Topic Identification and Evolution Analysis of Open Source Software

Authors: Dong Pingjun, Gao Xiangfei

Date: 2023-07-28T00:00:00+00:00

Abstract

[目的/意义] Software open source is an important production organization method and collaborative innovation movement in socialized software production. Through topic identification and evolution analysis of domestic and international software open source related research, this study explores the patterns of periodic hotspots and trend changes in the software open source research field, providing direction for scholars aiming to further optimize and promote the development of China's software open source innovation. [方法/过程] Using literature in the software open source field retrieved from the Web of Science database from 2001 to May 10, 2023 as the corpus, the perplexity metric was adopted to determine the number of topics, an LDA topic identification model was trained to obtain topic-word distributions and document-topic distributions, topics were labeled based on topic-word distributions, topic intensity was calculated based on document-topic distributions, and thereby hotspot topics were identified and evolution paths were summarized. [结果/结论] The topic identification results indicate that there are six important themes in the software open source research field, namely contribution motivation, business model, open source governance, collaboration model, open source license, and enterprise participation; from the perspective of topic evolution, software open source has relatively high research intensity in business model, open source governance, and enterprise participation themes in recent years, the research trend of open source license is relatively stable, and although the research intensity of contribution motivation and collaboration model shows a relatively declining trend, they have consistently maintained high attention throughout. Software open source research exhibits a development pattern from focusing on the individual dimension of spontaneous and autonomous open source motivation to the organizational dimension of enterprise and government participation. It is recommended that scholars pay attention to research on various themes of the open source ecosystem under the Chinese context, providing theoretical support for the healthy development of China's

open source ecosystem.

Full Text

Topic Identification and Evolutionary Analysis of Open Source Software Research

Dong Pingjun, Gao Xiangfei
Rising Sun School of Business Administration, Donghua University, Shanghai 200051

Abstract

[**Purpose/Significance**] Open source software represents a crucial production organization model and collaborative innovation movement in socialized software production. By identifying and analyzing the thematic evolution of open source software research both domestically and internationally, this study explores the phased hotspots and changing trends within the field, providing directional guidance for scholars aiming to promote the optimized development of China' s open source software innovation ecosystem.

[**Method/Process**] Using literature on open source software retrieved from the Web of Science database spanning 2001 to May 10, 2023 as the corpus, this study employs perplexity metrics to determine the optimal number of topics. An LDA topic identification model was trained to obtain topic-word and document-topic distributions. Topics were labeled based on the topic-word distribution, and topic intensity was calculated from the document-topic distribution to identify hotspots and summarize evolutionary trajectories.

[**Result/Conclusion**] The topic identification reveals six major themes in open source software research: contribution motivation, business models, open source governance, collaboration patterns, open source licensing, and enterprise participation. From an evolutionary perspective, business models, open source governance, and enterprise participation have garnered relatively high research interest in recent years. Research on open source licensing has remained relatively stable, while interest in contribution motivation and collaboration patterns, though declining relatively, has consistently maintained high attention. Open source software research has evolved from an initial focus on individual-level motivations toward organizational dimensions involving enterprise and government participation. Scholars are encouraged to investigate various themes within the Chinese open source ecosystem context to provide theoretical support for its healthy development.

Keywords: open source software; topic identification; topic evolution; LDA model
Classification Number: G353.1

In recent years, open source software has emerged as a significant production model and collaborative innovation movement in socialized software production, forming a rich open source ecosystem with increasingly profound impact. Under this model, various elements—including software users, developers, organizers, platforms, licenses, and researchers—mutually nourish and collaborate in continuous iteration: users freely download and trial software while providing feedback; individual or organizational developers can view, innovate, and redistribute source code under open source licensing terms; and open source organizations maintain ecosystem development through protocols and charters. Open source software embodies the sharing economy within software production, complementing proprietary software models to constitute the broader software production ecosystem.

Over the past decades, the open source movement has continuously driven software production development and innovation trends. From early internet-era innovations like the Linux operating system and MySQL relational database management system, to mobile-era technologies such as the Android OS, Apache web server, Vue frontend framework, Eclipse and VSCode development tools, and git project management systems, to contemporary breakthroughs like the Python language and ChatGPT—all have benefited from open source innovation. The movement has established platform communities typified by GitHub, extending beyond software into hardware, documentation, music, and broader domains. This global fusion of professional knowledge, skills, and production elements enables rapid iteration, unleashing tremendous innovation potential and driving technological and social progress.

