Scientific Knowledge Combination in Networks: New Perspectives on Analyzing Knowledge Absorption and Integration

Hongshu Chen[1],* Jingkang Liu[1] and Zikai Liu[2]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, China, 100081
[2] Ruixin Academy of Classic Learning, Beijing Institute of Technology, Beijing, China, 100081

Abstract

Recombinant innovation is considered a significant driver in generating new ideas, and it has been evidenced to have a higher rate of occurrence in scientific papers. Therefore, modeling and measuring the combination of scientific knowledge in articles has garnered widespread research interest. This paper aims to provide a new perspective to understand and measure the absorption and integration of scientific ideas and insights by leveraging knowledge networks. The references and content of the articles function as input for knowledge absorption and output for knowledge integration, respectively, in which the content refers to the substance or core elements found within the articles. These knowledge elements are extracted using KeyBERT, fused and consolidated with string fuzzy match and embedding-based semantic similarity provided by SciBERT, and labeled as supplied knowledge elements, absorbed knowledge elements, and generated knowledge elements. Knowledge networks are then constructed using the extracted elements and the cooccurrence of elements. Three types of metrics are developed to measure the structure and properties of knowledge networks, including descriptive statistics of nodes, degrees, edges, and components, network global structure metrics, and knowledge proximity calculated using document embedding. We finally use the key publications of the Nobel prize in physics to perform an empirical study.

Keywords

Knowledge Network, Knowledge Elements, KeyBERT, Knowledge Absorption, Knowledge Integration

1. Introduction

Innovation is the result of a combination of knowledge [1, 2]. The use of knowledge combination and recombination concepts has gained momentum in the literature over the past decade. As reviewed by Xiao, Makhija and Karim, more than 1,000 articles published in top management journals exploited the logic to some extent [1]. Since recombinant innovation is considered a significant driver in generating new ideas and has been evidenced to have a higher rate of occurrence in scientific papers [3, 4], it has garnered widespread interest to understand and measure the combination and integration of scientific knowledge in papers.

Scientific papers are one of the main carriers of innovative achievements. Researchers absorb data, information and knowledge by referencing the existing literature, and subsequently generate innovative ideas and insights with knowledge combination and integration [3, 5, 6]. The procedure of knowledge absorption is encoded by the content of the references [7], while the process of knowledge integration can be evaluated by examining the content of the articles themselves [8]. The content here refers to the substance or core elements of scientific papers. The prior literature considers IPC codes [9, 10], keywords [11], key phrases [12], MeSH terms [13], topics or predefined tags [14] as a proxy for knowledge elements. Building on these, the knowledge elements in this study refer to the integral and core concepts of a scientific article.

Although 'content' is a more direct reflection of knowledge absorption and integration, domain experts use citation patterns more frequently than using the body of knowledge to analyze evolutionary trajectories [13, 15-17]. Existing research argues that simple citation patterns provide noisy measurements of knowledge recombination because citation behavior is usually complicated [18-20]. With the development of text mining technologies, citation patterns that work with rough-grained document comprehension, such as article keywords or topics, and predefined categorizations, such as IPC codes, have been used to investigate the process of knowledge absorption [21, 22], yet both rough-grained topics and explicit taxonomies have

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online

Hongshu.Chen@bit.edu.cn (H. Chen); jkl200407@gmail.com (J. Liu); qcliu018@gmail.com (Z. Liu)

*Corresponding author.

limitations in directly indicating fundamental content and context. Moreover, recent studies have inspired discussions that beyond simple pairwise combinations, higher-order network structure is also important for understanding research contents and contexts for scientific innovations [23]. Thus, further research on reflecting knowledge elements and their structures thus is still warranted [24].

This paper aims to provide a new perspective of knowledge networks for understanding and measuring the absorption and integration of scientific ideas and insights. Knowledge elements are extracted using KeyBERT, and consolidated with string fuzzy match and semantic similarity provided by SciBERT, then labeled as supplied knowledge elements, absorbed knowledge elements, and generated knowledge elements. Knowledge networks are then constructed to reflect the structure of labeled knowledge elements in the absorption input and integration output. Three types of metrics are developed to measure the structure and properties of scientific knowledge combination in networks, including descriptive statistics, network global structure metrics, and knowledge proximity. We finally use the key publications of the Nobel prize in physics to perform an empirical study.

