Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies

Yi Zhang[1], Mengjia Wu[1], Wen Miao[2], Lu Huang[2,*], Jie Lu[1]

[1] yi.zhang@uts.edu.au; mengjia.wu@student.uts.edu.au; jie.lu@uts.edu.au
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[2] huanglu628@163.com *(Corresponding Email)*
School of Management and Economics, Beijing Institute of Technology, China

Abstract

Despite the tremendous contributions bibliometrics has made to profiling technological landscapes and identifying emerging topics, reliable methods for predicting potential technological changes are still elusive. To fill this gap, we propose a methodology based on bi-layer network analytics that characterizes emerging general-purpose technologies. The framework incorporates three novel indicators that quantify a technology's technical potential and social impacts, not just in one specific technological area but in a wide range of domains. Missing links in the network are extrapolated through a refined link prediction method, and a weighted resource allocation index ranks both current technologies and their predicted evolutions to reveal candidate innovations for further empirical and/or expert analysis. A case study on information science incorporating quanlitative and qualitative validations demonstrates the methodology to be feasible and reliable. Researchers and policymakers in information science and bibliometrics should find valuable decision support from the empirical insights presented.

Keywords: bibliometrics; network analytics; emerging technologies; general-purpose technologies; information science.

1 Introduction

Theoretical definitions of general-purpose technologies (GPTs) can be traced back to Paul David, who coined the idea from his observations of the widespread impact of electric dynamos on the productivity of the United States (US) during the 1920s (David, 1989). In the decade to follow, a well-recognized conception of what constitutes a GPT took hold – "pervasiveness, inherent potential for technical improvements, and innovational complementarities" (Bresnahan and Trajtenberg (1995) – and measuring the many and various aspects of GPTs became a topic of increasing interest for many economists. One indicator in particular, *generality*, has been widely applied (Hall & Trajtenberg, 2004), and its standard calculations consider patent citations and their technological classes. In the 2000s, however, the attention of the science, technology, and innovation (ST&I) community turned to the unique features of nanotechnologies, triggering discussions on the notion of *emerging* GPTs, i.e., EGPTs (Graham & Iacopetta, 2009; Youtie et al., 2008). Although much research has been undertaken to define what classifies a technology as emergent, characterizing the "general purpose" component of an EGPT has proven to be far more difficult. Of the few studies specific to EGPTs, all assume emergence before testing constructs like generality, e.g., Schultz & Joutz (2010). Conversely, of the studies that primarily explore emergence, Rotolo et al.'s (2015) systematic review defined five attributes of emerging technologies, theoretically guiding the development of further measurements. However, to our best knowledge, very few study has attempted to directly measure and classify emergence and generality at the same time. This, coupled with the lack of a quantitative measure for generality that does not solely depend on patents and citations, inspired us to establish a cohesive system of quantitative measures for identifying EGPTs that detects both emergence and generality.

Sharing a close interest with ST&I studies, bibliometrics is well recognized as a tool for supporting technology analysis and assessment. For example, it has been used to profile various technological areas (Chakraborty et al., 2015; Guo et al., 2010), identify emerging topics in science and technology (Glänzel & Thijs, 2012; Small et al., 2014), and track the pathways of technological change (Hou et al., 2018; Zhang et al., 2016; Zhou et al., 2014). More recently, the use of advanced data analytic technologies, such as topic models, streaming data analytics, and machine learning, have massively increased the amount of data traditional bibliometrics methods can process (Ding & Chen, 2014; Klavans & Boyack, 2017). They have also brought the ability to reveal hidden relationships (Zhang, Lu, et al., 2018; Zhang, Zhang, et al., 2017) and visualize complicated technological portfolios and innovation networks in highly interpretable ways (Börner et al., 2012; Suominen & Toivanen, 2016).

From a technical point of view, even though network analytics has long been a mainstay of social science (Borgatti et al., 2009), it was only introduced to bibliometric studies in the late 2000s. Originally used as a method for investigating research collaborations and the interactions between disciplines through bibliographic couplings (Yan et al., 2009; Yang et al., 2010), it has subsequently been combined with citation analysis to identify emerging topics and evaluate research impacts (Takeda & Kajikawa, 2009; Yan, 2015). Network analytics has also been explored for its ability to predict emerging technologies (Érdi et al., 2013) and to reveal hidden technological opportunities (Park & Yoon, 2018).

Yet, even with these techniques, developing a bibliometric model to identify EGPTs is still highly challenging. First, bibliometric models emphasize the use of historical data and have a natural connection with citation statistics, and thus are friendly for measuring generality. However, rapid developments in natural language processing in recent years are relaxing the field's dependence on patent archives to temper this philosophy of past as prologue, which may reveal insights directly from the semantics of the subject matter. Second, despite keen interest and many pilot studies on measuring and forecasting technical emergence (Carley et al., 2018), current bibliometric models are still falling short of truly "characterizing the potential of what is detected to be emerging" (Rotolo et al., 2015). Balancing generality with emergence to comprehensively characterize EGPTs further increases this challenge. Third, the bibliometric community is recognizing the benefits of link prediction as a way of identifying the likely technologies of tomorrow, but applying those methods to bibliographical information is not yet seamless. For example, theoretically mapping the key attributes of emerging technologies to the topological indicators of a bibliometric network can be problematic. Similarly, integrating heterogeneous bibliographical information into a single network so as to discover social impacts in addition to technological transitions still has issues.

Aiming to address these concerns, we propose a methodology based on bi-layer network analytics to quantitatively identify EGPTs. The methodology begins with the construction of a co-term network (the first layer) and a co-authorship network (the second layer). The two layers are then integrated into a bi-layer network that reflects both the substance of the technologies (i.e., terms) and the social entities (i.e., authors) engaged in their associated R&D. Typically, a traditional bibliometric network only reflects one indicator, e.g., term co-occurrence or co-authorship. The proposed bi-layer network charts both, offering a bibliometric solution that not only reveals the impact of key technologies, but also the authors and collaborative networks that are advancing these technologies. Integrating all this information into one analysis provides a novel perspective from which to draw comprehensive new insights.

To fully leverage this perspective, we adapted the five attributes of emerging technologies defined by Rotolo et al. (2015) into three new indicators capable of quantifying the topological structures in a bi-layer network, namely *fundamentality*, *speciality*, and *sociality*. Interestingly, among Rotolo et al.'s quintet of attributes, *prominent impact* is of particular interest to our endeavors. From their literature review, Rotolo et al. (2015) found that most scholars conceive of prominent impact as a force "exerted on the entire socio-economic system" – a concept, they add, that "comes very close to that of 'general-purpose technologies'". Discontented with the sweeping nature of this definition for the purposes of defining emergence, the authors proposed a more utilitarian version which acknowledges that an emerging technology's impact may be limited to one or a few domains. Thus, the intriguing argument was made that if we can measure prominent impact, we can measure generality as well. In part, this notion inspired the tripartite design of the above indicators.

