



Stepping beyond your comfort zone: Diffusion-based network analytics for knowledge trajectory recommendation

Yi Zhang  | Mengjia Wu  | Guangquan Zhang | Jie Lu

Australian Artificial Intelligence Institute,
Faculty of Engineering and Information
Technology, University of Technology
Sydney, Sydney, New South Wales,
Australia

Correspondence

Yi Zhang, Australian Artificial
Intelligence Institute, Faculty of
Engineering and Information Technology,
University of Technology Sydney, Sydney,
NSW, Australia.
Email: yi.zhang@uts.edu.au

Funding information

Australian Research Council,
Grant/Award Number: DE190100994

Abstract

Predicting a researcher's knowledge trajectories beyond their current foci can leverage potential inter-/cross-/multi-disciplinary interactions to achieve exploratory innovation. In this study, we present a method of diffusion-based network analytics for knowledge trajectory recommendation. The method begins by constructing a heterogeneous bibliometric network consisting of a co-topic layer and a co-authorship layer. A novel link prediction approach with a diffusion strategy is then used to capture the interactions between social elements (e.g., collaboration) and knowledge elements (e.g., technological similarity) in the process of exploratory innovation. This diffusion strategy differentiates the interactions occurring among homogeneous and heterogeneous nodes in the heterogeneous bibliometric network and weights the strengths of these interactions. Two sets of experiments—one with a local dataset and the other with a global dataset—demonstrate that the proposed method is prior to 10 selected baselines in link prediction, recommender systems, and upstream graph representation learning. A case study recommending knowledge trajectories of information scientists with topical hierarchy and explainable mediators reveals the proposed method's reliability and potential practical uses in broad scenarios.

1 | INTRODUCTION

Dating back to the early 1980s, the continuous and discontinuous technological changes drew attention from Dosi (1982). He defined the continuous changes as technological trajectories, emphasizing the cumulative process of technical advances in an established routine. When assembling scientific research and technological development as knowledge, *knowledge trajectories* refer to how knowledge is integrated and differentiated within this dynamic changing process (Barley et al., 2018).

Understanding the dynamics of knowledge trajectories is relevant to the broad interests of science, technology, and innovation (ST&I) studies. For example, disruptive innovation (Christensen et al., 2018) and recombinant innovation (Uzzi et al., 2013) investigate how knowledge interacts with each other in creating inventions. When differentiating the type of knowledge in an innovation process, *exploratory innovation* is known as a radical exploration with external knowledge, and *exploitative innovation* is an incremental exploitation that deepens internal knowledge and skills (Jansen et al., 2006).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

Social scientists recognize exploration and exploitation as alternative research strategies for scholars to shape/reshape knowledge trajectories and promote research productivity and impact (Foster et al., 2015; Huang et al., 2022). Besides science policies, the literature has identified several crucial determinants of exploratory innovation (Keshavarz & Shekari, 2020; Liu, Wang, et al., 2018a; Zeng et al., 2019), for example, career age and hot streaks, research achievement and reputation, and topic nature. More significantly, social interactions on the ST&I dynamics have been extensively investigated (Acar et al., 2019; Sun et al., 2013) and were specified into broad collaborations between multiple disciplines (Nicolini et al., 2012) and between academia and industry (Steinmo & Rasmussen, 2018). The insights inspire us that internal knowledge interactions and external social interactions can trigger the dynamics of knowledge trajectories, and thus covering both aspects could be essential for predicting knowledge trajectories.

Network analytics, particularly bibliometric network analytics, has been widely applied in ST&I studies, for example, measuring research impact (Yan & Ding, 2009) and tracing technological trends (Leydesdorff & Rafols, 2011). Significantly, link prediction, with its core assumption that two unconnected nodes can be linked in the future if common neighbors exist (Liben-Nowell & Kleinberg, 2007), surprisingly coincides with the recombination theory (Uzzi et al., 2013). Under this assumption, some studies used link prediction to recommend potential collaborators (Yan & Guns, 2014) and predict emerging technologies (Érdi et al., 2013; Zhou et al., 2019) by analyzing a homogeneous bibliometric network, for example, a co-term, co-citation, or co-authorship network. However, aiming to highlight the internal and external determinants of exploratory innovation, we argue that a heterogeneous network consisting of social elements (e.g., co-authorships) and knowledge elements (e.g., term co-occurrence) can comprehensively describe the scenario that knowledge diffuses through research collaboration and between similar technologies. Then, predicting missing links between these social and knowledge elements can identify how their interactions will evolve in the future and foresee future knowledge trajectories. However, modeling and measuring the complicated social and knowledge interactions in a heterogeneous bibliometric network is still challenging.

Following the definition given by Barley et al. (2018), this study considers the knowledge trajectories of an individual researcher as the historical and future changing processes of their research topics. It aims to predict the changing process by recommending research topics beyond their existing knowledge base to achieve exploratory innovation. Although developing knowledge trajectories is influenced by various internal and external

factors, particularly a scholar's research foci, our recommendations emphasize extending a scholar's existing knowledge base from the perspective of inter-/cross-/multi-disciplinary interactions; that is, what we call "stepping beyond the comfort zone." Even so, this study simulates social and knowledge interactions within knowledge diffusion to consider the role of a scholar's knowledge accumulation and interpersonal communications in establishing their knowledge trajectories.

This paper proposes a novel method for analyzing a heterogeneous bibliometric network and recommending knowledge trajectories to target researchers. The method begins by constructing a heterogeneous bibliometric network with a co-topic layer and a co-authorship layer. The interactions among homogeneous and heterogeneous nodes in this network are then predicted by a model of diffusion-based link prediction that relies on network-based inference (NBI) (Zhou et al., 2007). While modern deep learning-based approaches cannot sufficiently explain their results, the diffusion process provides clues for interpreting recommendations through involved mediators, that is, common neighbors such as collaborators and similar technologies. Furthermore, our model extends the scope of the inference from a bipartite network to a bi-layer network, since bipartite network analytics ignores knowledge diffusion between homogeneous nodes. Yet, as argued above, such interactions may reflect significant academic activities.

Ten baselines are selected for validation measurements: Six link prediction baselines, two recommendation baselines, and two upstream machine learning baselines on graph representation learning. We assembled two datasets for testing: a local dataset that contains 11,399 journal articles from the information science literature, and a global dataset comprising the complete set of the Digital Bibliography & Library Project (DBLP) database, covering 4.89 million research articles in computer science. The validation demonstrates the reliability of our method in recommending knowledge trajectories. Beyond our experiments, we conducted a case study using the local dataset to reveal the practical application of the proposed method. These insights can help provide empirical decision support to individual researchers, research institutions, and funding agencies in the information science discipline.

2 | RELATED WORK

2.1 | Bibliometric network analytics

In this paper, a *bibliometric network* refers to a network consisting of bibliometric entities (e.g., terms and

authors) and their relationships (e.g., co-occurrence). Information scientists leapt at the opportunity to apply network analytics to explore insights from network topologies (Björneborn, 2004). Previous bibliometric network analytics have: (a) used topological indicators (e.g., centrality) to identify key nodes, for example, influential researchers in a co-authorship network (Li et al., 2013; Yan & Ding, 2009); (b) used topology-based approaches (e.g., community detection and link prediction) to recognize specific behaviors and patterns, for example, collaborations (Yan & Guns, 2014), disciplinary interactions (Huang et al., 2020), and problem-solving patterns (Zhang, Wu, Hu, et al., 2021a); and (c) connected bibliometric networks with broad ST&I paradigms, for example, technology roadmaps (Jeong et al., 2021) and technology opportunity analysis (Ren & Zhao, 2021).

Sun and Han (2012) argued, “the interactions among multi-typed objects play a key role in disclosing the rich semantics that a network carries” and defined a meta path as sequential links between any two entities in a heterogeneous network. With pre-defined meta paths, link prediction has been widely recognized as a downstream task of heterogeneous network mining, which holds interpretable capabilities in inferring potential connections between pairwise nodes through their common neighbors (Dong et al., 2020). Rich studies have been observed on elaborating heterogeneous entities and relationships within meta paths for link prediction, for example, tracing co-authorship evolution using a knowledge graph with multi-entities and multi-relations (Zhang, 2017), measuring emerging technologies through a bi-layer network (Zhang, Wu, Miao, et al., 2021b), and recommending publication venues based on a network with multiple bibliometric entities (Kleśniński et al., 2021).