2. Methodology

2.1. Framework

The challenge of measuring the absorption and integration of scientific knowledge combination requires modeling supplied knowledge elements, absorbed knowledge elements, generated knowledge elements and their structures. According to the theory of recombination, continuous combination and reconstruction of knowledge elements in a knowledge network led to innovation. We propose a knowledge network model to analyze knowledge absorption and integration.

Building on prior literatures, the knowledge elements in this study refer to the integral and core concepts of a scientific article. The framework of knowledge network model is shown in Figure 1. The data used to construct the network are scientific papers from the Web of Science. One target paper may potentially have n references. The title and abstract of a target paper are merged as the 'output' of knowledge integration, whereas the titles and abstracts of references are seen as the 'input' of knowledge absorption. Specifically, n-gram terms are extracted from textual data using KeyBERT as candidate knowledge elements. Then they are cleaned with a stop word list and filtered using TFIDF values. This makes the selected elements capable of reflecting the main content of the article well in terms of semantics and having importance in statistical terms.

Furthermore, we finalize K knowledge elements for the target paper, and $K \times n$ elements for corresponding references. These elements are merged using string fuzzy match and embedding-based semantic similarity provided by SciBERT, to fuse knowledge elements with same concept and similar semantic meaning that exist in both absorption and integration phrases. A network containing all knowledge elements from the absorption and integration phases can then be constructed, in which each node represents a knowledge element, and the edges represent the co-occurrence relationships between knowledge elements. There are three types of nodes: supplied knowledge elements, absorbed knowledge elements, and generated knowledge elements. This paper develops three types of metrics for measuring knowledge elements and structure of knowledge networks, including descriptive statistics of nodes, node degrees and edges, network global structures metrics, and knowledge proximity calculated using document embedding.

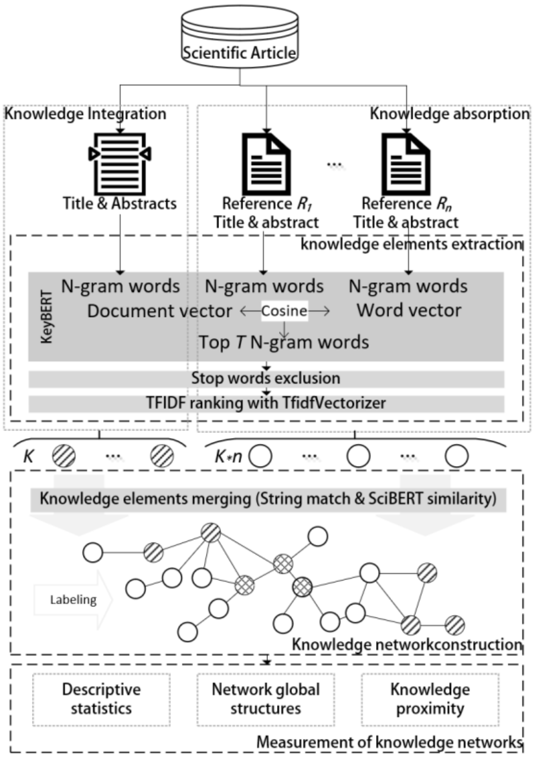


Figure 1: The Knowledge Network Model Framework for Analyzing Knowledge Absorption and Integration

2.2. Knowledge Elements Extraction with KeyBERT

The extraction of knowledge elements is the foundation for studying knowledge formation mechanisms and identifying innovation. Predefined taxonomies such as 'WoS Categories' offered by Web of Science and rough-grained topics have limitations in sculpting the details of knowledge flows. We refine the granularity of knowledge elements to n-gram terms in this paper and extract these key elements of knowledge using the Python KeyBERT package [25].