With the network constructed and the topological structures measured, candidate future innovations are identified with a refined link prediction algorithm, using a weighted index of resource allocation. The algorithm considers the links both within each network layer, i.e., co-term and co-authorship links, as well as between layers, i.e., author-term links. Whether or not a link is predicted is based on the weighted index, which is an amalgamation of frequency statistics, including term co-occurrence, co-authorships, and author-term co-occurrence. Ultimately, the differences between the current network and the predicted network are the key to forecasting technological changes and, of course, which technologies are most likely to be EGPTs in the near future.

A case study on 17,445 articles published in 15 journals and conference proceedings on information science between 1 Jan 1996 and 31 Dec 2018 demonstrate the feasibility and reliability of the method. Additionally, the empirical insights derived from the study should provide decision support to researchers and policymakers in information science disciplines.

The rest of this paper is organized as follows: Section 2 reviews previous studies on bibliometrics for analyzing emerging technologies, network analytics with bibliometric indicators, and theoretical discussion on characterizing EGPTs from a bibliometric perspective. In Section 3, we outline the research framework of the study and introduce the proposed methodology. Section 4 follows, presenting the data, results, validation measurements, and empirical insights derived from the case study. The article concludes in Section 5 with a discussion on the technical and practical implications of our findings, the limitations of the study, and possible future directions of research.

2 Literature review and theoretical background

As a tool for analyzing emerging technologies, bibliometrics has attracted common interest from the bibliometrics and ST&I communities. Further, the rising enthusiasm for social network analysis is solidifying the merit of bibliometrics in both breadth and depth. Therefore, what follows is a review of how bibliometrics has been used to analyze emerging technologies and an overview of network analytics and its bibliometric indicators. Furthermore, we discuss the theories and concepts of GPTs and emerging technologies, establishing a theoretical base for characterizing EGPTs from a bibliometric perspective.

2.1 Bibliometrics for analyzing emerging technologies

Since the 1980s, bibliometrics has played an increasingly active role in evaluating research performance (King, 1987). Beginning with patent analysis, bibliometric techniques have provided new angles for measuring technological change and supporting technology management (Basberg, 1987). Early methodologies from the 1990s and early 2000s included frameworks like: technology opportunity analysis and tech mining (Porter & Cunningham, 2004; Porter & Detampel, 1995); novel formats for investigating issues in technology management, such as science maps (Noyons & Van Raan, 1998) and technology roadmaps (Kostoff & Schaller, 2001); and methods introduced from the area of information retrieval, like topic detection and tracking (Allan, 2002). More recently, bibliometric studies on emerging technologies have coalesced into three broad categories: profiling technological opportunities and landscapes (Chakraborty et al., 2015; Guo et al., 2010); identifying technological topics and their relationships (Glänzel & Thijs, 2012; Small et al., 2014); and tracking the pathways of technological change (Hou et al., 2018; Zhang et al., 2016; Zhou et al., 2014).

Moreover, advances in data science technologies, such as topic models, streaming data analytics, and machine learning, are greatly enhancing traditional bibliometric techniques on a number of fronts. For instance, they have substantially increased the amount of information that can be analyzed (Klavans & Boyack, 2017), provided a means to discover hidden relationships (Zhang, Zhang, et al., 2017), and been used to visualize complicated technological/scientific landscapes and network structures (Börner et al., 2012; Suominen & Toivanen, 2016). These benefits are apparent in recent studies on emerging technologies, where intelligent analytic models are delivering much deeper insights into the dynamics of innovation through advancements like dynamic topic detection and tracking (Ding & Chen, 2014), determining the directions of technological change with semantic analytics (Guo et al., 2016), and investigating technological interactions with learning-enhanced bibliometric approaches (Zhang, Huang, et al., 2019). Zhang et al. (2020) refer to this nexus between data science and information science as "intelligent bibliometrics" and highlight that bibliometric indicators are increasingly being incorporated into intelligent models to support ST&I studies . If not already an established field in its own right, intelligent bibliometrics looms bright in our near future.

2.2 Network analytics with bibliometric indicators

Since the early 2000s, research interest in complex network analysis has rapidly expanded from applied physics to the computer and social sciences (Borgatti et al., 2009; Palla et al., 2005). Now more widely known as social network analysis, network analytics was a relative latecomer to bibliometrics. Initially, network analytics was used as a tool to investigate research collaborations and disciplinary interactions through bibliographic couplings (Yan et al., 2009; Yang et al., 2010). However, once network analytics began to be combined with citation networks, co-citation networks, and co-authorship networks, attention from the bibliometric community increased dramatically. Understanding the topological structures of these networks has provided solutions to many open research topics, such as collaboration and citation patterns (Ding, 2011; Liu et al., 2005). More recently, the introduction of co-word networks and semantic analysis has been providing fresh new angles to discover knowledge structures and identify research domains (Ravikumar et al., 2015; Zhang et al., 2012). Algorithms for community detection, link prediction, random walks, and others are also lending novel tools to increase the scope of traditional techniques and to undertake completely new types of analysis – for example, recommending potential collaborators (Huang, Zhu, et al., 2018; Yan & Guns, 2014), discovering technological opportunities (Park & Yoon, 2018), and detecting/predicting emerging topics and technologies (Érdi et al., 2013; Huang, Jia, et al., 2018).

Yet despite these endeavors to incorporate bibliometric indicators into network analytics, a stubborn focus on homogeneous networks, such as co-word networks, co-citation networks, and co-authorship networks, may mean

we are not gaining the maximum possible benefit for our efforts. As a motivating example for this study, it is reasonable to consider that "who did what" influences the evolution of "what". Hence, identifying and understanding the relationships within "who" should give us greater insights into "what". The bi-layer network proposed in this study is expected to provide such insights by characterizing both the technologies and the entities involved in their development.

2.3 Theoretical discussion on characterizing EGPTs from a bibliometric perspective

Although studies on EGPTs are scarce, there is a rich body of literature on the concepts and indicators surrounding emerging technologies and GPTs. Hence, this subsection begins with a review of the literature on the definition and measurement of GPTs and emerging technologies. These constructs establish the theoretical basis of the study. Building upon these concepts, we move to the definition of EGPTs and a discussion on the three characteristics for measurement from a bibliometric perspective.

2.3.1 Definitions and measurements of GPTs and emerging technologies

Economists first cast their eye on GPTs in the 1990s when attempting to describe a technology's role in economic growth. What emerged from these discussions was a commonly-held conception that GPTs transform daily life and business (Jovanovic & Rousseau, 2005) and various studies on example technologies that exemplify GPTs – steam engine (Rosenberg & Trajtenberg, 2004), electricity (Moser & Nicholas, 2004; Ristuccia & Solomou, 2014), and information technology (Basu & Fernald, 2007).