Recognizing open source' s positive role in software innovation and addressing China' s relatively lagging software development, China has actively built its open source ecosystem since 2009. By 2022, China ranked first globally in developer growth. According to CSDN, China' s leading developer community, registered users exceeded 35 million, with over 94% using open source software and more than 40% participating in open source projects. Gitee reported over 1.8 million new registered users in 2021, with cumulative open source developers surpassing 8 million. GitHub' s 2021 statistics showed China' s developer count reaching 7.55 million, ranking second globally. In 2022, Chinese open source demonstrated strong contribution enthusiasm in major international foundations: Chinese board members exceeded 40% in the Open Source Infrastructure Foundation; Chinese projects accounted for over 20% in the Cloud Native Computing Foundation; and the Apache Software Foundation hosted 24 active Chinese projects, including 14 top-level projects. Notably, in 2021, all five new projects entering the Apache incubator originated from China. Furthermore, increasing numbers of Chinese enterprises recognize open source' s importance, actively participating in projects. Tech giants like Huawei, Alibaba, and Baidu open-source their software and technologies while investing in ecosystem development. Government policies have also promoted open source development: the Ministry of Industry and Information Technology' s *Big Data Industry Development Plan (2016-2020)* explicitly encouraged participation in open source

projects to enhance China' s influence; open source was first included in China' s *14th Five-Year Plan* in March 2021; the *14th Five-Year Plan for Software and IT Services Industry* systematically outlined open source ecosystem development; and the *14th Five-Year Plan for Digital Economy Development* supported autonomous open source communities, platforms, and projects.

In summary, China' s open source innovation and operations are accelerating, with the ecosystem developing robustly. However, research on open source software remains relatively limited as a component of this ecosystem. This study conducts thematic identification and evolutionary analysis of international open source software research to understand ecosystem development patterns, providing insights and guidance for Chinese scholars to advance domestic open source sharing research and contribute Chinese wisdom to global open source development.

2. Related Research on Topic Identification and Evolution

Topics represent specific information content or concerns within texts, conceptualized as collections of related terms that help readers quickly grasp core content and provide better text summarization. Topic identification is a text mining technique that discovers hidden themes from text collections by analyzing semantic and contextual information to infer topics and their probabilities. Topic evolution refers to the changing development process of topics across different time periods within a certain timeframe. Topic identification and evolution analysis employs bibliometrics or natural language processing to identify and track domain topics, dynamically analyzing and visualizing development trends. Research typically involves evolutionary paths, keyword co-occurrence network formation and evolution, changes in core authors and institutions, and factors influencing topic evolution. Applications span public opinion monitoring, social media analysis, healthcare, business intelligence, and government decision-making, helping researchers comprehensively understand research progress and trends across fields and time periods.

Table 1 compares common methods for topic identification and evolution analysis. Bibliometric methods typically identify topics through word frequency, co-occurrence relationships, and citation analysis—simple and widely used approaches. Social network analysis is particularly favored for revealing inter-topic relationships. Xing Xiaozhao et al. proposed a disruptive technology identification method based on topic evolution using text mining and social network analysis, while Reza Vahidzadeh employed social network analysis to explore technical and non-technical themes in regional industrial symbiosis research. However, bibliometric methods, while capable of mining large-scale literature, often neglect semantic information, resulting in less rich outcomes. Consequently, increasing scholars employ machine learning-based topic models for mining and evolution analysis. Algorithms like Naive Bayes and Support Vector Machines enable classification and clustering for topic identification and tracking. Among these, probabilistic graphical models such as Latent Semantic

Analysis (LSA), Latent Dirichlet Allocation (LDA), and Dynamic Topic Models (DTM) transform texts into topic-word distributions, implicitly modeling relationships between documents, topics, and words. By learning probability distribution parameters, these models discover topic structures and evolution patterns, mining latent semantic relationships more effectively than traditional methods.

Table 1 Comparison of Common Methods for Topic Identification and Evolution Analysis

Method Type	Advantages	Disadvantages
Bibliometric Methods (Word frequency analysis; Co-occurrence analysis; Social network analysis; Factor analysis)	1. Simple and easy to use2. Reveals citation relationships and influence; obtains cooperation among authors and countries	1. Prone to high-frequency keyword extraction2. Ignores internal semantic information, focusing only on relationships and citations
Machine Learning Methods (Dynamic topic models; Latent semantic analysis; Latent Dirichlet allocation; Dynamic topic models)	1. Uses probabilistic graphical models to provide word probability distributions for each topic, revealing rich semantic content2. Handles large data volumes quickly and accurately; significantly reduces subjectivity in topic extraction	1. Requires large training datasets for model construction2. High computational complexity with large-scale texts

Existing literature on open source software exhibits several limitations. First, many studies focus on single projects or application domains, analyzing specific issues rather than the entire research landscape. Second, traditional bibliometric methods dominate literature reviews, with few employing unsupervised topic modeling from an LDA perspective, lacking in-depth evolutionary exploration. Additionally, some studies analyze only source code repositories to identify developer contribution themes for task matching, without leveraging other data sources. This study addresses these gaps by employing LDA topic modeling to deeply investigate open source software themes and evolution, exploring development trends and future directions. This research provides references for open source software development, promotes research in open source sharing, and offers insights for topic analysis in other domains.