This study follows the tradition of heterogeneous network mining. Its core method aligns with link prediction, highlighting (a) the design of meta paths reflecting knowledge diffusion with social and knowledge interactions in exploratory innovation; and (b) its interpretable capabilities in explaining prediction results through the mediators in a diffusion process.

2.2 | Scholarly recommendation

Recommending knowledge trajectories aligns with scholarly recommendations, targeting academic researchers and recommending academic outlets (Alhoori & Furuta, 2017) and counterparts (e.g., collaborators, reviewers, and supervisors) (Liu, Xie, & Chen, 2018b; Rahdari et al., 2020). Besides traditional content-based and collaborative filtering-based approaches, previous studies extensively facilitated the natural tie of scholarly recommendations with knowledge graphs, and introduced graph

representation learning to assemble heterogeneous attributes and represent entities in low dimensional vectors (Sun et al., 2021). For scholarly recommendations with bibliometric networks, Zhu et al. (2022) applied a translation-based approach to embed multiple bibliometric entities (e.g., authors, papers, and departments) and their relationships in a million-scale bibliometric network for co-authorship prediction. Aiming at research leadership recommendation, He et al. (2022) adopted an autoencoder model to represent authors with diverse features, for example, cognitive, geographical, and organizational proximities.

While traditional recommender systems usually consider homogeneous relationships, this study highlights the understanding of heterogeneous entities and their relationships (e.g., social and knowledge interactions) within the theoretical framework of exploratory innovation. Methodologically, we fully acknowledge the advantages of graph representation learning techniques (Dong et al., 2020) in large-scale recommendations. However, we also appreciate detailed meta-paths (e.g., diffusion) defined in a heterogeneous network, which avoid potential information loss and provide extra information for interpreting recommendations—tracing back along with a meta-path to identify core mediators. Additionally, with a shared focus on discovering user-item connections, link prediction and recommender systems are categorized as two overlapped downstream tasks in the computer science literature (Zhang et al., 2019). This study targets a recommendation task, but its core methodology is built on link prediction, highlighting the use of meta-paths defined from a heterogeneous bibliometric network.

3 | METHODOLOGY: DIFFUSION-BASED NETWORK ANALYTICS

3.1 | Theoretical basis: Exploratory innovation and knowledge diffusion

When the recombinant innovation theory well studied the role of knowledge interactions in an innovation process (Uzzi et al., 2013), the literature thoroughly discussed the positive correlations between a firm's collaborative network and its ability to exploratory innovation (Phelps, 2010), known as a radical innovation that “require(s) new knowledge or departure(s) from existing knowledge” (Jansen et al., 2006). Besides isolated interactions with either social or knowledge determinants, their synthesized impacts on the dynamics of ST&I are significant (Sun et al., 2013; Wang et al., 2014). In a large-scale science of science study, Huang et al. (2022) observed the

preference of productive and impactful researchers in exploratory innovation. We thus summarized:

Assumption 1. Exploratory innovation is positively correlated with research productivity and impacts, and is synergistically influenced by social and knowledge interactions.

Knowledge diffusion, known as the adaptations of scientific knowledge from scientific research to technological innovation (Sorenson & Fleming, 2004), follows a general process of innovation diffusion—“an innovation is communicated through different channels in a certain time among the members of a social system” (Rogers, 2003). The literature discussed the importance of knowledge features and the channels of transmitting knowledge in a diffusion process (Zanello et al., 2016). It highlighted close interpersonal ties promote knowledge diffusions, such as research collaboration, geographical localisation, and firm boundaries (Singh, 2005). Given that, we drew the following:

Assumption 2. Knowledge diffuses between knowledge elements, between social elements, and between social and knowledge elements. The three types of diffusion represent technological similarity, research collaboration, and knowledge adoption, respectively.

Following the two assumptions, this paper designs a bi-layer bibliometric network, consisting of a co-topic layer and a co-authorship layer, to describe scientific activities through a socio-technical system (Assumption 1). The diffusion-based network analytics simulates an innovation process in which knowledge diffuses among social and knowledge elements (Assumption 2), and predicts future interactions between a target social element and broad knowledge elements. The newly established interactions may create clues for exploratory innovation (Assumption 1), referring to inter-/cross-disciplinary recombinations with either upstream methodologies or downstream applications.

This study is on the trail of intelligent bibliometrics (Zhang et al., 2020)—developing computational models that elaborate artificial intelligence and data science techniques with bibliometric indicators for handling issues in ST&I studies. The research framework is given in Figure 1. It includes three phases: data pre-processing, bi-layer network construction, and diffusion-based prediction.

3.2 | Phase I: Data pre-processing

The proposed method targets bibliometric data, such as scientific documents, patents, and academic proposals.

More specifically, two fields of bibliographical information are focused: the combined text of an article's title and abstract, and its authorship information. Thus, two pre-processing tasks are required:

- *Pre-processing author names:* Author names in raw bibliometric data may vary hugely, appearing as, say, “Eugene Garfield,” “Garfield, Eugene,” “Garfield, E,” and “E Garfield.” Author name disambiguation is therefore required to consolidate variations and remove unidentified names.
- *Topic extraction and representation:* Targeting the titles and abstracts, topic extraction (e.g., topic models) is conducted to identify and label topics from the corpus.

Recently, most knowledge graphs (e.g., Microsoft Academic Graph, MAG) have disambiguated author names and identified topics, and thus appropriately facilitating these benefits may skip off this pre-processing phase.

3.3 | Phase II: Bi-layer network construction

Following Assumption 1 and highlighting the social and knowledge interactions in the process of exploratory innovation, the bi-layer network consists of a co-authorship layer and a co-topic layer. Briefly, on the co-authorship/co-topic layer, a node represents an author/topic. An edge represents the co-occurrence between connected authors/topics, weighted by their co-occurrence frequency. Notably, one of the most recent studies on topic taxonomy construction (Shang et al., 2020) raised a significant drawback of embedding in local text data, that is, word embedding techniques cannot effectively distinguish highly coupled words in a specific domain. We decided to keep the co-occurrence on the co-topic layer in our default setting. However, using a semantic layer to replace the topic player can be an alternative in cases with broad disciplinary interactions.

Referring to Figure 1, the bi-layer network is described as follows:

$$G\{G_a(V_a, E_{aa}), G_t(V_t, E_{tt}), E_{at}\}$$

where $G_a(V_a, E_{aa})$ represents a co-authorship layer with the author nodes V_a and the edges E_{aa} . $G_t(V_t, E_{tt})$ represents a co-topic layer, with topic nodes V_t and edges E_{tt} ; and E_{at} represents the edges between the two layers.

Let $|E|$ represent the weight of an edge E . Let V_a^i and V_a^c be nodes in the co-authorship layer and let V_t^k and V_t^j be nodes on the co-topic layer. Then the weights of the three types of edges can be represented as:

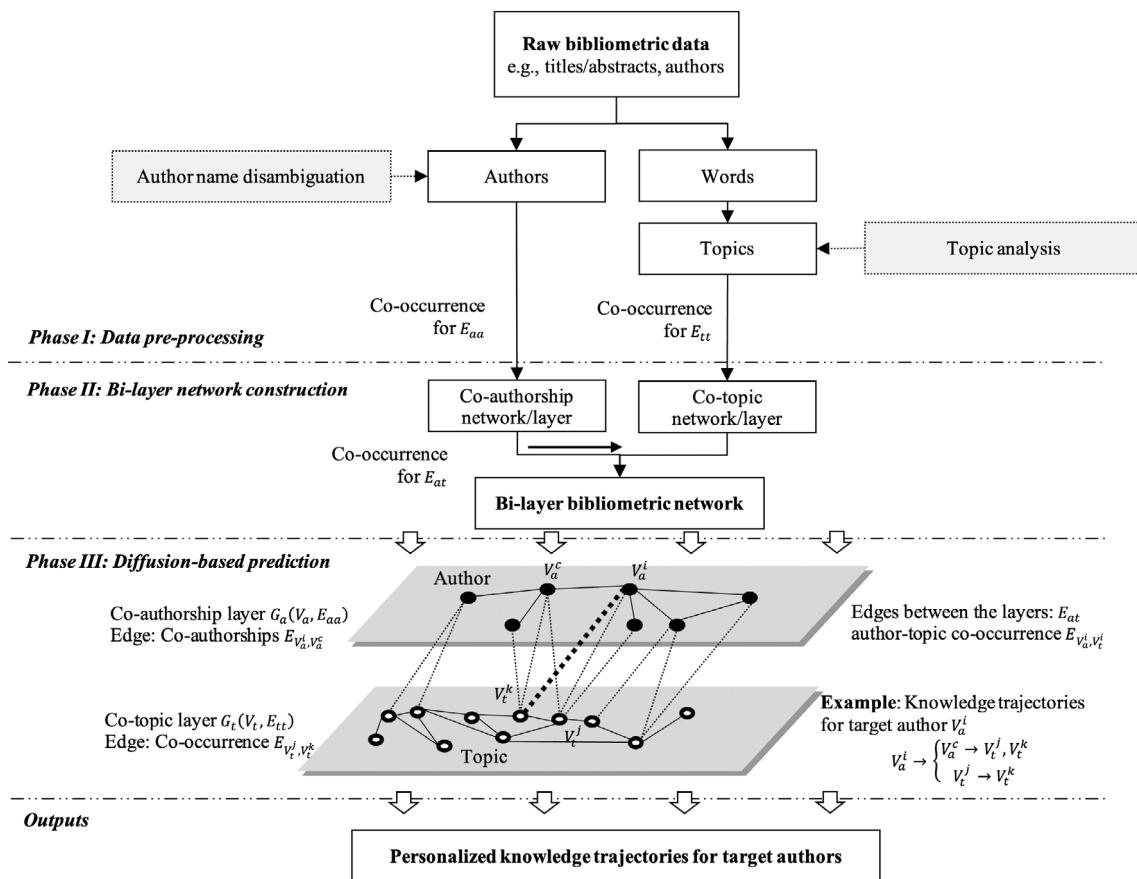


FIGURE 1 Research framework of diffusion-based network analytics for recommending knowledge trajectories.

$$|E_{V_a^i, V_a^c}| = \theta_{V_a^i, V_a^c} \quad (1)$$

$$|E_{V_t^k, V_t^j}| = \varphi_{V_t^k, V_t^j} \quad (2)$$

$$|E_{V_a^i, V_t^j}| = \mu_{V_a^i, V_t^j} \quad (3)$$

where $\theta_{V_a^i, V_a^c}$, $\varphi_{V_t^k, V_t^j}$, and $\mu_{V_a^i, V_t^j}$ are constants, representing the co-occurrent frequency between connected nodes.

3.4 | Phase III: Diffusion-based prediction

Following Assumption 2 and aiming to comprehensively describe knowledge diffusion among social and knowledge elements with diverse diffusion strategies, this study adopts the concept of resource allocation. Initially, in the NBI, Zhou et al. (2007) creatively designed a process of resource allocation in a bipartite network—a network consists of two sets of nodes, in which resources diffuse only between nodes from different sets and their common neighbors (CNs) serve as transmitters to distribute

resources (Ou et al., 2007). Introducing the resource allocation process to a bi-layer network highlights: (a) the core of research collaboration, that is, collaborators exchange and recombine ideas to achieve innovation (Wang et al., 2014); and (b) the basic assumption of CNs in social network analytics, that is, if two nodes have CNs, they may have relations (Yang & Zhang, 2016). Significantly, these CNs act as core mediators in a diffusion process and help interpret recommendations.

Within the NBI's framework, a user's potential preference for an item is measured by the number of resources the item eventually receives from the user (Zhou et al., 2007). Following this trail, our pilot study applied a resource allocation-based link prediction approach (Zhou et al., 2009) to recommend a researcher's potential research interests if the bipartite consisted of authors and terms (Zhang et al., 2018). We further developed a weighted index according to the diffusion strength and applied it to predict potential term-term and term-author connections (Zhang, Wu, Miao, et al., 2021b). However, when describing exploratory innovation in a bipartite network, we may encounter the following issues:

- The interactions between homogeneous nodes in a bipartite bibliometric network are essential for capturing the impacts of social and knowledge interactions on exploratory innovation (Assumption 1).
- Differentiating the diffusion strategies between homogeneous nodes and between heterogeneous nodes is necessary for reflecting actual knowledge diffusion between collaborators and between similar technologies (Assumption 2).

These two issues inspired the methodological development, including extending the scope of the resource allocation from a bipartite network to a bi-layer network and using different diffusion strategies to describe diverse diffusion processes. For this reason, we focus on analyzing three types of edges in the bi-layer network G : E_{aa} , E_{tt} , and E_{at} . Additionally, we redesigned the resource diffusion strategy to predict the potential edges E_{at} between a target author and topics. The algorithm of the diffusion-based prediction for a target author V_a^i is described as follows:

Step 1—Diffusion via author-topic edges $V_a^i \rightarrow V_t^j$: This is a typical resource allocation process designed by the NBI approach but with a weighting solution added to the diffusion strategy. If an author V_a^i holds the initial resources $r(V_a^i)$, a portion of those resources will spread to the connected topics V_t^j . The resource $f(V_t^j)$ that topic V_t^j will receive from author V_a^i can be calculated as:

$$f(V_t^j) = \frac{|E_{V_a^i, V_t^j}|}{\sum_{E_{V_a^i, V_t^p} \neq 0} |E_{V_a^i, V_t^p}|} r(V_a^i) \quad (4)$$

Step 2—Diffusion via author-author edges $V_a^i \rightarrow V_a^c$: Assuming academic researchers are willing to share knowledge with their co-authors, author V_a^i will “copy” the same amount of initial resources $r(V_a^i)$ and spread them to connected authors V_a^c (i.e., co-authors) based on their co-authorship strengths. The resources $f(V_a^c)$ that author V_a^c will receive from author V_a^i can be calculated as:

$$f(V_a^c) = \frac{|E_{V_a^i, V_a^c}|}{\sum_{E_{V_a^i, V_a^q} \neq 0} |E_{V_a^i, V_a^q}|} r(V_a^i) \quad (5)$$

Step 3—Diffusion via topic-topic edges $V_t^j \rightarrow V_t^k$: Assuming the co-occurrence between research topics indicates pairwise knowledge sharing, a topic V_t^j will spread the resources it has acquired to connected topics based on their co-occurrence strengths. The resource $f(V_t^k, V_t^j)$ that topic V_t^k will receive from topic V_t^j and the

total resource $f_t(V_t^k)$ that topic V_t^k will receive from connected topics can be calculated as:

$$f(V_t^k, V_t^j) = \frac{|E_{V_t^j, V_t^k}|}{\sum_{E_{V_t^j, V_t^p} \neq 0} |E_{V_t^j, V_t^p}|} f(V_t^j) \quad (6)$$

$$f_t(V_t^k) = \sum_{E_{V_t^k, V_t^p} \neq 0} f(V_t^k, V_t^p) \quad (7)$$

Step 4—Diffusion via author-topic edges $V_a^c \rightarrow V_t^k$: Repeat step 1 but for the target author's co-authors V_a^c , who will also diffuse their resources to connected topics. Thus, the resources $f(V_t^k, V_a^c)$ that topic V_t^k will receive from the author V_a^c and the total resources $f_a(V_t^k)$ that topic V_t^k will receive from the target author's co-authors can be calculated as:

$$f(V_t^k, V_a^c) = \frac{|E_{V_a^c, V_t^k}|}{\sum_{E_{V_a^c, V_t^p} \neq 0} |E_{V_a^c, V_t^p}|} f(V_a^c, V_a^i) \quad (8)$$

$$f_a(V_t^k) = \sum_{E_{V_a^c, V_t^k} \neq 0} f(V_t^k, V_a^c) \quad (9)$$

Step 5—Resource finalization: Since the objective of this prediction is to recommend new research topics to a target author—that is, topics beyond their comfort zone—our focus is solely on topics unconnected to the target author, that is, topics V_t^k . Thus, the final resource $f(V_t^k)$ that topic V_t^k will receive can be calculated as:

$$f(V_t^k) = f_t(V_t^k) + f_a(V_t^k) \quad (10)$$

Outputs—Ranking and personalized recommendation: The output of the proposed method is a ranking list R containing a list of the target author's V_a^i unconnected topics V_t^k , ranked by their final resource $f(V_t^k)$. This list is personalized, since this list R is generated based on this target author's co-authorships and their research topics. Such a list of recommendations will be different case by case.