Conceptually, David (1989) recognized GPTs when exploring the electric dynamo's impact on increasing the US productivity during the 1920s. He observed that the dynamo had spread across a broad range of industries and reflected that the phenomenon was neither unique to the dynamo nor the US. Bresnahan and Trajtenberg (1995) followed this trail, arguing that a GPT's key features are "pervasiveness, inherent potential for technical improvements, and innovational complementarities". These three characteristics were to inform the foundations of almost every definition of GPTs that was to come. Lipsey et al. (1998), for instance, simply paraphrased their definition, highlighting GPT's "scope for improvement", "wide usefulness", and "technological complementarities". Jovanovic and Rousseau (2005) embraced the concept of pervasiveness, noting that GPTs also have long-term positive effects on economic growth and they usually emerge infrequently. Graham and Iacopetta (2009) subsequently explained this infrequency from the perspective of a technology's dissemination process; that is, a new GPT only becomes suitable for wide use when new secondary technologies are developed or adopted to support it.

In tandem with the documenting of theories to define GPTs, interest in measuring GPTs also grew. The backward-looking formulation of generality, which gauges the disseminative patterns in patent citations (Hall & Trajtenberg, 2004), was one that quickly became a baseline for GPT measurement. Youtie et al. (2008) interpreted generality as a GPT's most fundamental feature with the claim that GPTs have a "substantial and pervasive effect across the whole society". From these threads, economists delved into patent statistics, using generality as a basis for developing more diverse indicators for analyzing GPTs (Petralia, 2020). However, in the early 2000s, a new conversation surfaced alongside the rise of nanotechnology. GPTs had attracted the attention of ST&I researchers, but it was unclear as to whether, or even if, emerging technologies like nanotechnology could meet the generally-accepted criteria for being a GPT (Graham & Iacopetta, 2009; Youtie et al., 2008). Thus, the notion of an EGPT was born and several pilot studies were conducted to test the question (Schultz & Joutz, 2010). However, all stand on the assumption that nanotechnology is an emerging technology first, and then draw conclusions as to whether the generality needed to be considered a GPT. In fact, to the best of our knowledge, there is still no unified approach in the literature for concurrently determining whether a technology is both emergent and a GPT. Further, truly capturing the potential of nascent technologies may require indicators of generality that are not so heavily tied to patents. Leveraging the hidden knowledge in scientific papers and other early-stage development documents via text analytics may be a more fruitful line of inquiry.

From their review, Rotolo et al. (2015) settled upon five different attributes to describe emergence: radical novelty, fast growth, coherence, prominent impact, and uncertainty/ambiguity. In the five years since this study, their work has greatly shaped the design and implementation of bibliometric approaches for measuring technological emergence (Carley et al., 2018; Chung & Sohn, 2020). In this vein, prominent impact has particular relevance to

this study. As mentioned, Rotolo et al. (2015) identified some significant overlaps between prominent impact and generality, but contended that, in the context of emerging technologies, the parameters of what constitutes prominent impact must be relaxed from wide-ranging domains and entire socio-economic systems to also include certain narrow scopes in one or a few domains if required. We assert that this argument provides a key bridge between generality and emergence that can be exploited to build a unified framework for identifying EGPTs, as discussed in greater detail next.

2.3.2 Measurable characteristics of EGPTs

In simple terms, measuring an EGPT requires measuring emergence and generality. Would it were that practice was as simple as theory. Taking generality first, the original form of generality proposed by Hall and Trajtenberg (2004) considers the diversity of sector classifications that a patent's citations belong to. Thus, calculating this metric requires two elements: patent citation statistics and a classification key, such as codes from the International Patent Classification (IPC) system. Finding equivalent proxies in other types of bibliometric data is not straightforward. In terms of emergence, work to quantifying Rotolo et al.'s (2015) five attributes of emerging technologies is still in its early stages, especially from a bibliometric perspective. For example, touching on radical novelty, fast growth, and coherence, Carley et al. (2018) proposed a general indicator of technical emergence to identify emergent terms. The idea rests on tracking the most recent dynamics in term frequency and authorships – by recent, the authors mean months or a year or two at most. This strategy inspired us to combine text analytics with the established concepts of GPTs and emerging technologies so as to develop a set of adaptable indicators to identify EGPTs for broad bibliometric data.

Figure 1 depicts the evolution of our three indicators from concepts in the literature. GPTs and generality were the starting point. Rotolo et al.'s connection between generality and prominent impact and the revised scope of impact to include specialized purposes inspired the first split in indicators from some more whole conception into fundamentality and speciality. Fundamentality represents general impact – the extent to which a technology has or could fundamentally impact a wide range of domains, whether they be industries, research areas, or other technologies. Speciality represents the extent to which a technology has been highly successful at the thing(s) it was designed to do, which may include the ability to create relationships between other technological components in the same area. Both could be measured easily from a standard co-term network. Prior bibliometric studies have measured a technology's general impact in knowledge production processes through the topological structures in a citation or co-term networks, such as centrality (Furukawa et al., 2015; Shibata et al., 2009). However, while the principles of centrality apply seemed to apply well to our general conception of fundamentality, the same cannot be said of speciality. Therefore, fundamentality is constructed as a weighted amalgam of three different centrality measures, as outlined in Section 3.2.2.1. But, to measure speciality, we turned to the literature on community, adopting both definitions from network theory and clustering algorithms to design the indicator.

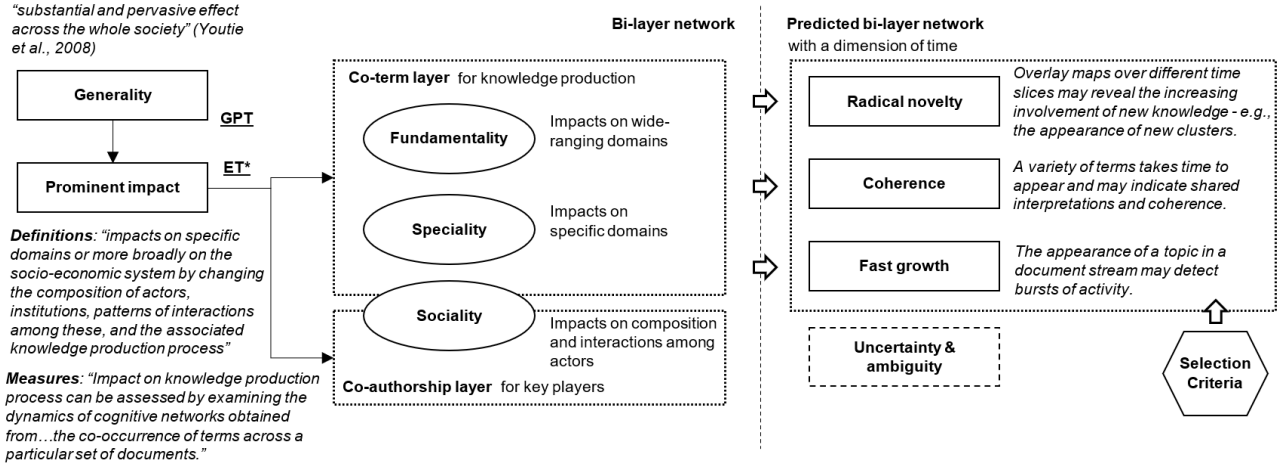


Figure 1. Theoretical background for the three measurable characteristics of EGPTs
ET stands for emerging technology. All non-referenced quotes were taken from Rotolo et al. (2015).