3. Research Methodology

3.1 Data Collection and Acquisition

This study selected the Web of Science (WOS) database as the data source, retrieving literature on May 11, 2023, covering 2001 to the retrieval date. The document type was limited to "Article" to exclude conferences and newspapers. The search query TS=("open source software" OR "free software" OR "libre software") yielded 1,146 initial papers. After reviewing titles and abstracts to exclude technically-focused papers unrelated to this study' s research orientation, 769 bibliographic records were retained, including title, keywords, extended keywords, abstract, journal source, publication year, author, institution, and country. The annual distribution is shown in Figure 1 [Figure 1: see original paper].

Figure 2 [Figure 2: see original paper] displays the top fifteen source journals by publication count. *Research Policy*, *Journal of Systems and Software*, and *Information Systems Research* contribute relatively large numbers, accounting for 30.7% of publications among the top fifteen journals.

3.2 Data Preprocessing

Before topic identification and evolution analysis, literature data requires preprocessing to enhance quality and accuracy. This study extracted titles, keywords, extended keywords, and abstracts from WOS bibliographic records as the corpus for topic modeling. The NLTK natural language processing toolkit performed tokenization with the following steps: (1) Excluding special characters such as numbers and punctuation; (2) Restoring abbreviations to full forms to merge concept variants (e.g., "OSS" to "Open Source Software," "PR" to "Pull Request"); (3) Converting words to lowercase, performing part-of-speech tagging and filtering to remove non-meaningful words like adjectives and adverbs; (4) Constructing stopword lists, synonym lists, and domain-specific dictionaries for open source software research to remove stopwords and merge synonyms, continuously expanding these lists during model training to optimize tokenization; (5) Performing lemmatization and stemming.

3.3 Research Method

This study employs the Latent Dirichlet Allocation (LDA) model—a probabilistic graphical model-based topic modeling technique—to analyze thematic characteristics of open source software, revealing key topics and research hotspots. Blei et al. proposed LDA in 2003, identifying latent topics through joint modeling of word-topic and topic-word distributions in texts. LDA can model large volumes of open source software texts, identifying key topics and interrelated terms to uncover core research issues. Its advantage lies in automatic topic discovery without predefined topics or manual annotation, learning relationships between topics and documents through statistical analysis. LDA is particularly useful

for large-scale text processing and can discover topic evolution across different time periods, though it requires pre-specifying topic numbers.

LDA is a three-level Bayesian model (Figure 3 [Figure 3: see original paper]) representing each document as a probability distribution over topics, with each topic as a probability distribution over words. LDA assumes each word in an open source research article is generated by first selecting a topic with certain probability, then selecting a word from that topic. Document-topic and topic-word distributions depend on Dirichlet priors determined by parameters α and β.

As shown in Figure 4 [Figure 4: see original paper], LDA is now widely used for discovering latent topics in document collections. This study treats a paper' s title, keywords, extended keywords, and abstract as one document. LDA assumes that among M (769) open source documents, each document comprises N words drawn from K topics; each document follows a multinomial distribution over topics, and each topic follows a multinomial distribution over words. The prior for document-topic distributions is a Dirichlet distribution with parameter α, while the prior for topic-word distributions is a Dirichlet with parameter β.

Based on these assumptions, LDA' s generative process for each document is: (1) Sample a multinomial topic distribution for document m from a Dirichlet distribution with parameter α; (2) Generate topic z_{mn} for the nth word in document m according to this multinomial distribution; (3) Sample a multinomial word distribution for topic k from a Dirichlet distribution with parameter β; (4) Integrate word distributions across topics. Repeating this process generates documents of N words, ultimately producing M (769) documents across K topics. Gibbs sampling estimates parameters, training document-topic distributions and corresponding word distributions ϕ.

3.4 LDA Parameter Determination

Gensim, a Python package for text analysis, topic modeling, and word embeddings, was used to train the LDA model. Before training, three hyperparameters must be determined: α, β, and k. Parameter α represents the Dirichlet prior for document-topic distributions, controlling sparsity and topic diversity per document. Smaller α values result in fewer topics per document. Parameter β defines the Dirichlet prior for topic-word distributions, determining word diversity within topics. Smaller β values concentrate topics on a few high-frequency words, while larger β values distribute words more evenly. As topic results are not highly sensitive to α and β, Gensim' s default values are typically used—fixed symmetric priors of 1/k, where k represents the optimal number of topics.