Steps 1 and 2 describe a scenario where authors are open to sharing knowledge with their co-authors. Step 3 reveals that co-occurred topics can act as a mediator for knowledge sharing. Both scenarios are designed to effectively simulate knowledge diffusion in real-world scientific activities, and these involved mediators will be identified for interpreting recommendations. Eventually, the model recommends research topics the target author

has never touched, referring to future knowledge trajectories beyond their comfort zone.

3.5 | Validation measurements (10 baselines and 3 measures)

1. Data splitting strategies for training and test sets

Facilitating the publication year of scientific articles, we divided the data into two sub-datasets—one for training and the other for testing. Following the use of 3- or 5-year citation windows to track research impact (Aksnes et al., 2019), we split the data with two strategies for robust check:

- Strategy 1: Articles published in the most recent 5 years as the testing set and the remaining “old” data for training.
- Strategy 2: Splitting the data with the threshold of the most recent 3 years.

2. Baselines and measures

The key contribution of the proposed method is to develop a link prediction-based approach to recommend knowledge trajectories for target researchers. Thus, we compared the proposed method with total 10 baselines in link prediction, recommender systems, and upstream graph representation learning.

Our method was built on link prediction, and thus we selected the most mainstream link prediction baselines, covering traditional models and some recent developments:

- Jaccard Coefficient (JC): A common neighbor (CN)-based algorithm that calculates the proportion of common neighbors between two unlinked nodes.
- Adamic-Adar Index (AA): A CN-based algorithm that assigns more weights to common neighbors with smaller degrees (Adamic & Adar, 2003).
- Preferential Attachment (PA): An algorithm assuming that the more connected a node is, the more likely it is to receive new links (Newman, 2001).
- Resource Allocation (RA): A CN-based algorithm that allocates resources according to the degree of their CNs (Zhou et al., 2009).
- Weighted Resource Allocation (WRA): A refined RA algorithm that uses a weighted index to involve edge weights (Zhang, Wu, Miao, et al., 2021b).
- Semantic Diffusion (SD): To examine whether the drawback of embedding in local text data (Shang et al., 2020) exists in our local bibliometric dataset, we followed the general process of the proposed method

but constructed a semantic layer to replace the co-topic layer. We generated node vectors using word embedding (Mikolov et al., 2013) and then measured their semantic similarities.

Considering the overlaps between link prediction and recommender systems, we specifically chose two typical recommender system baselines:

- Content-based (*Content*): Recommending topics similar to an author’s current foci.
- Collaborative filtering (CF): Recommending the topics of co-authors.

Since machine learning, particularly deep learning, has been widely applied to either link prediction or recommender systems, we selected two upstream machine learning baselines using state-of-the-art graph representation learning techniques:

- *Node2Vec*: Considering the bi-layer network as a homogeneous graph, we represented nodes via node embedding (Grover & Leskovec, 2016) and trained a Support Vector Machine (SVM)-based model to predict the possible connections between researchers and their unconnected topics.
- Heterogeneous graph neural network (*HetGNN*): Considering the bi-layer network as a heterogenous graph with extra features, for example, papers and venues, we transformed the graph into low-dimensional embeddings using HetGNN (Zhang et al., 2019). Then, similar to Node2Vec, an SVM-based model was applied for predicting researcher-topic connections.

In terms of validation measures, we exploited three measures as follows:

- Receiver operating characteristics (ROC) and area under the curve (*AUC*).
- *Precision*: Given a test set with N edges that only exist in this test set, we measured the proportion of these edges appearing in the top N prediction list.
- *Top k hits*: Given a relatively small k , we measured the proportion of edges correctly predicted in the top k prediction list.

4 | RESULTS

4.1 | Data description and pre-processing

The DBLP database¹ is well known for covering research articles published in major computer science (CS) journals

and proceedings, highlighting the CS community's specific recognition in high-quality journals and reputable conferences. With the open data platform AMiner (Tang et al., 2008), we collected 4,894,081 articles indexed by DBLP on April 9, 2020, and before—that is, the DBLP-Citation-network v12.

We chose AMiner since its released data have already been pre-processed and stored in knowledge graphs. Specifically: (1) AMiner has worked on author name disambiguation for years and achieved appealing accomplishments (Tang et al., 2011). We directly used their disambiguated names, and retrieved 4,398,138 distinctive authors identified in the collected dataset; and (2) DBLP articles are linked to MAG's topic tags, called the field of study (FoS). The FoS tags were created by hierarchical topic modeling (Shen et al., 2018), with each article containing one or more FoS tags. We directly translated these well-recognized FoS tags as topics, identifying 89,504 distinctive topics. On average, each topic was mentioned in around 54.68 papers to indicate the topic scale.

In addition to the entire DBLP dataset, we retrieved 11,399 articles on the information science (IS) disciplines from nine representative IS journals, defined by Hou et al. (2018)—that is, *JASIST*, *Information Processing & Management*, *Journal of Informetrics*, *Information Research*, *Library & Information Science Research*, *Scientometrics*, *Research Evaluation*, *Journal of Documentation*, and *Journal of Information Science*. This sub-dataset contained 14,521 distinctive authors and 7,028 FoS tags, and became our “local” dataset.

Two sets of experiments were designed to examine the performance of the proposed method in diverse data scenarios:

- *Experiment I*: Local dataset (the sub-dataset for the information science disciplines)—It contains a controllable number of articles with relatively high coupling but not-too-narrow topics, and a general preference for research collaboration.
- *Experiment II*: Global dataset² (the DBLP dataset)—As a large-scale dataset covering distinct research topics, the DBLP dataset spans seven of the Web of Science research areas³: artificial intelligence, cybernetics, information systems, software engineering, theory and methods, hardware and architecture, and interdisciplinary applications.

For each experiment, we split the data with two strategies: Articles published in 2015/2018 and before as the training sets, and articles after 2015/2018 as the testing sets. We used authors, FoS tags, and their co-occurrences to build up the co-authorship layer and the co-topic layer, as well as edges connecting authors and topics. With these steps completed, we constructed two bi-layer

networks, one for the training purpose and the other for testing. The statistical information of the two experiments is given in Table 1.

4.2 | Experimental results

We applied the diffusion-based network analytics to Experiments I and II, that is, conducting heterogeneous network analytics, and scoring and ranking candidate edges (e.g., >1 million in Experiment I and ~30 billion in Experiment II, see Table 1). Considering the data scale, we practised two different strategies: For Experiment I, we conducted a full-set validation by ranking all candidate edges; but for Experiment II, we practised two sampling strategies:

- Experiment II (a)—random sampling: We randomly selected 25,000 positive edges and 25,000 negative edges, and composed a testing set.
- Experiment II (b)—distribution-retained sampling: Following the original distribution of the positive and negative edges in the global dataset, we randomly selected 1% common authors and 1% common topics to compose a test set. Generally, the 2015 set contains 2.7 million candidate edges, and the 2018 one has 1.3 million.

Sampling-based validation strategies have been widely applied in network analytics and graph learning (Grover & Leskovec, 2016; Zhang et al., 2019). Our sampling strategies followed the validation schemes prevalently used by Zhou et al. (2009). However, compared to randomly selected edges from the entire network, we sampled candidate edges from the test set—the time window between the training and test sets may reflect the actual innovation process. More importantly, aiming to examine the robustness, for each sampling strategy in Experiment II, we practised 10-fold cross-validations and measured the performance via the average values of the three measures. Table 2 presents the validation results for Experiments I and II, crossing two data-splitting strategies and the two sampling strategies of Experiment II. Besides that, we draw the ROC curves and AUC values of the proposed method and the 10 baselines in all experiments; see Figures⁴ in the Supporting Information.