In terms of a technology's social impact, there has been a long history of bibliometricians analyzing co-authorship networks to investigate the role of key players in a field and to identify changes in their collaborative patterns within a knowledge production process (Ding, 2011; Hicks et al., 1986). This gave rise to the third split in prominent impact – sociality. We conceived sociality as a measure of the extent to which a technology could be transferred among key players based on the idea that it is relatively easy for a technology to transfer from a researcher to his/her collaborators and then to create broad socio-economic impacts with an increasing number of stakeholders. Given that this construct traces the interactions between people and technologies, calculating the value for this indicator demands a bi-lateral co-term/co-author network.

These three measurable characteristics establish a comprehensive framework through which to analyze a technology's prominent impact in multiple respects. Combined, they should reveal technologies that fundamentally impact wide-ranging technological domains but also specialize in a given technological area and have the potential to transfer knowledge among key players. That said, conceptually, such a measurement may be biased toward generality since generating impact takes time, which historical data has in abundance. By contrast, recognizing emergence without the benefit of hindsight requires a future-oriented analysis. For this, we turned to network analytics, and, more specifically, link prediction. The idea is to simulate the near future by predicting likely connections within and between entities and terms, and then examine the differences between the current network and the predicted future. Here, three more of Rotolo et al.'s five indicators offered useful criteria for structuring the comparison: radical novelty, fast growth, and coherence.

Taking inspiration from Glänzel and Thijs (2012), we defined three criteria based on changes in a technology's impact scores, i.e., its fundamentality, speciality, and sociality now and as predicted in the future. As part of developing a method to cluster emerging topics, Glänzel and Thijs (2012) devised a set of categories consistent with Rotolo et al.'s indicators. The three categories are: 1) completely new topics (but with their roots in existing topics); 2) topics showing exceptional growth; and 3) shift topics from existing ones. The first category does not apply to our case. However, growth, dissemination, and shifts in rank do. A deeper investigation of the literature revealed the following insights.

- A node appearing in the top ranks of the predicted network only may reflect radical novelty. This conjecture partially coincides with Yan (2015) definition of novelty, which contends that new topics share low variation with existing ones.
- An increase in a node's rank from the current network to a near-future prediction of one would reflect fast growth. This measure is straightforward, with most previous studies detecting either the number of terms/documents involved or the size of a cluster (Carley et al., 2018; Ohniwa et al., 2010).
- A node with high ranks now and in the future may reflect coherence. This measure stands on the arguments in the literature that emphasize the traceable roots of emerging topics from existing knowledge bases (Furukawa et al., 2015; Glänzel & Thijs, 2012).

As a last point, Rotolo et al.'s (2015) fifth indicator of emergence is uncertainty/ambiguity. For several reasons, we decided to skip this attribute. First, Rotolo et al. themselves reported that "the evaluation of uncertainty and ambiguity remains largely unexplored". From our review of the literature, not much has changed in the intervening five years. Second, of the studies that do touch on this area, many are based on prior assumptions about a technology that are beyond the scope of this study. For example, "the diversity of a new domain's terminology may indicate uncertainty" (Lucio & Leydesdorff, 2009), but in our studies such a diversity of terms is unmeasurable since all nodes (i.e., terms) in the predicted network are retained the same as the current network.

The full method and specific construction of each of these indicators follows in the next section.

3 Methodology

An overview of the EGPT framework is given in Figure 2. As illustrated, the methodology involves three key phases: data and pre-processing, bi-layer network analytics, and validation measurements.

3.1 Data and pre-processing

The framework supports a range of bibliometric data, including scientific articles, patents, academic proposals, etc., and two types of bibliographical information: authors and co-authors (and their affiliations); and terms extracted from the titles and abstracts of documents via natural language processing (NLP) techniques.

The process begins by removing noise and consolidating synonyms through a term clumping process that comprises a set of case-driven strategies (Zhang et al., 2014). For example, meaningless single words are removed, low-frequency terms are discarded, words with the same root are consolidated, and so on. A similar process is also applied to the author names. Here, the clumping procedure focuses on issues like disambiguation and standardizing the names into the same "first-name-then-surname" format. The output of this phase consists of a list of terms and their co-occurrence statistics, a list of authors and their co-authorships, and a matrix of the number of times each author has mentioned a term.