Four common methods determine the optimal topic number k: (1) **Perplexity**: Measures model prediction accuracy on new data; lower perplexity indicates better performance. (2) **Coherence Score**: Evaluates topic quality through consistency metrics; higher scores indicate better models. (3) **Text Clustering**: Derives topic numbers from clustering results, though this is susceptible to sam-

ple characteristics and algorithmic influences, requiring manual intervention. (4) **Visualization and Manual Assessment**: Uses topic-word distribution visualizations, domain knowledge, and research objectives to evaluate topic quality, inter-topic relationships, and interpretability.

Perplexity, measuring uncertainty in document-topic assignment, is the most popular method for determining topic numbers. This study adopts perplexity, selecting the k value where perplexity is minimized or at an inflection point, indicating strong generalization. The perplexity formula is:

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

where M is the total number of documents in the open source domain, p(w_d) is the generation probability of document d, and N_d is the word count in document d. The generation probability p(w_d) is calculated as the product of each word' s generation probability:

$$p(w_d) = \prod_{n=1}^{N_d} p(w_{dn})$$

Restricting topic numbers to 2-20 and calculating perplexity for each yields the trend shown in Figure 5 [Figure 5: see original paper]. Perplexity is minimized at 6 topics. Coherence score validation (Figure 6 [Figure 6: see original paper]) shows the first inflection at 3 topics, but this number is too small for adequate generalization. A second inflection occurs at 6 topics with relatively high coherence, confirming 6 as the optimal topic number for open source software research.

3.5 LDA Topic Modeling

With topic number set to 6, LDA training yields two results: topic-word distribution and document-topic distribution. The topic-word distribution reveals which words belong to each topic and their probabilities, enabling analysis of inter-topic relationships, identification of specific topic content, and revelation of evolutionary paths. The document-topic distribution shows which topics each document contains and their weights, primarily used for calculating topic intensity, identifying phased hotspots, and determining evolution paths.

This study employs LDAvis for interactive visualization based on Sievert and Shirley' s (2014) web-based topic visualization method, facilitating better understanding and analysis of topics, words, and weight distributions from an integrated perspective. Figure 7 [Figure 7: see original paper] shows six bubbles of varying sizes representing the six topics. Bubble size corresponds to document count, indicating relative importance. Inter-topic distances reflect

similarity—closer topics share more similar word distributions. Minimal overlap indicates good classification performance.

4. Results and Analysis

4.1 Topic Identification and Analysis

Based on LDA' s topic-word distribution, topics were labeled by examining the top ten highest-probability words per topic (Table 2).

Table 2 LDA Topic Modeling Results

Topic 1 (Contribution Motivation)	Topic 2 (Business Models)	Topic 3 (Open Source Governance)
project	development	project
motivation	project	community
community	developer	developer
developer	contribution	development
contribution	study	innovation
study	development	research
development	innovation	governance
innovation	research	model
research	model	ecosystem
model	study	system

Topic 4 (Collaboration Patterns)	Topic 5 (Open Source Licensing)	Topic 6 (Enterprise Participation)
project	license	community
developer	study	project
development	innovation	developer
network	process	study
community	model	design
knowledge	ecosystem	model
model	research	development
study	development	research
impact	adoption	analysis
community	study	model

Based on high-probability words, topics were labeled as follows:

Topic 1: Contribution Motivation
High-frequency words like project, motivation, developer, and contribution align with research on developer participation motivations. This is among the earliest research topics, with many scholars investigating Lerner and Tirole' s question:

"Why do thousands of top developers contribute for free to create public goods?" Alexander Hars and Shaosong Ou systematically studied motivations through email surveys, revealing intrinsic and extrinsic factors. Learning also drives participation. Von Hippel and von Krogh proposed a "private-collective" innovation model from an organizational science perspective. Shaul Oreg and Oded Nov categorized motivations by instrumental levels. Von Krogh et al. (2012) constructed a motivation-practice framework incorporating long-term pursuits beyond short-term rewards. Moqri et al. confirmed both non-monetary and future monetary reward incentives.

Topic 4: Collaboration Patterns
High-frequency words like developer, team, network, and impact relate to developer collaboration patterns. Raymond' s *The Cathedral and the Bazaar* systematically explained open source collaboration, contrasting Linux' s "bazaar" model with traditional "cathedral" models. Open source developers organize and contribute more freely, creating evolving social networks. Strong prior collaboration increases project joining likelihood. Collaboration patterns critically determine project success and growth.