Our method demonstrates recognizable advantages across the two experiments and the three measures, compared to the 10 baselines. We made the following interpretations:

- The prior and consistent performance of the proposed method in both local and global datasets, both data

TABLE 1 Statistical information of Experiments I and II.

	Experiment I				Experiment II			
	Training set		Test set		Training set		Test set	
	# Nodes ¹	# Edges	# Nodes	# Edges	# Nodes	# Edges	# Nodes	# Edges
2015								
Co-authorship layer	11,836	15,510	3,415	5,525	3,173,445	9,089,406	1,692,287	5,746,287
Co-topic layer	6,497	122,531	2,348	24,226	83,563	13,493,950	65,904	6,757,798
E_{at}	18,333	137,779	5,763	30,329	3,257,008	56,547,787	1,758,191	26,175,654
# Papers	9,908 (86.9%)		1,491 (13.1%)		3,610,096 (73.8%)		1,283,985 (26.2%)	
# Common authors/common topics/ possible edge ²	730/1,817/1,326,410				467,594/59,960/28,036,936,240			
# Existing edges (training set) ³	14,184				23,158,903			
# Positive edges ⁴	4,955				8,982,483			
# Negative edges ⁵	1,307,271				28,004,794,854			
# Candidate edges ⁶	1,312,226				28,013,777,337			
2018								
Co-authorship layer	13,612	18,823	1,281	2,046	4,067,201	12,797,994	658,528	18,636,98
Co-topic layer	6,885	133,313	1,019	7,256	88,290	15,669,947	41,826	2,365,718
E_{at}	20,497	157,167	2,300	9,229	4,155,491	73,125,911	700,354	7,340,919
# Papers	10,948 (96.0%)		451 (4.0%)		4,578,978 (93.6%)		315,103 (6.4%)	
# Common authors/common topics/ possible edges	372/876/325,872				327,591/40,609/13,303,142,919			
# Existing edges (training set)	6,366				20,794,415			
# Positive edges	1,569				2,439,640			
# Negative edges	317,937				13,279,908,864			
# Candidate edges	319,506				13,282,348,504			

Note: (a) # represents the number of related items. (b) Common items are items appearing in training and testing sets, and possible edges represent the maximum number of edges that can appear between common authors and topics. (c) The subset of possible edges that exist in the training set. (d) The subset of possible edges that exist in the testing set but not in the training set. (e) The subset of possible edges that exist in neither the test set nor the training set. (f) The subset of possible edges that do not exist in the training set. Candidate edges = Possible edges – Existing edges.

splitting strategies, and both sampling strategies, demonstrates its reliability and robustness. Particularly, the method is superior in the top 100 hits, highlighting its accuracy in top-ranked topics and reflecting its usefulness in actual recommendations—since top-ranked items are easier to get user attention, their accuracy could be more practically crucial than that of the overall recommendation.

- Resource allocation-based models (e.g., the proposed method, RA, and WRA) can achieve relatively favorite performance, since they algorithmically highlight the role of meta paths in capturing knowledge diffusion in exploratory innovation. In contrast, simply relying on semantic similarities (e.g., Content and SD), co-authorships (e.g., CF), and common neighbors (e.g., JC and AA), and sloppily embedding

heterogeneous features (e.g., Node2Vec and HetGNN) are insufficient.

- SD's unpreferred performance has been expected, demonstrating the drawback of embedding techniques in a local bibliometric dataset. Compared to a large sparse co-topic layer, a well-connected semantic layer might introduce much noise.

In general, the full-set validation of Experiment I illustrates the remarkable advantage of our method in recommending knowledge trajectories in a local dataset, highlighting the nature of interdisciplinary interactions in exploratory innovation. Despite a sampling-based validation, the results of Experiment II can still statistically demonstrate its prior performance on a large-scale global dataset.

TABLE 2 Validation results for Experiments I and II.

Method	Experiment I			Experiment II (a)—10-fold			Experiment II (b)—10-fold		
	AUC	Precision	Top k hit2	AUC	Precision	Top k hit	AUC	Precision	Top k hit
2015									
CF ³	0.6997	0.0327	0.120	0.9147	0.8590	1.000	N/A	N/A	N/A
Content	0.7191	0.0307	0.030	0.9000	0.8339	0.999	N/A	N/A	N/A
AA	0.7385	0.0379	0.080	0.9274	0.8683	1.000	0.9352	0.0428	0.070
JC	0.3824	0.0024	0.000	0.4841	0.4765	0.864	0.4278	0.0020	0.003
PA	0.7301	0.0575	0.140	0.9223	0.8564	0.996	0.9291	0.0613	0.119
RA	0.7536	0.0555	0.100	0.9542	0.8928	0.996	0.9585	0.0840	0.156
WRA	0.7688	0.0823	0.300	0.9653	0.9127	1.000	0.9678	0.1106	0.234
SD	0.5981	0.0508	0.000	N/A	N/A	N/A	N/A	N/A	N/A
Node2Vec	0.6336	0.0823	0.300	0.8493	0.9128	1.000	0.8354	0.1107	0.233
HetGNN	0.7175	0.0823	0.300	0.8280	0.9128	1.000	0.8350	0.1107	0.233
Diffusion	0.7740	0.1009	0.530	0.9704	0.9139	1.000	0.9685	0.1166	0.279
2018									
CF	0.6350	0.0255	0.050	0.8981	0.8299	1.000	N/A	N/A	N/A
Content	0.6439	0.0198	0.000	0.8690	0.8010	1.000	N/A	N/A	N/A
AA	0.6706	0.0242	0.040	0.9028	0.8259	1.000	0.8862	0.0141	0.020
JC	0.4340	0.0064	0.020	0.4247	0.4249	0.945	0.3459	0.0012	0.002
PA	0.6471	0.0529	0.160	0.9078	0.8367	1.000	0.9016	0.0277	0.031
RA	0.6862	0.0389	0.050	0.9445	0.8766	1.000	0.9384	0.0402	0.051
WRA	0.6964	0.0637	0.210	0.9621	0.9055	1.000	0.9560	0.0533	0.059
SD	0.6537	0.0742	0.350	N/A	N/A	N/A	N/A	N/A	N/A
Node2Vec	0.6110	0.0637	0.210	0.8432	0.9057	1.000	0.8203	0.0535	0.059
HetGNN	0.6965	0.0637	0.210	0.8502	0.9057	1.000	0.8357	0.0535	0.059
Diffusion	0.7009	0.0784	0.370	0.9650	0.9062	1.000	0.9627	0.1058	0.151

Note: (a) We ran each sampling experiment 10 times and recorded the average values of the measures. (b) $K = 100$. (3) We skipped off CF and Content in Experiment II(b) and SD in Experiment II due to the extremely high computational cost and its relatively unappealing performance in Experiment I.

4.3 | Algorithm complexity analysis

Computational efficiency is not a key pursuance of most information studies, as well as ours, but this complexity analysis demonstrates the balanced performance of our method in seeking the trade-offs between effectiveness and efficiency. The experiments were performed on a high-performance computing server: Intel Xeon Gold 6238R 2.2GHz 28cores (26 cores enabled); 38.5 MB L3 Cache (Max Turbo Freq. 4.0GHz, Min 3.0GHz); 180GB RAM (Six Channels). Table 3 reports the experimental time of our method and the 10 baselines in Experiments I and II (a).

Generally, the proposed method is not the most efficient one among the baselines, but it is far more efficient than the two traditional recommendation baselines and the three baselines with neural networks (i.e., Node2Vec,

HetGNN, and SD). Considering the trade-off between efficiency and performance (Table 2), our method is superior to the baselines.

One interesting observation reveals that our method calls data beyond the sample, and algorithmically, it relies on the topological structure of the network (i.e., reaching nodes and edges out of the sample data through the diffusion process), which in some sense endorses the reliability of our sampling-based validation. Specifically, compared to the 1.3 million candidate edges in Experiment I, Experiment II(a) uses 50,000 candidate edges (i.e., 25,000 positive edges and 25,000 negative edges). So, in terms of data scale, Experiment II(a) is smaller than Experiment I, and thus most baselines require less time in Experiment II(a). In contrast, our method requires a much longer experimental time, indicating its complexity is related to the network structure rather than the sample scale.