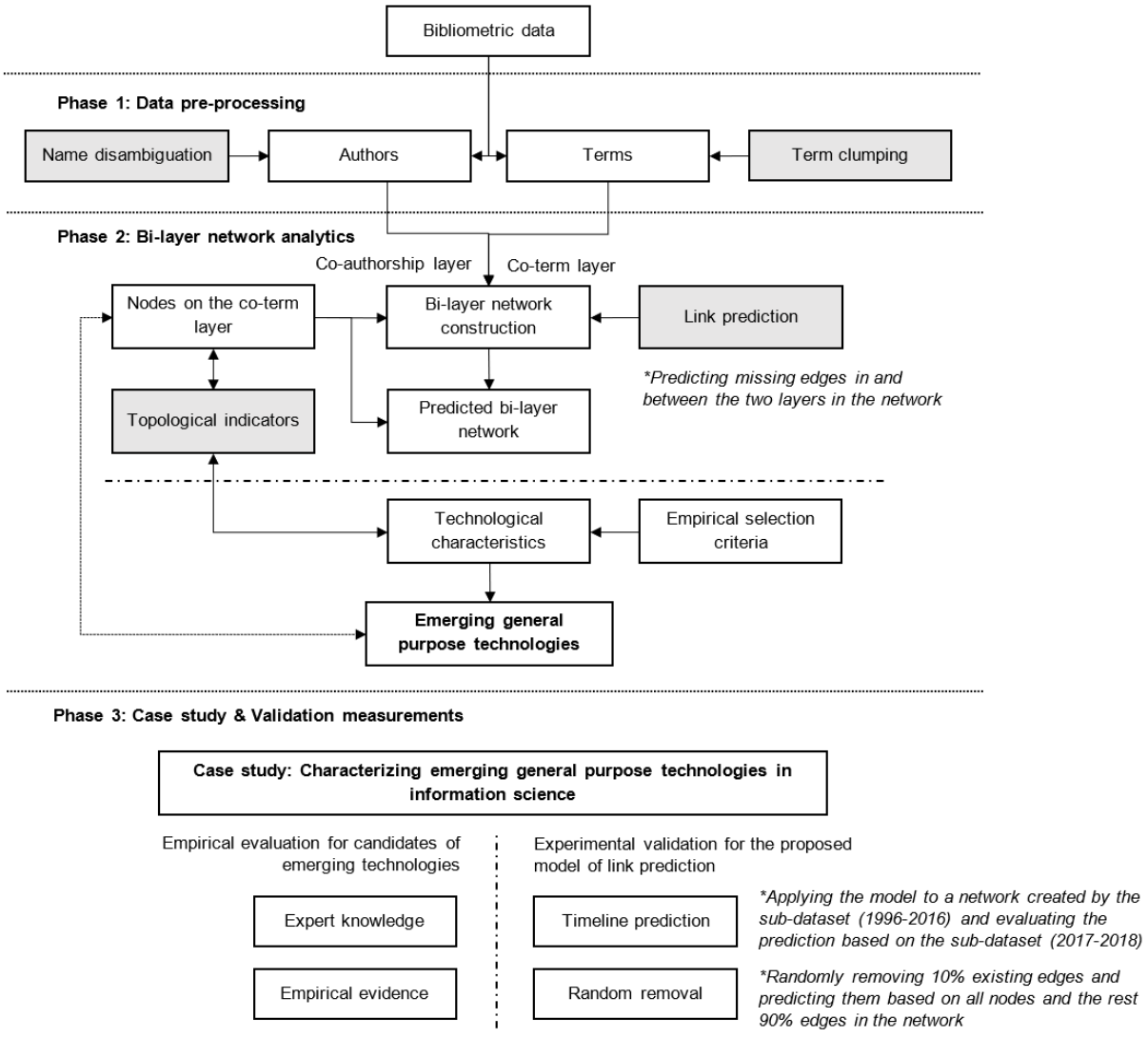


Figure 2. Research framework.

3.2 Bi-layer network analytics

Phase two of the methodology involves constructing a bi-layer network, quantifying the technological characteristics, predicting missing links, and identifying the EGPTs according to the selection criteria.

3.2.1 Construction of a bi-layer network

Compared to the homogeneous networks common to bibliometrics, such as co-term and co-authorship networks, bi-layer networks are heterogenous. Integrating two dimensions of bibliographical information into one framework allows us to leverage new angles and new topological structures to discover more comprehensive insights. Figure 3 provides a simple illustration.

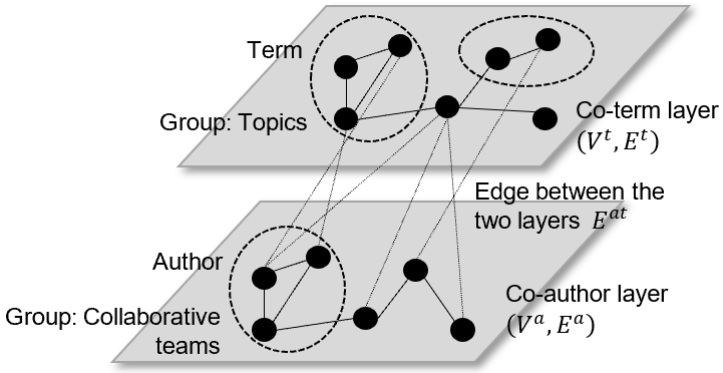


Figure 3. A simple example of a bi-layer network

The first layer of the network is a term co-occurrence network (co-term for short), and the second is a co-authorship network (co-author for short). Both layers are undirected graphs. The nodes on the first layer represent terms that reflect technological components, such as materials, functions, manufacturing processes, applications, etc. The links represent term co-occurrence. Likewise, the nodes on the second layer represent authors, and the links denote collaborations. Authors are tracked at the individual level, not the institutional level, although their affiliations are preserved. The links between the layers indicate the number of times an author mentions a term as a representation of research activity. All co-occurrence statistics are derived from the pre-processing phase, and all links are weighted by the normalized frequency of the co-occurrence relationship.

The network is formulated as $N = \{(V^t, E^t), (V^a, E^a), E^{at}\}$, where (V^t, E^t) denotes the sets of nodes and links on the co-term layer, and (V^a, E^a) denotes the same on the co-author layer. E^{at} is the set of links between the two layers.

Although our goal is to identify which nodes in the co-term layer are EGPTs via a ranking strategy, the co-authorship layer and the links between the two layers are an important factor in this assessment as these are integral to quantifying fundamentality, speciality, and sociality. The specific formulations follow.

3.2.2 Quantification of technological characteristics

(1) Fundamentality

Fundamentality measures the extent of the wide-range domains, research areas, disciplines, or sectors a technology impacts. It operates on the co-term layer and is a combination of Freeman's (1978) three indicators of *centrality* – degree, closeness, and betweenness. A traditional measure of network topology, each strand of centrality reflects a different aspect of the power relationships hold – degree measures involvement, closeness measures access to other nodes, and betweenness measures the ability to control the flow of the network. From a systematic examination of the advantages and disadvantages of the three indicators, Opsahl et al. (2010) find that degree only considers the local structures around a node, closeness is usually limited to the largest part of a network, and betweenness may ignore nodes in unimportant positions, in which a large number of nodes will receive a score of 0. Applying these findings to our context, we realized that many EGPTs might fall into one of these categories. Therefore, rather than grapple with each indicator separately, we decided to integrate all three into a new weighted index, i.e., fundamentality. Standard formulations of the three individual indicators are given below, followed by the steps of our integration and weighting procedure.

- Degree reflects the number of connections to a node (Freeman, 1978). The higher the degree, the greater the importance of a node to its neighborhood. From Opsahl et al. (2010) observations, we know this indicator

captures local structures and so it best reflects a technology's impact on closely-connected technological components. The equation for calculating the degree of node v_i^t is

$$DC(v_i^t) = \frac{\sum_{j=1}^{|V^t|} w_{v_i^t, v_j^t}}{|V^t|-1} \quad (1)$$

where $|V^t|$ is the number of nodes in the layer, and $w_{v_i^t, v_j^t}$ is the weight of the link between node v_i^t and node v_j^t.

- Closeness measures the length of the paths from a node to all other nodes (Opsahl et al., 2010) and characterizes a node's engagement across the largest component of a network. Thus, this indicator reflects a technology's impact on a field, which can be thought of as the extent to which a technology is an essential component of other relevant technologies (Aarstad et al., 2015). The closeness of node v_i^t is calculated by

$$CC(v_i^t) = \frac{|V^t|-1}{\sum_{j=1}^{|V^t|} d_{v_i^t, v_j^t}} \quad (2)$$

where $d_{v_i^t, v_j^t}$ is the shortest distance between node v_i^t and node v_j^t.

- Betweenness counts the number of times a node serves as the shortest path between two other nodes (Newman, 2005). It measures the power of a node to control the flow of the network. In other words, betweenness asks: How critical is this node to the network structure? If it were to disappear, would part of the network topology break down? This indicator operates in the entire network, reflecting the level of entire technological systems, and is calculated by

$$BC(v_i^t) = \frac{2\sum \frac{\sigma(v_i^t)_{v_s^t, v_p^t}}{\sigma_{v_s^t, v_p^t}}}{(|V^t|-1)(|V^t|-2)}, v_i^t \neq v_s^t \neq v_p^t \quad (3)$$

where v_s^t and v_p^t are two different nodes on the layer, $\sigma_{v_s^t, v_p^t}$ represents the number of the shortest paths between nodes v_s^t and v_p^t, and $\sigma(v_i^t)_{v_s^t, v_p^t}$ is the number of the shortest paths between nodes v_s^t and v_p^t, crossing node v_i^t.