KeyBERT leverages BERT embeddings to extract the keywords most similar to a given document [26]. The BERT model Sentence-Transformers is applied to obtain vector representation at the document level first, then cosine similarity is used to find the top T most similar keywords that best describe the entire document. As shown in Figure1, T n-gram keywords are extracted with the KeyBERT from merged title and abstract of target paper and each referenced article. In this research, the length of terms is set as one word and two words ($n = 1, 2$); and T is set to 50 to select the top terms that can best deliver the semantic content of each document. Then a stop word list is applied to clean the selected term list and TFDIF values are computed for all these terms to show their statistical significance. We finally set $K = 10$, which means that most representative and important 10 terms are maintained for each document and will be used for knowledge network construction.

2.3. Knowledge Network Construction

According to the theory of knowledge recombination, continuous combination and reconstruction of knowledge elements in a knowledge network led to innovation. In a typical knowledge network, $G_K(V_{elements}, E)$, each node represents an extracted knowledge element, and the edges represent the co-occurrence relationships between knowledge elements.

As shown in Figure 1, the knowledge element extraction module extracted a total of $K(n+1)$ knowledge elements. These terms are first merged using fuzzy string match and semantic similarity calculated using SciBERT embedding, in order to fuse knowledge elements with the same concept and similar semantic meaning in the complete process of knowledge absorption and integration. The knowledge elements that have been finalized act as the nodes within the network, while their co-occurrence in the article and references form the edges of the network. As shown in Figure 2, these knowledge elements are labeled as supplied knowledge elements that are provided by references; absorbed knowledge elements that exist in both the target paper and its references; and generated knowledge elements that exist only in the title & abstract.

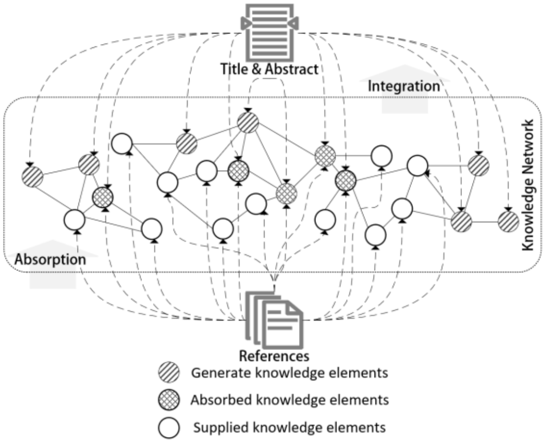

Title & Abstract
Integration
Knowledge Network
Absorption
References
Generate knowledge elements
Absorbed knowledge elements
Supplied knowledge elements

Figure 2: Schematic Diagram of the Knowledge Network

2.4. Measurement of Knowledge Networks

2.4.1. Descriptive Statistics

Descriptive statistical metrics are used to measure the total number of nodes, edges, and components in the knowledge network to analyze the network scale. We compute the knowledge absorption efficiency using formular (1), in which $V_{Absorbed\ elements}$ represents the number of absorbed knowledge elements and $V_{elements}$ represents the total number of unique knowledge elements in the network. We also compute the knowledge integration efficiency via formular (2), in which $V_{Generated\ elements}$ represents the number of generated knowledge elements and $V_{elements}$ is the total number of unique elements. In addition, we also calculate the degree distribution, which is the probability distribution of degrees over the knowledge network, to measure the complexity of the networks.

$$Absorption = V_{Absorbed\ elements}/V_{elements} \quad (1)$$

$$Integration = V_{Generated\ elements}/V_{elements} \quad (2)$$

2.4.2. Global Network Metrics

The relationships among elements within a network are reflected by the network structures. As a result, structural characteristics significantly influence future interactions of elements. The proposed model considers metrics related to macro-structural characteristics of the network, including network density, average path length, and clustering coefficient [27, 28].