TABLE 3 Complexity analysis.

Method	Experiment I Time (s)	Experiment II (a) Time (s)
CF	36,743	28,562
Content	36,461	3,332
JC	71	270
AA	11	12
PA	11	1
RA	128	12
Node2Vec	28,746 (T) + 65 (I) ²	72,368 (T) + 15 (I)
HetGNN	3,548 (T) + 80 (I)	129,653 (T) + 21 (I)
WRA	132	13
SD	21,658 (T) + 135 (I)	
Diffusion	126	1,098

Note: For the same reason in Table 2, we skipped off SD in Experiment II(a). Abbreviations: I, inference time; T, training time.

4.4 | Case study: Recommending knowledge trajectories for information scientists

To further verify the practical merits of the method, we conducted a case study with the local dataset. Our goal was to chart the knowledge trajectories of information scientists in the dataset. The results of the study not only showcase the qualitative substance of the proposed method, but also demonstrate a feasible way to support decision-making for individual researchers, policy-makers, and entrepreneurs.

Unlike Experiment I, in the case study, we used the entire local dataset to construct the bi-layer network, which included 14,521 authors and 7,028 FoS tags. We recommended knowledge trajectories for these 14,521 authors to help them step beyond their comfort zones. The statistical information of the bi-layer network is given in Table 4.

As noted, MAG's FoS tags were based on hierarchical topic modeling (Shen et al., 2018). One example regarding this topical hierarchy is “computer science (Level 1)—library science (Level 1.1)—scientific communications (Level 1.1.1)”. See a sample of FoS' topical hierarchy in Figure 2. In this case study, we fully facilitated this hierarchy and used the following strategy to recommend knowledge trajectories supporting target researchers to step beyond their comfort zones: Given a target researcher has already published articles on Topic 1.1, we marked the sub-branch of Topic 1.1, including Topics 1, 1.1, and 1.1.1, as the comfort zone topics (Green). We recommended two types of topics beyond the comfort zone:

TABLE 4 Statistical information of the bi-layer networks (case study).

	# Node	# Edge
Co-authorship layer	14,521	20,704
Co-topic layer	7,028	137,088
E_{at}	21,549	165,222
Total	21,549	536,299

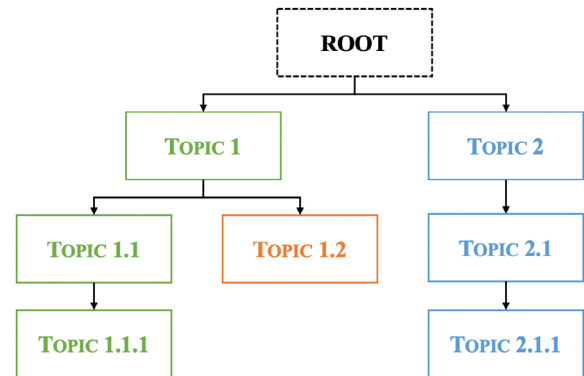


FIGURE 2 Sample of the topical hierarchy of FoS tags.

- *Neighbor topics* (Orange)—topics in the same branch but different sub-branches, which may refer to inter-disciplinary exploration.
- *Outsider topics* (Blue)—topics in other branches, which may represent new knowledge beyond established knowledge base.

The complete list of this recommendation can be found in Table S1, including the top 100 neighbor topics and the top 100 outsider topics for the 14,521 information scientists. We also traced the diffusion process and leveraged the mediators (i.e., collaborators who have ever been involved in this topic and may offer helps, and similar topics which share certain common features with the recommended topic) in the meta paths to interpret recommendation results, see Table S2a,b.

Aiming to further showcase the performance of our recommendation, we specifically chose the recommendation lists and their explanations for Dr Ying Ding (Figure 3) and Dr Alan L. Porter (Figure 4), and visualized them in a hierarchical tree—neighbor topics are in orange and blue nodes represent outsider topics; items after the recommended topics and linked with dash lines are related mediators, interpreting why the model recommends. The topical hierarchy (i.e., topic levels and their upstream-downstream relationships) strictly follows MAG's FoS system.

We strategically chose the two showcases—both researchers are involved in inter-/cross-disciplinary

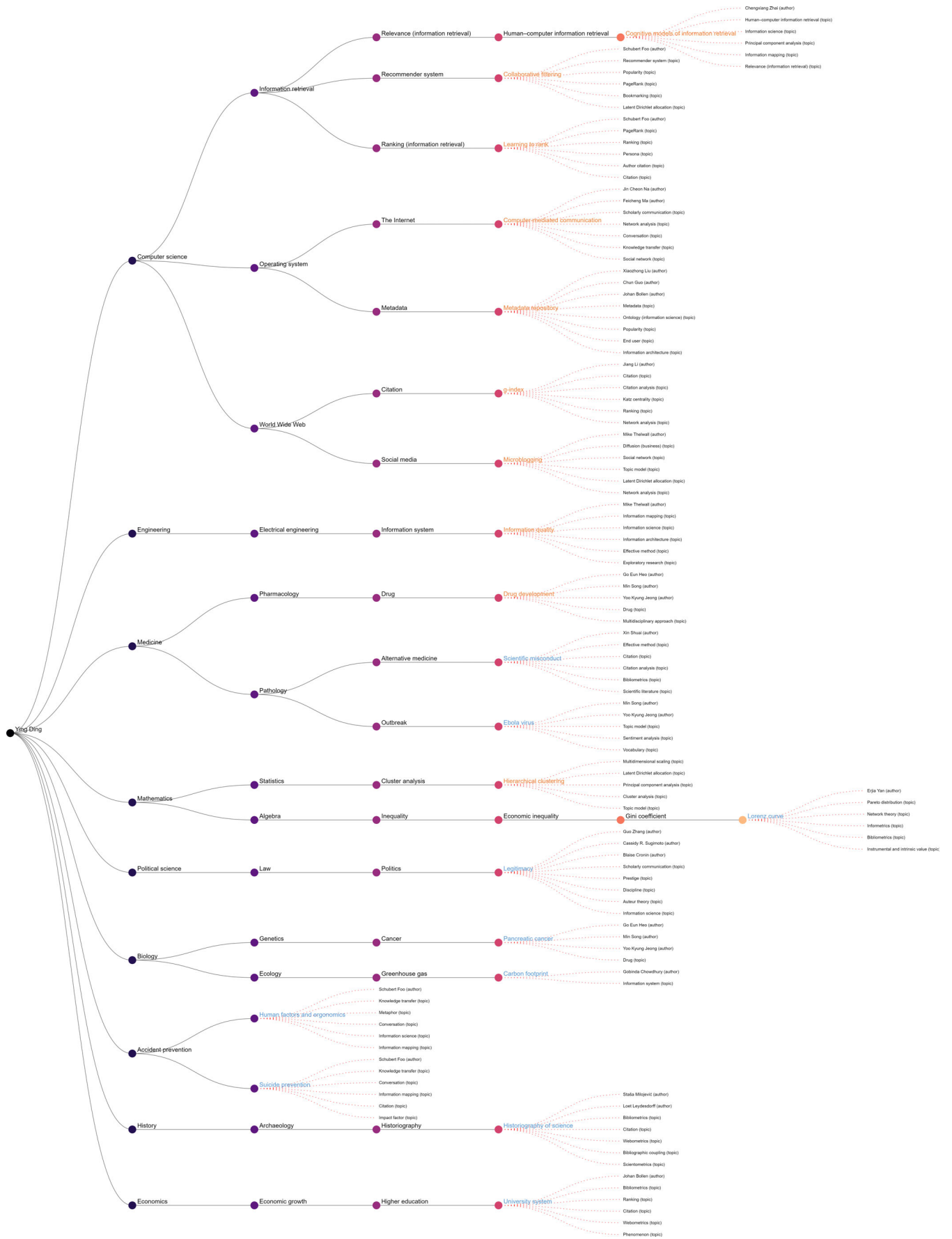


FIGURE 3 Showcase 1—Recommendations for Dr Ying Ding.