Topic 5: Open Source Licensing
High-frequency words like license, innovation, and success reveal research on license selection and project success. License choice, interacting with organizational sponsorship, affects contributor interest and development activity. Project success also relates to sustained developer participation, community culture, and ideology. License compliance management is crucial—non-compliance damages organizational reputation. When derivative development requires substantial effort, less restrictive licenses better facilitate success. Popular licenses include MIT, GNU GPL, Apache, BSD, and Mozilla Public License, each with distinct characteristics for different scenarios.

Topic 2: Business Models
High-frequency words like business, model, and company align with open source business model research. Scholars have examined factors influencing model selection, design approaches, and systematic classifications. Common models include: (1) Open core—open-sourcing core code while charging for plugins or runtime materials; (2) Support and consulting—providing technical support, training, and customized solutions; (3) Delayed open source—new versions remain proprietary while older versions are open-sourced; (4) Dual licensing—offering both open source and commercial licenses; (5) Donations and sponsorship—securing funding from enterprises, organizations, or individuals. Red Hat, MongoDB, Elastic, and Docker exemplify successful open source commercialization.

Topic 3: Open Source Governance
High-frequency words like governance, model, network, and ecosystem fit open source community governance research. Virtual communities require different governance mechanisms than traditional organizations. Vishal Midha proposed a two-dimensional governance classification (participation and responsibility management) demonstrating impacts on maintenance outcomes. Saerom Lee

explored governance strategies for allocating developers across multiple projects. Early governance research focused on communities, but increasing corporate participation has shifted attention to enterprise open source governance, including trade-offs between open and proprietary software and time allocation for open source participation.

Topic 6: Enterprise Participation
High-frequency words like community, company, and developer reflect research on corporate employee participation in open source projects. Enterprise participation provides access to technological developments, knowledge sharing, innovation promotion, cost reduction, and brand reputation enhancement. Companies contribute through technical support, training services, community cultivation, code contributions, and event sponsorship.

4.2 Hot Topic Identification

Topic intensity reflects a topic' s relative importance or prominence within the document collection over time. Hot topics appear frequently, exhibiting relatively high intensity values during specific periods. After LDA modeling, the document-topic matrix contains each document' s probability values for each topic. Summing and averaging these probabilities for all documents in a given year yields annual topic intensity (Formula 3):

$$\text{Topic Intensity}(k,t) = \frac{\sum_{d=1}^{M_t} p(k|d)}{M_t}$$

where M_t is the document count in year t (or total documents for overall intensity), and p(k|d) is the probability of topic k in document d. The hot topic threshold is defined as the average intensity across all topics; topics exceeding this threshold are considered hot topics (Figure 8 [Figure 8: see original paper]).

Contribution motivation, business models, and collaboration patterns emerge as hot topics. Contributor motivations have long interested researchers, with participants including individuals, enterprises, foundations, and governments —individual motivations receiving the most attention, followed by enterprises. Collaboration pattern research remains hot due to global distribution and complex technical/knowledge/communication challenges. Business model research is also active, increasingly focusing on commercialization, community governance, and corporate participation. Licensing research receives relatively less attention.

While academic interest in open source software grows globally, Figure 9 [Figure 9: see original paper] reveals a significant gap between China and international research. Strengthening theoretical research is essential to support China's open source development.

4.3 Topic Evolution Analysis

Topic intensity changes reveal trends and importance in open source evolution. Calculating annual topic intensity using Formula 3 and applying three-period moving averages to reduce volatility, we classify evolution patterns as rising, declining, or stable based on prior research (Figures 9-11 [Figure 11: see original paper]).

Rising Trends (Figure 10 [Figure 10: see original paper]): Business models, open source governance, and enterprise participation show fluctuating upward trends. As open source software permeates industries, research interest grows in creating commercial value and sustainable business models. Community governance research also gains attention, as successful projects require complete management processes integrating open source, collaborative workflows, and quality management. Large communities typically have five roles: (1) Leaders who guide project development with decision-making authority; (2) Maintainers who handle daily operations and management; (3) Committers who submit code and handle project affairs; (4) Contributors who participate through various means (PRs, issues, documentation, community engagement); and (5) Users who provide feedback and requirements. While China has many open source communities, most governance and operations remain rudimentary, requiring further research. Additionally, corporate participation research increases as companies recognize benefits including technological growth, reputation enhancement, and talent recruitment.

