0.06	0.68, 0.25, 0.07
2016	0.61, 0.32, 0.07	0.57, 0.32, 0.11	0.70, 0.26, 0.05	0.68, 0.27, 0.05	0.61, 0.31, 0.08
2017	0.65, 0.26, 0.09	0.43, 0.46, 0.10	0.62, 0.35, 0.03	0.64, 0.29, 0.07	0.59, 0.32, 0.08
2018	0.64, 0.27, 0.09	0.53, 0.40, 0.07	0.71, 0.26, 0.03	0.66, 0.29, 0.05	0.65, 0.29, 0.07
2019	0.69, 0.20, 0.11	0.53, 0.38, 0.09	0.71, 0.26, 0.03	0.67, 0.28, 0.05	0.61, 0.32, 0.08

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant Nos. 72004009, 61872033, and 71804016, and also supported by Beijing Institute of Technology Research Fund Program for Young Scholars Beijing and Nova Program (Z201100006820015) from the Beijing Municipal Science and Technology Commission.

References

Aagaard, K., Mongeon, P., Ramos-Vielba, I., & Thomas, D. A. (2021). Getting to the bottom of research funding: Acknowledging the complexity of funding dynamics. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0251488

Acharya, A., Singh, S. K., Pereira, V., & Singh, P. (2018). Big data, knowledge co-creation and decision making in fashion industry. *International Journal of Information Management, 42*, 90–101.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Brennecke, J., & Rank, O. (2017). The firm's knowledge network and the transfer of advice among corporate inventors—A multilevel network study. *Research Policy, 46*, 768–783.

Chang, S.-H. (2017). The technology networks and development trends of university–industry collaborative patents. *Technological Forecasting and Social Change, 118*, 107–113.

Chen, Y. L., Dong, Y. T., Zeng, Y., Yang, X. Y., Shen, J. T., Zheng, L., Jiang, J. W., Pu, L. M., & Bao, Q. L. (2020). Mapping of diseases from clinical medicine research—A visualization study. *Scientometrics, 125*, 171–185.

Chen, H., Jin, Q., Wang, X., & Xiong, F. (2022). Profiling academic–industrial collaborations in bibliometric-enhanced topic networks: A case study on digitalization research. *Technological Forecasting and Social Change, 175*, 121402.

Chen, H., Wang, X., Pan, S., & Xiong, F. (2021). Identify topic relations in scientific literature using topic modeling. *IEEE Transactions on Engineering Management, 68*, 1232–1244.

Chen, H. S., Zhang, G. Q., Zhu, D. H., & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change, 119*, 39–52.

Cheng, Y., Chen, K., Sun, H. M., Zhang, Y. P., & Tao, F. (2018). Data and knowledge mining with big data towards smart production. *Journal of Industrial Information Integration, 9*, 1–13.

Colatat, P. (2015). An organizational perspective to funding science: Collaborator novelty at DARPA. *Research Policy, 44*, 874–887.

De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics, 103*, 413–433.

Gao, J. P., Su, C., Wang, H. Y., Zhai, L. H., & Pan, Y. T. (2019). Research fund evaluation based on academic publication output analysis: The case of Chinese research fund evaluation. *Scientometrics, 119*, 959–972.

Greiner-Petter, A., Youssef, A., Ruas, T., Miller, B. R., Schubotz, M., Aizawa, A., & Gipp, B. (2020). Math-word embedding in math search and semantic extraction. *Scientometrics, 125*, 3017–3046.

Grimpe, C. (2012). Extramural research grants and scientists' funding strategies: Beggars cannot be choosers? *Research Policy, 41*, 1448–1460.

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics, 11*, 407–422.

Guerzoni, M., Taylor Aldridge, T., Audretsch, D. B., & Desai, S. (2014). A new industry creation and originality: Insight from the funding sources of university patents. *Research Policy, 43*, 1697–1706.

Heinrich, G. (2005). *Parameter estimation for text analysis*. Technical Report.

Hellsten, I., & Leydesdorff, L. (2020). Automated analysis of actor–topic networks on Twitter: New approaches to the analysis of socio-semantic networks. *Journal of the Association for Information Science and Technology, 71*, 3–15.

Hu, K., Luo, Q., Qi, K. L., Yang, S. L., Mao, J., Fu, X. K., Zheng, J., Wu, H. Y., Guo, Y., & Zhu, Q. B. (2019). Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Information Processing and Management, 56*, 1185–1203.

Hu, Y.-H., Tai, C.-T., Liu, K. E., & Cai, C.-F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics, 14*, 101004.