The challenge to combining these three measures into one is retaining their diverse emphases. Therefore, rather than simply summing or averaging the three values, we apply an entropy weighting first approach (Grupp, 1990). This captures the dynamics of the indicators considered (Zhang, Qian, et al., 2017) - briefly, the key assumption of entropy weighting is that the more common an indicator is, the less weight it has. As such, entropy weighting is driven by the data.

Using degree centrality as an example, the weight w_{dc} of $DC(v_i^t)$ can be calculated via the following steps.

Step 1: Normalize $DC(v_i^t)$ as $f_{dc}(v_i^t)$ using the following max-min normalization approach:

$$f_{dc}(v_i^t) = \frac{DC(v_i^t) - \min\left(DC(v^t)\right)}{\max\left(DC(v^t)\right) - \min\left(DC(v^t)\right)} \quad (4)$$

where $\max\left(DC(v^t)\right)$ and $\min\left(DC(v^t)\right)$ respectively represent the maximum and minimum value of $DC(v_i^t)$ in the co-term layer.

Step 2: Calculate the entropy H_{dc} as follows:

$$H_{dc} = -\frac{1}{\ln|V^t|}\sum_i^{|V^t|} f_{dc}(v_i^t) ln f_{dc}(v_i^t) \quad (5)$$

Step 3: Convert the entropy value H_{dc} into a weight w_{dc} for $DC(v_i^t)$ via

$$w_{dc} = \frac{1-H_{dc}}{3-\sum H} \ (0 \leq w_{dc} \leq 1) \quad (6)$$

where $\sum H$ represents the sum of the entropies of the three forms of centrality.

The weights for all three forms of centrality can be calculated following the above steps, after which the fundamentality $F(v_i^t)$ of node v_i^t can be derived via

$$F(v_i^t) = w_{dc}f_{dc}(v_i^t) + w_{cc}f_{cc}(v_i^t) + w_{bc}f_{bc}(v_i^t) \quad (7)$$

where w_{cc} and w_{bc} represents the weight for $CC(v_i^t)$ and $BC(v_i^t)$, respectively, existing $w_{dc} + w_{cc} + w_{bc} = 1$.

(2) Speciality

Speciality also applies to the co-term layer. It measures the capacity of a technology to create relationships between other technological components in the same area. Whereas fundamentality operates at a range of granularities, speciality introduces the concept of communities, which roughly equate to Rotolo's narrower scope of one or a few domains. Formally, the concept of community here follows the definition given by Girvan and Newman (2002) – that is, a community is a set of proximally located nodes in a network that share similar features and where the density of edges in that region is higher than in the immediately surrounding regions. Topologically, a community can represent a smaller part of a network than closeness would capture or a larger area than captured by degree centrality, resulting in overlaps. For example, speciality might manifest as relevance to a technological process as opposed to a discipline.

Measuring speciality is a two-step process. First, a community detection algorithm clusters the terms into communities G^t. (We chose the smart local moving algorithm (Waltman & Van Eck, 2013) due to its established recognition from the bibliometric community.) Then, speciality $Sp(v_i^t)$ of node v_i^t is calculated

$$Sp(v_i^t) = \frac{\sum_{j=1}^{|G^t(v_i^t)|} w_{v_i^t, v_j^t}}{|G^t(v_i^t)|} \quad (8)$$

where $|G^t(v_i^t)|$ represents the number of nodes in the community to which node v_i^t belongs.

(3) Sociality

The sociality of a technology takes both terms and authors into consideration. From a macro perspective, if a technology is owned by more than one enterprise or research institution and can be easily transferred between those owners, or even between different sectors, it is likely to have high socio-economic impact. Thus, both the co-term and the co-authorship layers contribute to quantifying the sociality of a node $E(v_i^t)$ on the co-term layer, as outlined in Equations (9) and (10).

$$L(v_m^a) = \sum_{n=1}^{|V^a|} w_{v_m^a, v_n^a} \quad (9)$$

$$E(v_i^t) = \sum_{n=1}^{|V^a|} w_{v_i^t, v_n^a} \times L(v_n^a) \quad (10)$$

where v_m^a is a node on the co-authorship layer and $|V^a|$ is the number of nodes on the co-authorship layer.

3.2.3 Further expression of the three measureable characteristics

Fundamentality, speciality, and sociality aptly reflect emergence and generality in the context of an EGPT. Fundamentality measures a technology's impact, and/or potential for impact, in the broad technological landscape by leveraging the local and global topological structure of the co-term layer. Speciality measures a technology's impact in a specific technical area based on the attributes of a community. And sociality measures a technology's social engagement as an indicator of the potential for socio-economic impact through the co-author layer and the between-layer interactions of the bi-layer network.

These three indicators can be used in a number of ways. First, we can use a 3D map to highlight one, two, or all three of these values. Second, we can use multi-objective optimization to roll all three indicators into one and generate a ranked list of EGPTs. Third, different weighting approaches beyond the standard frequency weights can be applied to suit particular purposes, such as entropy or standard deviation. Last, any of these options can be coupled with selection criteria and expert knowledge to focus and/or enrich the analysis.

However, while each indicator is a strong measure of different sorts of impact, their capacity to represent both emergence and potential needs bolstering. These are the purposes of the link prediction method, discussed next.

3.2.4 Link prediction

Incorporating link prediction into the methodology introduces a time dimension through which to trace emergence and to reveal potential. Technically, link prediction approaches could fill in any missing connections in the bi-layer network. The result is a predicted bi-layer network. Generating two ranked lists of technologies – one from the original network and one from the new predicted network – and comparing the differences between the two reveals the potential impact an existing technology may have in future and, most importantly, which terms most probably represent EGPTs.

Compared to algorithms like common neighbors and Adamic-Adar (Lü & Zhou, 2010), the resource allocation (RA) algorithm yields better precision. The RA algorithm assumes that every node in a network has an amount of resources, and common neighbors serve as transmitters to evenly distribute one node's resources to connecting nodes (Ou et al., 2007). There are no constraints on the types of nodes that can serve as transmitters; rather, resources are treated more like heat, diffusing from a hot place to a cold place and back again. Inspired by this, Zhou et al. (2007) proposed an approach of personal recommendation by tracking the flows of resource allocation in a bipartite network to produce personal recommendations. They constructed a network consisting of two sets of nodes; however, resources were only allowed to spread between the two sets of nodes, not amongst the nodes within one set. Following this thread, Zhang, Wang, et al. (2018) applied the same approach to two networks independently, first to a co-term network then to an author-term network, to predict future scientific activity, such as the possible future directions of a research topic and the prospective next topics of interest for specific researchers.

Working along the same lines, we have simply extended the scope of Zhang, Wang et al. (2018) from a bipartite network to a bi-layer network. The strategy can be thought of as applying the same technique to three bipartite networks, i.e., a co-term network, a co-authorship network, and an author-term network. In the author-term network, each author is assumed to initially hold an amount of resources. In the first-round allocation, the resources will diffuse to terms the author has mentioned. Then, in the second-round, the resources will diffuse back to other authors that have also mentioned these terms. Hence, the initial resource allocations diffuse from one author to other authors sharing similar research interests. Similarly, in the co-term/co-author networks, the initial resource allocations move from a term to co-occurring terms and from an author to their co-authors.