Declining Trends (Figure 11): Contribution motivation and collaboration patterns show relative decline. Early research focused on these foundational mechanisms, but as communities matured, attention shifted to commercialization, governance, and security. Moreover, expanding community scale and increasing interaction complexity require more comprehensive and in-depth research approaches beyond simple motivation and collaboration studies.

Stable Trends (Figure 12 [Figure 12: see original paper]): Open source licensing research remains relatively stable. As a fundamental issue, license selection and project success factors have maintained consistent attention. Developers must comply with licenses even when obtaining free source code. The OSI has certified over 80 open source licenses, legally regulating usage, modification, copying, and distribution. License selection involves balancing rights and obligations, significantly impacting legal compliance and commercial sustainability. Research on success factors helps contributors and organizations improve quality, attract contributors, and increase user participation.

5. Conclusions and Recommendations

5.1 Conclusions

Open source software production has profoundly impacted the software innovation landscape, accelerating knowledge creation. Open source research covers

six major themes. Overall intensity analysis shows contribution motivation, business models, and collaboration patterns exceed the intensity threshold, constituting hot topics. Evolutionary trends align with ecosystem development, progressing from individual-level motivations toward organizational dimensions involving enterprise participation and government governance. Research on business models, governance, and enterprise participation shows rising trends, while licensing research remains stable. Contribution motivation and collaboration pattern research maintain stable absolute values but declining relative proportions.

5.2 Recommendations

The evolutionary patterns of open source research themes reflect changing focal points in the open source movement. Although China is a major software producer, gaps remain in foundational and professional software domains. Building a socially collaborative open source innovation ecosystem is essential. A healthy ecosystem requires joint efforts from all participants. Scholars are encouraged to conduct China-context research to serve domestic software industry development.

At the individual level, research should examine Chinese contributors' motivations and design effective incentive mechanisms and cultural atmospheres. At the enterprise level, studies should investigate motivations and mechanisms for corporate participation to guide strategic decisions. At the platform and organizational level, research should design governance frameworks suitable for China' s context. Additionally, studying open source-based entrepreneurship cases and models will encourage open source startups, providing global communities with Chinese practices and wisdom.

References

[1] Zhou Tao, Wang Chao. Research on knowledge contribution behavior of users in open source software communities[J]. Science Research Management, 2020, 41(02): 202-209.
[2] National People' s Congress of China. *The 14th Five-Year Plan for National Economic and Social Development and Long-Range Objectives Through 2035*[EB/OL]. http://www.npc.gov.cn/, 2021-03-13.
[3] Ministry of Industry and Information Technology of China. *14th Five-Year Plan for Software and Information Technology Services Industry*[EB/OL]. http://www.miit.gov.cn/, 2021-11-30.
[4] Central People' s Government of China. *14th Five-Year Plan for Digital Economy Development*[EB/OL]. https://www.gov.cn/, 2022-01-12.
[5] Lu Guoqiang, Huang Wei, Sun Yue, et al. Research on ontology evolution intensity of network public opinion in major emergencies based on separation of opinion objects and ontology[J]. Library and Information Service, 2023, 67(05): 119-129. DOI:10.13266/j.issn.0252-3116.2023.05.011.