Network density describes the connectivity of a network. It is calculated by dividing the total number of connections E by the total number of possible connections E_{Max} with the same number of nodes, which can be computed using formular $E_{Max} = n(n-1)/2$ [29].When the network density is lower, the connections between the knowledge elements are sparser, suggesting that the behaviors of elements are less influenced by the network structure.

$$D = E/E_{Max} \quad (3)$$

Average path length L can be calculated as the average length of the shortest path between any two nodes, as shown in (4), where the distance between node i and node j, d_{ij}, represents the total number of links connecting the shortest path between the two nodes. The value of N denotes the total number of nodes in the network. It helps distinguish negotiable networks from comparatively inefficient ones.

$$L = \frac{2}{N(N-1)} \sum_{i \neq j} d_{ij} \quad (4)$$

The global clustering coefficient is calculated via formular (5), in which C stands for local clustering coefficient, and $C_i = 2E_i/k_i(k_i - 1)$, and E_i represents the number of edges for vertex i, and $\mathrm{k_i}$ indicates the degree of vertex i [30]. A higher clustering coefficient is an indication of a small world.

$$C = \frac{\sum_i C_i}{n} \quad (5)$$

2.4.3. Knowledge Proximity

The cosine distance of embedding-based vectors derived from scientific articles and patents can be used to quantify the proximity of knowledge [31]. Embedding-based approaches avoid creating high-dimensional sparse vectors in mapping massive textual data, thus having great potential for feature extraction and knowledge representation. We apply doc2vec to map documents into fixed-length numeric vectors, translate latent semantics into low-dimensional dense space [32, 33], and measure the proximity of knowledge of the target article and its references. The knowledge proximity metric shows the semantic distance between the content of the references and the target paper. The greater the semantic distance, the larger the changes in the integrated content after the absorption phrase.

3. Empirical study

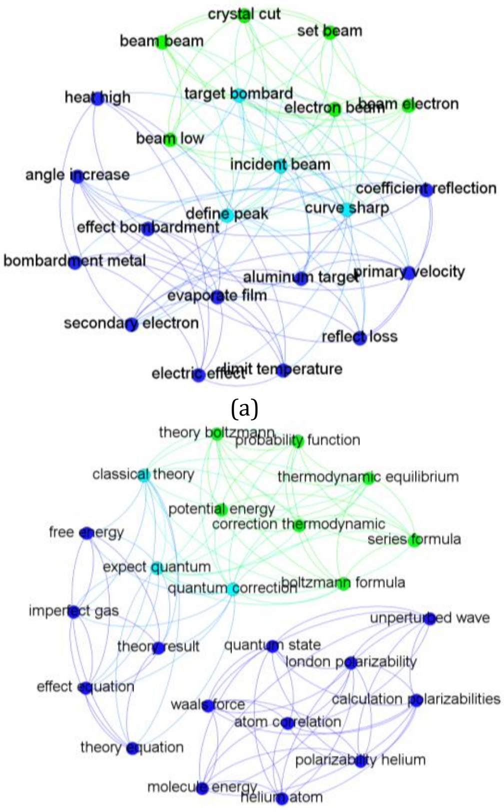
In this paper, we choose the key publications of Nobel prize in physics[1]. Only papers published before 2004 were included. Initially, we collected 179 prize winning papers. In order to analyze the knowledge absorption and integration using content of publications, we only keep papers that have three or more references, and the language of writing is limited to English. We finalize 124 Nobel prize papers in physics as the dataset for empirical study.

3.1. Descriptive Statistics of Knowledge Networks of Nobel Prize-Winning Papers

After generating the corresponding knowledge network using the model proposed in this paper, the 124 Nobel Prize-winning papers have an average of 31.85 supplied knowledge elements, 5.54 absorbed knowledge elements and 2.64 generated knowledge elements. There are 120 connected knowledge networks and the other 4 are unconnected ones. The average knowledge absorption efficiency is 0.14, and the average knowledge integration efficiency is 0.09.