The net effect is that resources can be reallocated in all three ways – between terms, between authors, and between terms and authors. As such, the RA index of two nodes is the sum of the resources allocated to those two nodes by all their common neighbors, calculated as per Equation (11).

$$RA(v_x, v_y) = \sum_{v_z \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{|\Gamma(v_z)|} \tag{11}$$

where v_x and v_y are two different, unlinked nodes in either the co-term or co-authorship layer of the bi-layer network, and $\Gamma(v_x)$ denotes the set of nodes neighboring v_x.

The approach taken in several previous studies has been to evenly distributed resources to connected nodes (Zhang, Wang, et al., 2018; Zhou et al., 2007). However, with a bi-layer network, we wondered whether a stronger connection between nodes might warrant a higher chance to gain resources. For instance, a term frequently mentioned by an author could be more important to that author than terms mentioned only once. Indeed, as the later experiments show, weighting each link in the network according to its co-occurrence frequency marginally improves the accuracy of the final predictions. Hence, links on the co-authorship layer are weighted according to the number of co-authored papers, and links between the two layers are weighted according to how frequently a term has been mentioned by the author. The weighting formula is:

$$WRA(v_x, v_y) = \sum_{v_z \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{w_{v_x,v_z} + w_{v_y,v_z}}{\sum_{v_k \in \Gamma(v_z)} w_{v_k,v_z}} \tag{12}$$

The final output of the link prediction procedure is a ranked list of all links in the new network, i.e., the predicted network, including missing links, to be used for subsequent comparison with the current network.

3.2.5 Selection criteria for identifying EGPTs

As discussed in Section 2.3, the three technological characteristics provide strong quantitative evidence for a technology's historical influence. However, although historical data can provide some clues as to future potential, emergence can only truly be captured from a future-oriented analysis. The link prediction approach described in the previous section provided the data, but the next step is to conduct the actual analysis, i.e., to assess the differences between the two ranked lists of technologies and judge which should be deemed EGPTs. To assist with the analysis, we designed three criteria for shortlisting the candidates:

- *Criterion 1*: A technology only appears in the predicted list (List B) and with a high rank.
- *Criterion 2*: A technology in List B has a dramatically higher rank than in the original list (List A).
- *Criterion 3*: A technology appears among the top ranks of both lists.

These selection criteria complement the three technological characteristics, creating a semi-automatic solution for identifying EGPTs and reducing the expert labor required to produce a definitive list. More specifically, many candidate technologies could be immediately eliminated from consideration by simply deciding thresholds for what constitutes a high rank for Criteria 1 and 3, and the minimum difference in ranks for Criterion 2. The terms remaining could either be deemed EGPTs, or manually reviewed by experts as a final step in the selection process.

3.3 Validation

For a comprehensive evaluation of the framework, we conducted two different validation protocols: one experimental to validate the link prediction model, and the other empirical to assess the final shortlist of candidate EGPTs. Details of the two procedures follow.

3.3.1 Experimental validation

To assess the efficacy of the link prediction model and, more specifically, the weighted resource allocation (WRA) algorithm, we selected four baselines, conducted two link prediction scenarios, and compared the results to a ground truth. The chosen baselines were:

- *Common Neighbors* (CN), which simply counts the number of common neighbors. This is the most basic and direct form of measuring neighborhood overlaps.
- *Jaccard Coefficient* (JC), which calculates the proportion of common neighbors between two unlinked nodes. This is a more comprehensive approach than CN.
- *Adamic-Adar Index* (AA) – a refined version of the CN algorithm that assigns more weight to common neighbors with smaller degrees.
- *Resource Allocation* (RA) – the unweighted version of resource allocation.

The results were analyzed in terms of receiver operating characteristic (ROC) and the area under the curve (AUC) following the design of Fawcett (2006). The two link prediction scenarios were designed as follows:

Experiment 1: Random Removal. We randomly removed 10% of the existing links and ran the four baselines plus the proposed algorithm over the network. We then compared how many of the missing links each algorithm was able to identify.

Experiment 2: Timeline Prediction. Using the publication year of each document as a time stamp, we divided the entire dataset into time windows to simulate a data stream. Records in the 1996-2016 set were used as the training set; records from 2017-2018 served as the test set.

3.3.2 Empirical evaluation

Our evaluation process comprised both expert analysis and a comparison with evidence summarized in previous similar studies. We design the empirical evaluation for examining those EGPTs characterized in the case study, in which domain experts will be invited to mark those selected technologies (i.e., terms and phrases) based on their expertise. Additionally, as a supplementary approach, we will also empirically compare our key findings with evidence summarized from previous studies in similar cases.

4 Results

Our chosen discipline for this case study is information science. The choice to analyze one discipline may seem unusual given that we are, at least in part, testing the methodology's efficacy at predicting technologies that span a broad range of disciplines. Our reasoning here is that information science has, for some time, been a spearhead for cross-disciplinary research. Information science connects fundamental studies, such as mathematics, physics, and computer science, with the real-world needs discussed in the social sciences. Therefore, investigating information science "kills two birds with one stone", so to speak. It provides the opportunity to examine EGPTs that originate from one discipline but are, or will become, key components in others. It also provides the opportunity to explore how the methodology works when focused on an individual discipline, shedding light on how adaptable the methodology might be to any given discipline. Lastly, as information scientists ourselves, it is relatively easy for us to find willing experts to help with validation.

4.1 Data collection and pre-processing

To assemble our corpus, we followed the search strategy proposed by Hou et al. (2018), retrieving 17,445 articles published between 1 Jan 1996 and 31 Dec 2018 from the 15 journals and conference proceedings listed in Table 1.

Table 1. Selected journals and conference proceedings.

Journal Name	*Journal Name*
Annual Review of Information Science and Technology	Journal of Documentation
Information Processing & Management	Scientometrics
Journal of the Association for Information Science and Technology	Information Research
Library Resources & Technical Services	Journal of Informetrics
Journal of Information Science	Research Evaluation
Library & Information Science Research	The Electronic Library
ASIS&T Annual Meeting Proceedings	Information Technology and Libraries
Program: Automated Library and Information Systems	

Note: Previous journal names were considered when collecting data.

We extracted 238,789 terms from the titles and abstracts of the 17,445 articles using the NLP function in VantagePoint[1]. The data was then cleaned to remove noise and consolidate synonyms using the term clumping process (Zhang et al., 2014). The stepwise results of this process are given in Table 2. The final 4,773 terms were used to construct the co-term network.

Table 2. Stepwise results of the term clumping process

Step	*Description*	*#Terms*
0	Raw terms retrieved with NLP	238,789
1	Remove terms starting/ending with non-alphabetic characters, e.g., "step 1" or "1.5 m/s"	211,909
2	Remove common terms in scientific articles, e.g., "research framework"	204,984
3	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	202,666
4	Consolidate synonyms based on expert knowledge, e.g., "co-word analysis" and "word co-occurrence analysis"	190,772
5	Consolidate terms with the same stem, e.g., "information system" and "information systems"	171,285
6	Remove single-word terms, e.g., "information"	161,255
7	Remove terms appearing less than five times	4,773

[1] VantagePoint is a software platform for bibliometrics-based text analytics and knowledge management owned by Search Technology Inc. More details can be found at the website: www.vantagepoint.com.