[6] Ma Xiaoyue, Xue Pengzhen, Chen Yijin, et al. Construction and trend analysis of social media crisis theme evolution model[J]. Library and Information Service, 2021, 65(13): 77-86. DOI:10.13266/j.issn.0252-3116.2021.13.008.
[7] Huangfu L, Mo Y, Zhang P, et al. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling[J]. Journal of medical Internet research, 2022, 24(2): e31726.
[8] Qian Y, Liu Y, Sheng Q Z. Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence[J]. Journal of Informetrics, 2020, 14(3): 101047.
[9] Zhang Ling, Yun Chengtao, Yin Sili, et al. Comparative analysis of theme evolution between China' s research integrity policies and literature[J]. Journal of Modern Information, 2023, 43(06): 108-120.
[10] Li Xiuxia, Cheng Jiejing, Han Xia. Priority ranking of disciplinary research themes based on fusion of publication and citation trends—taking Chinese information science as an example[J]. Library and Information Service, 2019, 63(11): 88-95.
[11] Tang Guoyuan. Construction of a research method for disciplinary theme evolution based on co-word analysis[J]. Library and Information Service, 2017, 61(23): 100-107.
[12] Ding Shengchun, Liu Xiaoying, Li Zhen. Research on evolution of network public opinion hot themes integrating comment influence[J]. Modern Information, 2021, 41(08): 87-97.
[13] Huang C, Yang C, Wang S, et al. Evolution of topics in education research: A systematic review using bibliometric analysis[J]. Educational Review, 2020, 72(3): 281-297.
[14] Xing Xiaozhao, Ren Liang, Lei Xiaoping, et al. Research on disruptive technology identification based on patent topic evolution—taking brain-inspired intelligence as an example[J]. Information Science, 2023, 41(03): 81-88.
[15] Vahidzadeh R, Bertanza G, Sbaffoni S, et al. Regional industrial symbiosis: A review based on social network analysis[J]. Journal of Cleaner Production, 2021, 280: 124054.
[16] Zeng Ziming, Chen Siyu. Evolution analysis of network public opinion on public health emergencies based on LDA and Bert-BiLSTM-Attention model[J/OL]. Information Studies: Theory & Application: 1-13[2023-05-22].
[17] Zhou Jian, Zhang Jie, Qu Ran, et al. Topic mining and evolution analysis of blockchain at home and abroad based on LDA[J]. Journal of Intelligence, 2021, 40(09): 161-169.
[18] Liu J, Nie H, Li S, et al. Tracing the pace of COVID-19 research: topic modeling and evolution[J]. Big Data Research, 2021, 25: 100236.
[19] Zhang Liu, Wang Hui, Xiang Mengmeng. Topic heat and evolution analysis of emergency management based on LDA[J/OL]. Information Science: 1-20[2023-05-22].
[20] Chen Guangpei, Wei Jiang, Li Tuoyu. Open source community: research context, knowledge framework, and prospects[J]. Foreign Economics & Management, 2021, 43(02): 84-102.
[21] Wang Z, Perry D E, Xu X. Characterizing individualized coding contri-

butions of OSS developers from topic perspective[J]. International Journal of Software Engineering and Knowledge Engineering, 2017, 27(01): 91-124.
[22] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
[23] Zhou, H., Yu, H. & Hu, R. Topic evolution based on the probabilistic topic model: a review. Front. Comput. Sci. 11, 786–802 (2017).
[24] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 178-185.
[25] Xia Mengmeng, Ru Xuwei, Zhang Hongjun. Research theme evolution analysis of industrial innovation ecosystem based on LDA model[J]. China University Science & Technology, 2022, No.409(09): 41-46.
[26] Chen Qi, Zhang Jundong, Zheng Wanting, et al. Topic evolution analysis of artificial intelligence in traditional Chinese medicine based on LDA model[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2022, 24(09): 3315-3324.
[27] He Liang, Li Fang. Scientific literature topic discovery and trend analysis based on topic models[J]. Journal of Chinese Information Processing, 2012, 26(02): 109-115.
[28] Ran Congjing, Li Wang. Construction of enterprise competitor identification model based on LDA—taking NIO as an example[J/OL]. Information Studies: Theory & Application: 1-11[2023-05-25].
[29] Tan Chunhui, Xiong Mengyuan. Comparative analysis of hot topic evolution in data mining research at home and abroad based on LDA model[J]. Information Science, 2021, 39(4): 174-185.
[30] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics[C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014: 63-70.
[31] Lerner J, Tirole J. Some simple economics of open source[J]. The journal of industrial economics, 2002, 50(2): 197-234.
[32] Alexander Hars S O. Working for free? Motivations for participating in open-source projects[J]. International journal of electronic commerce, 2002, 6(3): 25-39.
[33] Ye Y, Kishida K. Toward an understanding of the motivation of open source software developers[C]//25th International Conference on Software Engineering, 2003. Proceedings. IEEE, 2003: 419-429.
[34] Eric von Hippel, Georg von Krogh. Open source software and the "private-collective" innovation model: Issues for organization science[J]. Organization Science, 2003, 14(2): 209-223.
[35] Oreg, S and Nov, O. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values[J]. Computers in Human Behavior, 2008, 24(5): 2055-2073.
[36] Krogh G V, Haefliger S, Spaeth S, et al. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development[J]. MIS Quarterly, 2012, 36(2): 649-676.