Figure 3 (a) depicts a representative connected knowledge network, which is constructed based on the data of the article identified as WOS-000201553700001. All knowledge elements can be reached by a single random walk within the knowledge network. Figure 3 (b) illustrates an unconnected knowledge network, created using data from the article identified as WOS-000201591300009. Supplied knowledge elements are marked in deep blue, absorbed knowledge elements are highlighted in light blue and generated knowledge elements are colored in green.

We then compute the degree distribution of each knowledge network to measure their complexity. To summarize the degree distribution in the dataset, we illustrate the average node degree distribution in Figure 4. The distribution is right-skewed, showing majority of the node degree values are concentrated in the range of 9 to 14, and some of the nodes have even higher degree values. There is no isolated vertex nor pendant vertex in the knowledge networks.



(a)

(b)

Figure 3: (a) One Example of a Connected Knowledge Network; (b) One Example of an Unconnected Knowledge Network

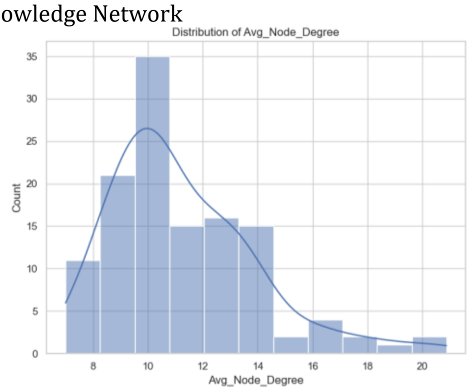


Figure 4: Average Node Degree Distribution for 124 Nobel Prize-Winning Papers

3.2. Global Network Metrics of Knowledge Networks of Nobel Prize-Winning Papers

To measure the structural characteristics of knowledge networks, we calculate the density distribution for all the papers and present the result in Figure 5. As this metric is measured on a scale of 0 to 1, lower values indicate knowledge networks with fewer relationships and higher values represent knowledge networks with more relationships. A value closer to 0 illustrates a sparser network with fewer connections, while a value closer to 1 indicates a denser network with stronger connections between nodes. Figure 5

[1] https://www.nobelprize.org/prizes/

also shows a right-skewed distribution, which means that the majority of the knowledge networks are sparse ones. There is still a lot of potential to have more connections in knowledge networks.

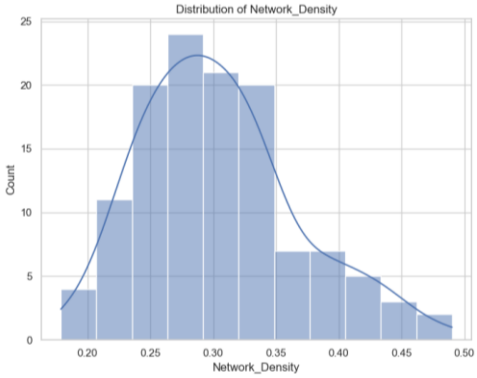


Figure 5: Density Distribution for 124 Nobel Prize-Winning Papers

The average shortest path length distribution and average clustering coefficient distribution are then computed and shown in Figure 6 and Figure 7 correspondingly. These networks demonstrate strong small-world properties, characterized by a high clustering coefficient and a short average path length.