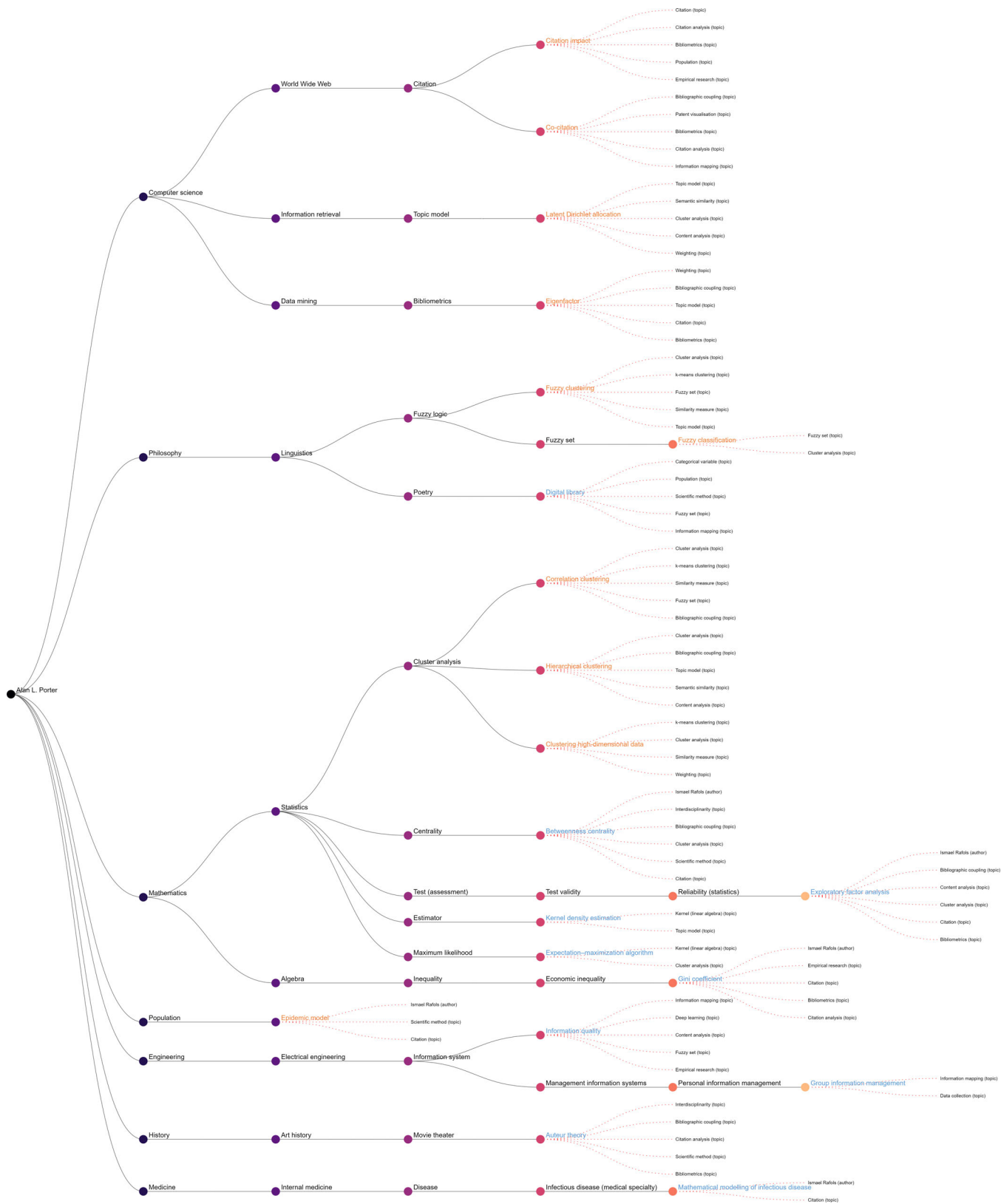


FIGURE 4 Showcase 2—Recommendations for Dr Alan L. Porter.

studies, but with diverse emphases, for example, Dr Porter highlights ST&I studies using bibliometrics and text mining techniques, and Dr Ding introduces data

analytical models from computer science disciplines for methodological development in the science of science studies. Interestingly, in Figure 3, the recommended

neighbor topics coincide with Dr Ding's close tie with data science, particularly information retrieval and semantic web. Figure 4 recommends upstream data analytical models and measures to Dr Porter, in line with his application-driven research foci. Significantly, we collected some inspiring feedback on the two showcases, in terms of the practical use of the proposed method:

- Different understandings on exploratory innovation: Application-driven researchers may hold interest on developing a universal analytical framework for broad cases and thus prefer exploring cross-disciplinary applications with outsider topics. In contrast, methodology-driven researchers are attracted by handleable models connecting with their current foci, that is, interdisciplinary innovation with neighbor topics.
- System interactivity: Current interpretation may provide hints on who may have knowledge on the topic and can offer helps, but annotations to explain "what the topic is" will help understand which kind of specific problems this technique can handle and how target researchers can adopt this novel tool to their practical uses.

5 | DISCUSSION AND CONCLUSIONS

Following the assumptions on exploratory innovation and knowledge diffusion, this study developed a method of diffusion-based network analytics for recommending knowledge trajectories. We constructed a heterogeneous bibliometric network consisting of a co-authorship layer and a co-topic layer, analyzed the process of knowledge diffusion between authors and research topics, and recommended personalized topics for target authors, which lie out of their current research foci and could be their future knowledge trajectories to help step beyond their comfort zones.

5.1 | Technical and practical implications

Highlighted as a method of heterogeneous link prediction with a novel diffusion process, this study brings several technical contributions and practical implications to the literature:

Non-parametric and explainable recommendations: This method contains no super parameters requiring human intervention or extra experiments. More significantly, it inherits the benefits of heterogeneous network mining, which transparently analyses topological

structures with pre-defined meta paths and achieves explainable recommendations through mediators in the diffusion process.

A diffusion process among heterogeneous and homogeneous nodes in a bibliometric network: This method designs a diffusion process to reflect real-world social and knowledge interactions through pre-defined diffusion strategies with heterogeneous nodes (e.g., author-term) and homogeneous nodes (e.g., author-author and term-term). A heterogeneous link prediction with novel diffusion strategies is then developed, which is new to the literature.

Comfort-zone topics: Compared to emerging topics, comfort-zone topics might not be new and have existed for decades. However, the key point is they have never been studied by the target researcher, and may enlighten the recombination with their existing knowledge base and eventually achieve exploratory innovation. Practically, different understandings of exploratory innovation (e.g., application-driven vs. method-driven) may lead to various preferences, and despite explanations provided by the topic hierarchy and the diffusion mediators, further interpretations on what the recommended topic is and how to use it could be value-added.

5.2 | Limitations and future directions

As with all studies, ours has limitations that provide opportunities for future research. (a) As a key drawback of heterogeneous network mining, the efficiency of the proposed method could be critical in a large-scale bibliometric network (see the complexity analysis in Table 3). Despite its consistent prior performance in different data splitting strategies and sampling strategies, using distributed systems and parallel computing techniques to refine our algorithm could be among future directions. (b) Despite acceptable reasons for using the DBLP data as a global dataset, applying the proposed method to some well-recognized global datasets (e.g., MAG, Web of Science, and Scopus) may create practical significance, such as understanding multidisciplinary interactions. (c) One interesting follow-up direction to this study includes introducing dynamic network analytics to capture the cumulative changes over time when knowledge trajectories emphasize the dynamic process of scientific and technological evolution. (d) A function to recognize and evaluate emerging/trendy topics will add extra practical significance.

ACKNOWLEDGEMENTS

This work is supported by the Australian Research Council under Discovery Early Career Researcher Award

DE190100994. Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

ORCID

Yi Zhang  <https://orcid.org/0000-0002-7731-0301>

Mengjia Wu  <https://orcid.org/0000-0003-3956-7808>

ENDNOTES

¹ <https://dblp.org/>

² Although the DBLP data focuses on computer science and may not be a typical global dataset, we argue that the rapid development of information technology over past decades has led to active disciplinary interactions that cross relatively broad and diverse research areas.