Huang, M. H., & Huang, M. J. (2018). An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries. *Scientometrics, 115*, 833–847.
Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science, 64*, 2833–2855.
Huang, Y., Zhang, Y., Youtie, J., Porter, A. L., & Wang, X. (2016). How does national scientific funding support emerging interdisciplinary research: A comparison study of big data research in the US and China. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0154509
Lamba, M., & Madhusudhan, M. (2019). Mapping of topics in DESIDOC Journal of Library and Information Technology, India: A study. *Scientometrics, 120*, 477–505.
Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning (ICML-14)*, 2014 (pp. 1188–1196).
Lee, Y.-Y., Ke, H., Yen, T.-Y., Huang, H.-H., & Chen, H.-H. (2020). Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. *Journal of the Association for Information Science and Technology, 71*, 657–670.
Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211–225.
Leydesdorff, L. (2003). The mutual information of university–industry–government relations: An indicator of the Triple Helix dynamics. *Scientometrics, 58*, 445–467.
Li, J., Xie, Y., Wu, D., & Chen, Y. (2017). Underestimating or overestimating the distribution inequality of research funding? The influence of funding sources and subdivision. *Scientometrics, 112*, 55–74.
Liu, W. (2020). Accuracy of funding information in Scopus: A comparative case study. *Scientometrics, 124*, 803–811.
Liu, Q., Srinivasan, A., Hu, J. K., & Wang, G. J. (2017). Preface: Security and privacy in big data clouds. *Future Generation Computer Systems: The International Journal of Escience, 72*, 206–207.
Liu, W., Tang, L., & Hu, G. (2020). Funding information in Web of Science: An updated overview. *Scientometrics, 122*, 1509–1524.
Ma, T. C., Li, R. N., Ou, G. Y., & Yue, M. L. (2018). Topic based research competitiveness evaluation. *Scientometrics, 117*, 789–803.
Mejia, C., & Kajikawa, Y. (2018). Using acknowledgement data to characterize funding organizations by the types of research sponsored: The case of robotics research. *Scientometrics, 114*, 883–904.
Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of workshop at ICLR*, 2013.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Moody, C. E. (2016). Mixing Dirichlet topic models and word embeddings to make lda2vec. arXiv: Computation and Language.
Munari, F., & Toschi, L. (2021). The impact of public funding on science valorisation: An analysis of the ERC Proof-of-Concept Programme. *Research Policy, 50*, 104211.
National Academies of Sciences, Engineering and Medicine. (2020). *The endless frontier: The next 75 years in science.* The National Academies Press.
Naumanen, M., Uusitalo, T., Huttunen-Saarivirta, E., & Van der Have, R. (2019). Development strategies for heavy duty electric battery vehicles: Comparison between China, EU, Japan and USA. *Resources, Conservation and Recycling, 151*, 104413.
Nooy, W. D., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek*. Cambridge University Press.
Okeke, D. C., & Ukonze, I. (2019). Conceptualizing urban space (environment) for the delivery of sustainable urban development in Africa: Evidence from Enugu City in Nigeria. *Land Use Policy, 87*, 104074.
Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks. *Journal of Management, 38*, 1115–1166.
Qadir, J., Sainz-de-Abajo, B., Khan, A., Garcia-Zapirain, B., de la Torre-Diez, I., & Mahmood, H. (2020). Towards mobile edge computing: Taxonomy, challenges, applications and future realms. *IEEE Access, 8*, 189129–189162.
Song, B., & Suh, Y. (2019). Identifying convergence fields and technologies for industrial safety: LDA-based network analysis. *Technological Forecasting and Social Change, 138*, 115–126.
Stahlman, G. R., & Heidorn, P. B. (2020). Mapping the "long tail" of research funding: A topic analysis of NSF grant proposals in the division of astronomical sciences. *Proceedings of the Association for Information Science and Technology, 57*, e276.
Tang, L., Hu, G., & Liu, W. (2017). Funding acknowledgment analysis: Queries and caveats. *Journal of the Association for Information Science and Technology, 68*, 790–794.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*, 523–538.

Wang, J., Lee, Y. N., & Walsh, J. P. (2018). Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy, 47*, 1070–1083.

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: A study on 10 countries. *Scientometrics, 91*, 591–599.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports, 714*, 1–73.

Zhang, Y., Huang, Y., Porterc, A. L., Zhang, G. Q., & Lu, J. (2019). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change, 146*, 795–807.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H. S., & Zhang, G. Q. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*, 1099–1117.

Zhang, X. B., Sun, J. Y., Fei, Y. N., & Wei, C. (2020). Cooler rooms on a hotter planet? Household coping strategies, climate change, and air conditioning usage in rural China. *Energy Research and Social Science, 68*, 101605.

Zhao, R. Y., Li, X. L., Liang, Z. S., & Li, D. Y. (2019). Development strategy and collaboration preference in S&T of enterprises based on funded papers: A case study of Google. *Scientometrics, 121*, 323–347.

Zhao, S. X., Lou, W., Tan, A. M., & Yu, S. (2018). Do funded papers attract more usage? *Scientometrics, 115*, 153–168.

Zhou, P., & Tian, H. (2014). Funded collaboration research in mathematics in China. *Scientometrics, 99*, 695–715.