[37] Moqri, M.; Mei, X.; Qiu, L.; and Bandyopadhyay, S. Effect of "following" on contributions to open source communities[J]. Journal of Management Information Systems, 2018, 35(4): 1188–1217.
[38] Raymond E. The cathedral and the bazaar[J]. Knowledge, Technology & Policy, 1999, 12(3): 23-49.
[39] Hong Q, Kim S, Cheung S C, et al. Understanding a developer social network and its evolution[C]//2011 27th IEEE international conference on software maintenance (ICSM). IEEE, 2011: 323-332.
[40] Hahn J, Moon J Y, Zhang C. Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties[J]. Information Systems Research, 2008, 19(3): 369-391.
[41] Stewart K J, Ammeter A P, Maruping L M. Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects[J]. Information Systems Research, 2006, 17(2): 126-144.
[42] Fang Y, Neufeld D. Understanding sustained participation in open source software projects[J]. Journal of Management Information Systems, 2009, 25(4): 9-50.
[43] Gamalielsson J, Lundell B. Sustainability of Open Source software communities beyond a fork: How and why has the LibreOffice project evolved?[J]. Journal of Systems and Software, 2014, 89: 128-145.
[44] Sha Z, Petrov A, Tian Y, et al. Analyzing the Robustness of Open Source Software Ecosystems to the Loss of Contributors: A Case Study[J]. Available at SSRN 4082801.
[45] SL Daniel L M, Maruping L, Cataldo M, et al. The impact of ideology misfit on open source software communities and companies[J]. Management information systems quarterly, 2018, 42(4).
[46] Maruping L M, Daniel S L, Cataldo M. Developer centrality and the impact of value congruence and incongruence on commitment and code contribution activity in open source software communities[J]. MIS Quarterly, 2019, 43(3): 951-976.
[47] Tang T Y, Fang E E, Qualls W J. More Is Not Necessarily Better: An Absorptive Capacity Perspective on Network Effects in Open Source Software Development Communities[J]. MIS Quarterly, 2020, 44(4).
[48] Gangadharan G R, D' Andrea V, De Paoli S, et al. Managing license compliance in free and open source software development[J]. Information Systems Frontiers, 2012, 14: 143-154.
[49] Sen R, Subramaniam C, Nelson M L. Open source software licenses: Strong-copyleft, non-copyleft, or somewhere in between?[J]. Decision support systems, 2011, 52(1): 199-206.
[50] Perr J, Appleyard M M, Patrick P. Open for business: emerging business models in open source software[J]. International Journal of Technology Management, 2010, 52(3/4): 432-456.
[51] Belenzon S, Schankerman M. Motivation and sorting of human capital in open innovation[J]. Strategic Management Journal, 2015, 36(6): 795-820.
[52] Shahrivar S, Elahi S, Hassanzadeh A, et al. A business model for commer-

cial open source software: A systematic literature review[J]. Information and Software Technology, 2018, 103: 202-214.
[53] Ferraz I N, Santos C D. Transformation of free and open source software development projects: governance between the cathedral and bazaar[J]. Revista de Administração de Empresas, 2022, 62.
[54] Von Krogh G, Von Hippel E. The promise of research on open source software[J]. Management science, 2006, 52(7): 975-983.
[55] Midha V, Bhattacherjee A. Governance practices and software maintenance: A study of open source projects[J]. Decision Support Systems, 2012, 54(1): 23-32.
[56] Lee S, Baek H, Jahng J. Governance strategies for open collaboration: Focusing on resource allocation in open source software development organizations. International Journal of Information Management. 2017. pp. 431–437.
[57] Schaarschmidt M, Walsh G, von Kortzfleisch H F O. How do firms influence open source software communities? A framework and empirical analysis of different governance modes[J]. Information and Organization, 2015, 25(2): 99-114.
[58] Harutyunyan N, Riehle D. Getting started with corporate open source governance: A case study evaluation of industry best practices[J]. 2021.
[59] Wang Y, Chen Y, Koo B. Open to your rival: Competition between open source and proprietary software under indirect network effects[J]. Journal of Management Information Systems, 2020, 37(4): 1128-1154.
[60] Mehra A, Mookerjee V. Human capital development for programmers using open source software[J]. MIS quarterly, 2012: 107-122.
[61] Rolandsson B, Bergquist M, Ljungberg J. Open source in the firm: Opening up professional practices of software development[J]. Research Policy, 2011, 40(4): 576-587.
[62] Dahlander L, Magnusson M. How do firms make use of open source communities?[J]. Long range planning, 2008, 41(6): 629-649.
[63] Lin Lili, Ma Xiufeng. Research theme discovery and evolution analysis of domestic library and information science based on LDA model[J]. Information Science, 2019, 37(12): 87-92. DOI:10.13833/j.issn.1007-7634.2019.12.013.
[64] Fan Xiaoqing. Research on motivations of participants in China' s open source movement[J]. Education Media Research, 2019(01): 18-25.
[66] Eghbal N. Working in public: the making and maintenance of open source software[M]. San Francisco: Stripe Press, 2020.

Author Contributions:

Dong Pingjun: Conceptualization, supervision, revision, and final approval of the manuscript.

Gao Xiangfei: Literature investigation, data collection, data processing, and manuscript drafting.

Note: Figure translations are in progress. See original paper for figures.