³ <https://incites.help.clarivate.com/Content/Research-Areas/wos-research-areas.htm>

⁴ Two figures for Experiment I with the two data-splitting strategies; Forty figures for Experiment II with 10 folders, two data-splitting strategies, and two sampling strategies.

REFERENCES

- Acar, O. A., Tarakci, M., & Van Knippenberg, D. (2019). Creativity and innovation under constraints: A cross-disciplinary integrative review. *Journal of Management*, 45(1), 96–121.
- Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211–230.
- Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1), 2158244019829575.
- Alhoori, H., & Furuta, R. (2017). Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics*, 11(2), 553–563.
- Barley, W. C., Treem, J. W., & Kuhn, T. (2018). Valuing multiple trajectories of knowledge: A critical review and agenda for knowledge management research. *Academy of Management Annals*, 12(1), 278–317.
- Björneborn, L. (2004). *Small-world link structures across an academic web space: A library and information science approach* (PhD thesis), Royal School of Library and Information Science. Citeseer.
- Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive innovation: An intellectual history and directions for future research. *Journal of Management Studies*, 55(7), 1043–1078.
- Dong, Y., Hu, Z., Wang, K., Sun, Y., & Tang, J. (2020). *Heterogeneous network representation learning* (pp. 4861–4867). IJCAI.
- Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy*, 11(3), 147–162.
- Érdi, P., Makóvi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zálányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics*, 95(1), 225–242. <https://doi.org/10.1007/s11192-012-0796-4>
- Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5), 875–908.
- Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 855–864.
- He, C., Wu, J., & Zhang, Q. (2022). Proximity-aware research leadership recommendation in research collaboration via deep neural networks. *Journal of the Association for Information Science and Technology*, 73(1), 70–89.
- Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics*, 115(2), 869–892.
- Huang, L., Liu, F., & Zhang, Y. (2020). Overlapping community discovery for identifying key research themes. *IEEE Transactions on Engineering Management*, 68(5), 1321–1333.
- Huang, S., Lu, W., Bu, Y., & Huang, Y. (2022). Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing & Management*, 59(6), 103110.
- Jansen, J. J., Van Den Bosch, F. A., & Volberda, H. W. (2006). Exploratory innovation, exploitative innovation, and performance: Effects of organisational antecedents and environmental moderators. *Management Science*, 52(11), 1661–1674.
- Jeong, Y., Jang, H., & Yoon, B. (2021). Developing a risk-adaptive technology roadmap using a Bayesian network and topic modeling under deep uncertainty. *Scientometrics*, 126(5), 3697–3722.
- Keshavarz, H., & Shekari, M. R. (2020). Factors affecting topic selection for theses and dissertations in library and information science: A national scale study. *Library & Information Science Research*, 42(4), 101052.
- Kleśniński, R., Kazienko, P., & Kajdanowicz, T. (2021). Where should I publish? Heterogeneous, networks-based prediction of paper's citation success. *Journal of Informetrics*, 15(3), 101200.
- Leydesdorff, L., & Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology*, 62(5), 846–860.
- Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9), 1515–1530.
- Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019–1031.
- Liu, L., Wang, Y., Sinatra, R., Giles, C. L., Song, C., & Wang, D. (2018a). Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559(7714), 396–399.
- Liu, Z., Xie, X., & Chen, L. (2018b). Context-aware academic collaborator recommendation. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1870–1879.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 3111–3119.
- Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2), 025102.
- Nicolini, D., Mengis, J., & Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. *Organisation Science*, 23(3), 612–629.

- Ou, Q., Jin, Y.-D., Zhou, T., Wang, B.-H., & Yin, B.-Q. (2007). Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Physical Review E*, *75*(2), 021102.
- Phelps, C. C. (2010). A longitudinal study of the influence of alliance network structure and composition on firm exploratory innovation. *Academy of Management Journal*, *53*(4), 890–913.
- Rahdari, B., Brusilovsky, P., Babichenko, D., Littleton, E. B., Patel, R., Fawcett, J., & Blum, Z. (2020). Grapevine: A profile-based exploratory search and recommendation system for finding research advisors. *Proceedings of the Association for Information Science and Technology*, *57*(1), e271.
- Ren, H., & Zhao, Y. (2021). Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technovation*, *101*, 102196.
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.
- Shang, J., Zhang, X., Liu, L., Li, S., & Han, J. (2020). Nettato: Automated topic taxonomy construction from text-rich network. *Proceedings of the Web Conference 2020*, 1908–1919.
- Shen, Z., Ma, H., & Wang, K. (2018). A web-scale system for scientific knowledge exploration. *arXiv*, 1805.12216.
- Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, *51*(5), 756–770.
- Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. *Research Policy*, *33*(10), 1615–1634.
- Steinmo, M., & Rasmussen, E. (2018). The interplay of cognitive and relational social capital dimensions in university-industry collaboration: Overcoming the experience barrier. *Research Policy*, *47*(10), 1964–1974.
- Sun, J., Zhu, M., Jiang, Y., Liu, Y., & Wu, L. (2021). Hierarchical attention model for personalised tag recommendation. *Journal of the Association for Information Science and Technology*, *72*(2), 173–189.
- Sun, X., Kaur, J., Milojević, S., Flammini, A., & Menczer, F. (2013). Social dynamics of science. *Scientific Reports*, *3*(1), 1–6.
- Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *3*(2), 1–159.
- Tang, J., Fong, A. C., Wang, B., & Zhang, J. (2011). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 975–987.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, 990–998.
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472.
- Wang, C., Rodan, S., Fruin, M., & Xu, X. (2014). Knowledge networks, collaboration networks, and exploratory innovation. *Academy of Management Journal*, *57*(2), 484–514.
- Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A co-authorship network analysis. *Journal of the Association for Information Science and Technology*, *60*(10), 2107–2118.
- Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, *8*(2), 295–309.
- Yang, J., & Zhang, X.-D. (2016). Predicting missing links in complex networks based on common neighbors and distance. *Scientific Reports*, *6*(1), 1–10.
- Zanello, G., Fu, X., Mohnen, P., & Ventresca, M. (2016). The creation and diffusion of innovation in developing countries: A systematic literature review. *Journal of Economic Surveys*, *30*(5), 884–912.
- Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 1–11.
- Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 793–803.
- Zhang, J. (2017). Uncovering mechanisms of co-authorship evolution by multi-relations-based link prediction. *Information Processing & Management*, *53*(1), 42–51.
- Zhang, Y., Porter, A. L., Cunningham, S. W., Chiavetta, D., & Newman, N. (2020). Parallel or intersecting lines? Intelligent bibliometrics for investigating the involvement of data science in policy analysis. *IEEE Transactions on Engineering Management*, *68*, 1259–1271.
- Zhang, Y., Wang, X., Huang, L., Zhang, G., & Lu, J. (2018). Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology. *2018 Annual Meeting of the Association for Information Science and Technology*.
- Zhang, Y., Wu, M., Hu, Z., Ward, R., Zhang, X., & Porter, A. (2021a). Profiling and predicting the problem-solving patterns in China's research systems: A methodology of intelligent bibliometrics and empirical insights. *Quantitative Science Studies*, *2*(1), 409–432.
- Zhang, Y., Wu, M., Miao, W., Huang, L., & Lu, J. (2021b). Bi-layer network analytics: A methodology for characterising emerging general-purpose technologies. *Journal of Informetrics*, *15*(4), 101202.
- Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*(4), 623–630.
- Zhou, T., Ren, J., Medo, M., & Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, *76*(4), 046115.
- Zhou, X., Huang, L., Zhang, Y., & Yu, M. (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. *Scientometrics*, *121*(2), 699–737.
- Zhu, Y., Quan, L., Chen, P. Y., Kim, M. C., & Che, C. (2022). Predicting co-authorship using bibliographic network embedding. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.24711>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang, Y., Wu, M., Zhang, G., & Lu, J. (2023). Stepping beyond your comfort zone: Diffusion-based network analytics for knowledge trajectory recommendation. *Journal of the Association for Information Science and Technology*, *74*(7), 775–790. <https://doi.org/10.1002/asi.24754>