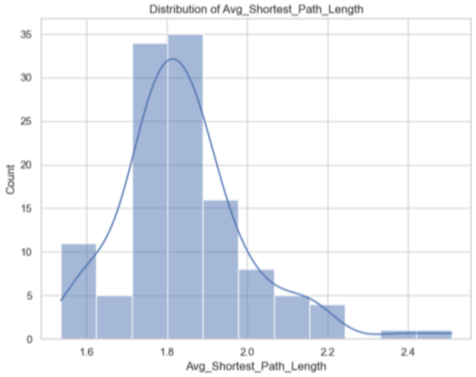


Figure 6: Average Shortest Path Length Distribution

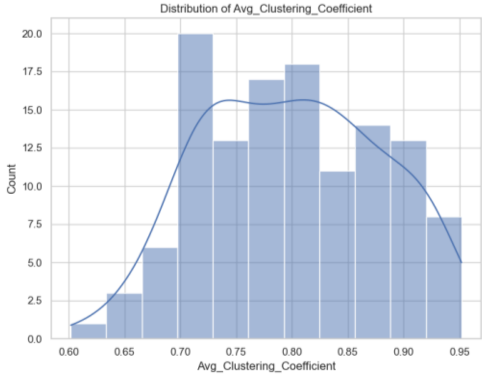


Figure 7: Average Clustering Coefficient Distribution

In addition, to evaluate the knowledge proximity of absorption and integration phrase of the knowledge combination in these papers. We then compute cosine distance of 150-dimension embedding-based vectors of each target paper and their references. These vectors are generated using doc2vec. The average semantic similarity is 0.80.

4. Conclusion, Limitations and Future Work

In this paper, we explore a new perspective on modeling knowledge absorption and integration with knowledge networks. The references and content of the articles function as input to knowledge absorption and output to knowledge integration, respectively. Although this paper provides heuristic research that can potentially be used to model and measure the process and result of knowledge combination, it has several limitations that need to be explored in future research: First and foremost, (1) at this stage, this study has not established a control group for the experimental group to further investigate whether the indicators provided by the model can effectively reflect the effects of knowledge integration and innovation; in addition, (2) the number of references indirectly affects the size of the current knowledge network, and this influence needs to be minimized by further adjusting the network's nodes and edges; (3) there are limitations in constructing network edges solely based on term co-occurrence relationships; (4) more metrics need to be design to measure the efficiency and structure of knowledge absorption and integration.

In future research, a control group that can be well matched with Nobel Prize-winning papers needs to be established. We will address the above concerns in future research so as to keep improving the methodology in representing and analyzing the complex system of knowledge combination and recombinant innovation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 72004009).

References

[1] Xiao, T., Makhija, M. & Karim, S. A Knowledge Recombination Perspective of Innovation: Review and New Research Directions. Journal of Management 48 (2022) 1724-1777, doi:10.1177/01492063211055982.

[2] Savino, T., Messeni Petruzzelli, A. & Albino, V. Search and Recombination Process to Innovate: A Review of the Empirical Evidence and a Research Agenda. International Journal of Management Reviews 19 (2017) 54-75, doi:10.1111/ijmr.12081.

[3] Fortunato, S. et al. Science of science. Science 359 (2018) eaao0185, doi:10.1126/science.aao0185.

[4] Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical Combinations and Scientific Impact. Science 342 (2013) 468-472, doi:10.1126/science.1240474.

[5] Sternberg, R. J. The nature of creativity. Creativity Research Journal 18 (2006) 87-98, doi:10.1207/s15326934crj1801_10.

[6] Fleming, L. Recombinant Uncertainty in Technological Search. Management Science 47 (2001) 117-132, doi:10.1287/mnsc.47.1.117.10671.

[7] Jaffe, A. B., Trajtenberg, M. & Henderson, R. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. The Quarterly

Journal of Economics 108 (1993) 577-598, doi:10.2307/2118401.

[8] Kaufman, J. C. & Beghetto, R. A. Beyond Big and Little: The Four C Model of Creativity. Review Of General Psychology 13 (2009) 1-12, doi:10.1037/a0013688.

[9] Wang, C. L., Rodan, S., Fruin, M. & Xu, X. Y. Knowledge Networks, Collaboration Networks, And Exploratory Innovation. Acad. Manage. J. 57 (2014) 484-514, doi:10.5465/amj.2011.0917.

[10] Brennecke, J. & Rank, O. The firm's knowledge network and the transfer of advice among corporate inventors-A multilevel network study. Research Policy 46 (2017) 768-783, doi:10.1016/j.respol.2017.02.002.

[11] Guan, J. C., Yan, Y. & Zhang, J. J. The impact of collaboration and knowledge networks on citations. Journal of Informetrics 11 (2017) 407-422, doi:10.1016/j.joi.2017.02.007.

[12] Jee, J., Park, S. & Lee, S. Potential of patent image data as technology intelligence source. Journal of Informetrics 16 (2022) 19, doi:10.1016/j.joi.2022.101263.

[13] Wang, S. et al. Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. Journal of the Association for Information Science and Technology 74 (2023) 150-167, doi:10.1002/asi.24719.

[14] Li, Y. N., Wu, Y., Zhang, L. & Chen, J. W. The Impact of Network Structure on Knowledge Adoption: A Network Text Analysis on Knowledge-Sharing Platforms. IEEE Trans. Comput. Soc. Syst. (2023) 16, doi:10.1109/tcss.2023.3255588.

[15] Mugabushaka, A.-M., Sadat, J. & Faria, J. C. D. In Search of Outstanding Research Advances: Prototyping the creation of an open dataset of" editorial highlights". arXiv preprint arXiv:2011.07910 (2020).

[16] Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. Nature 566 (2019) 378-382, doi:10.1038/s41586-019-0941-9.

[17] Roach, M. & Cohen, W. M. Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. MANAGEMENT SCIENCE 59 (2013) 504-525, doi:10.1287/mnsc.1120.1644.

[18] Meyer, M. Does science push technology? Patents citing scientific literature. Research Policy 29 (2000) 409-434, doi:10.1016/S0048-7333(99)00040-2.

[19] Zhang, L., Sun, B. B., Chinchilla-Rodriguez, Z., Chen, L. X. & Huang, Y. Interdisciplinarity and collaboration: on the relationship between disciplinary diversity in departmental affiliations and reference lists. Scientometrics 117 (2018) 271-291, doi:10.1007/s11192-018-2853-0.

[20] Zhang, G., Ding, Y. & Milojević, S. Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology 64 (2013) 1490-1503, doi:10.1002/asi.22850.

[21] Wagner, C. S. et al. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. Journal of Informetrics 5 (2011) 14-26, doi:10.1016/j.joi.2010.06.004.

[22] Bu, Y., Li, M., Gu, W. & Huang, W.-b. Topic diversity: A discipline scheme-free diversity measurement for journals. Journal of the Association for Information Science and Technology 72 (2021) 523-539, doi:10.1002/asi.24433.

[23] Shi, F. & Evans, J. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nature Communications 14 (2023) 1641.

[24] Seibert, S. E., Kacmar, K. M., Kraimer, M. L., Downes, P. E. & Noble, D. The Role of Research Strategies and Professional Networks in Management Scholars' Productivity. Journal of Management 43 (2017) 1103-1130, doi:10.1177/0149206314546196.

[25] Trappey, A. J. C. et al. Patent landscape and key technology interaction roadmap using graph convolutional network – Case of mobile communication technologies beyond 5G. Journal of Informetrics 17 (2023) 101354, doi:10.1016/j.joi.2022.101354.

[26] Reimers, N., & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv: 1908.10084. (2019).

[27] Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G. & van den Oord, A. Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. Research Policy 37 (2008) 1717-1731, doi:10.1016/j.respol.2008.08.010.

[28] Fleming, L. & Frenken, K. The evolution of inventor networks in the silicon valley and Boston regions. Advances in Complex Systems 10 (2007) 53-71, doi:10.1142/S0219525907000921.

[29] Kong, X., Shi, Y., Yu, S., Liu, J. & Xia, F. Academic social networks: Modeling, analysis, mining and applications. Journal of Network and Computer Applications 132 (2019) 86-103, doi:10.1016/j.jnca.2019.01.029.

[30] Newman, M. E. The structure and function of complex networks. SIAM review 45 (2003) 167-256.

[31] Feng, S. J. The proximity of ideas: An analysis of patent text using machine learning. Plos One 15 (2020) 19, doi:10.1371/journal.pone.0234880.

[32] Le, Q. & Mikolov, T. Distributed representations of sentences and documents. in Proceedings of the 31st International Conference on Machine Learning - Volume 32. Beijing, China, 2014, pp. II–1188–II–1196.

[33] Zhang, Y. et al. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. Journal of Informetrics 12 (2018) 1099-1117, doi:10.1016/j.joi.2018.09.004.