Note that: 1) The expert knowledge in Step 4 was mostly based on information gleaned from previous experiments and experiences assembled into thesauri for automatic term consolidation. 2) Typically, we would only remove terms that do not appear more than once (Step 7). But, in this case, we increased the threshold to reduce the scale of terms. 3) From a manual review of the top thousand most-frequently mentioned terms on the list, we then marked 128 (with the help of our experts) that we felt would be interesting to examine in more detail subsequent to the initial analysis. Terms representing specific technologies, concepts, approaches, algorithms, and research topics made the list. Terms with relatively general meanings in isolation, such as country names (e.g., United States), affiliation names (e.g., National Library), and jargon like scientific publications or research output, did not.

In terms of authors, we collected a raw list of 18,882 names and cleaned the list with the "light" disambiguation function in VantagePoint, which consolidates name variations like "Eugene Garfield", "Garfield, Eugene", and "E Garfield". We then removed authors with only one published paper, leaving 3,113 authors from which we constructed the co-authorship network.

4.2 Bi-layer network analytics-based prediction

We built the bi-layer network and ran the WRA algorithm to generate the predicted network. The descriptive statistics are given in Table 3. We then visualized a small section of the full network, both original and predicted, using Gephi, as shown in Figure 4. The entire network is simply too large to illustrate with any clarity using existing visualization tools. The section rendered includes 10 authors and 12 terms across four communities (distinguished with different colors). At this scale, new and interesting links become clear, such as a potential co-authorship between Lutz Bornmann and Ronald Rousseau, possible research topics in information retrieval that Mike Thelwall may be interested in, and a potential technological recombination between citation analysis and machine learning techniques.

Table 3. Descriptive statistics of the bi-layer network

		Count	*Weight**			
			Max.	*Min.*	*Mean*	*Standard deviation*
Node	Author	3,113	198	2	7.259	11.357
	Term	4,773	1,017	6	20.486	41.857
Links	Co-authorships	6,065	85	1	2.151	2.678
	Co-terms	250,363	188	1	1.294	1.291
	Term-author	102,741	36	1	1.231	0.837

Note that the weight of nodes represents the number of articles associated with this node, and the weight of links reflects the co-occurrent frequency between its connected nodes.

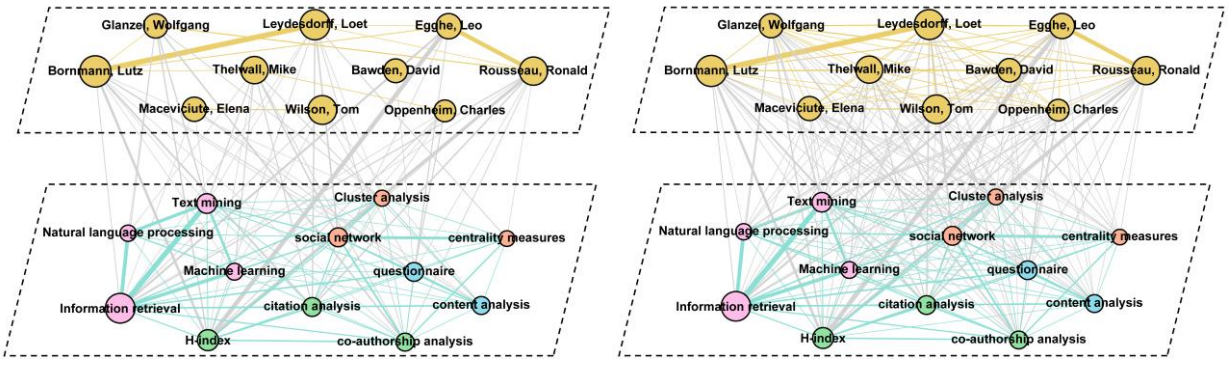


Figure 4. A sample of the bi-layer network of information science (1996 to 2018) – (left) for the current network and (right) for the predicted network

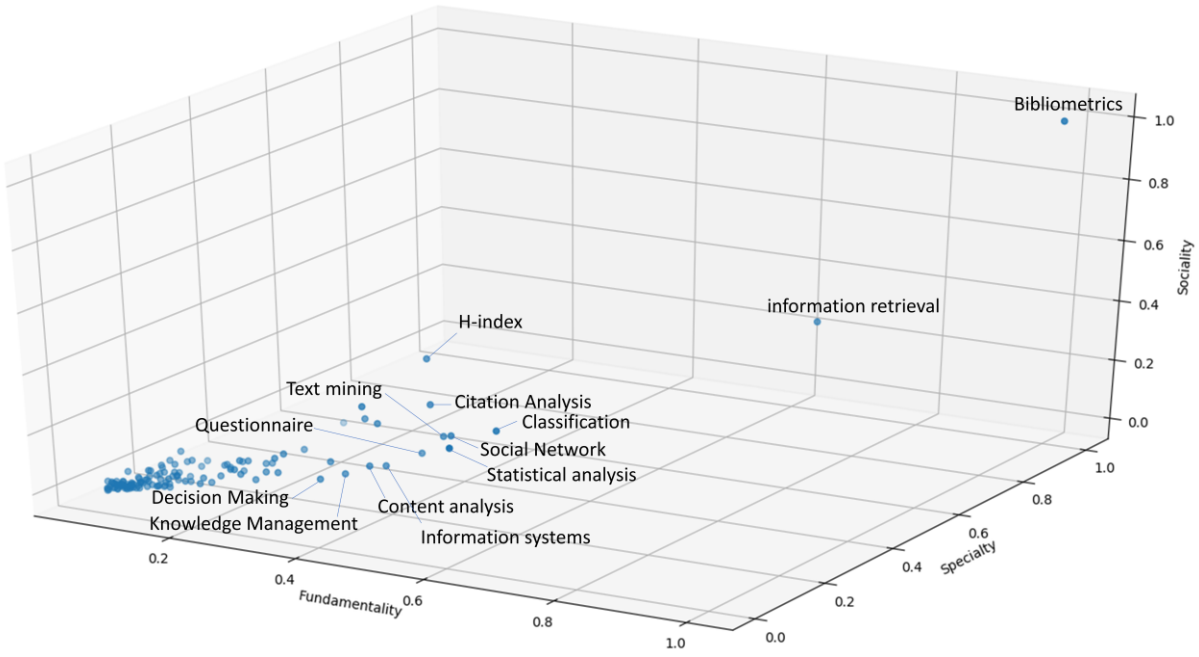
We then calculated the three technological characteristics for all 4,733 nodes on the co-term layer and compared changes in these three indicators between the current network and the predicted network. The descriptive statistics

appear in Table 4. Overall, adding in the predicted missing links increased the fundamentality values but did not significantly influence the other two indicators.

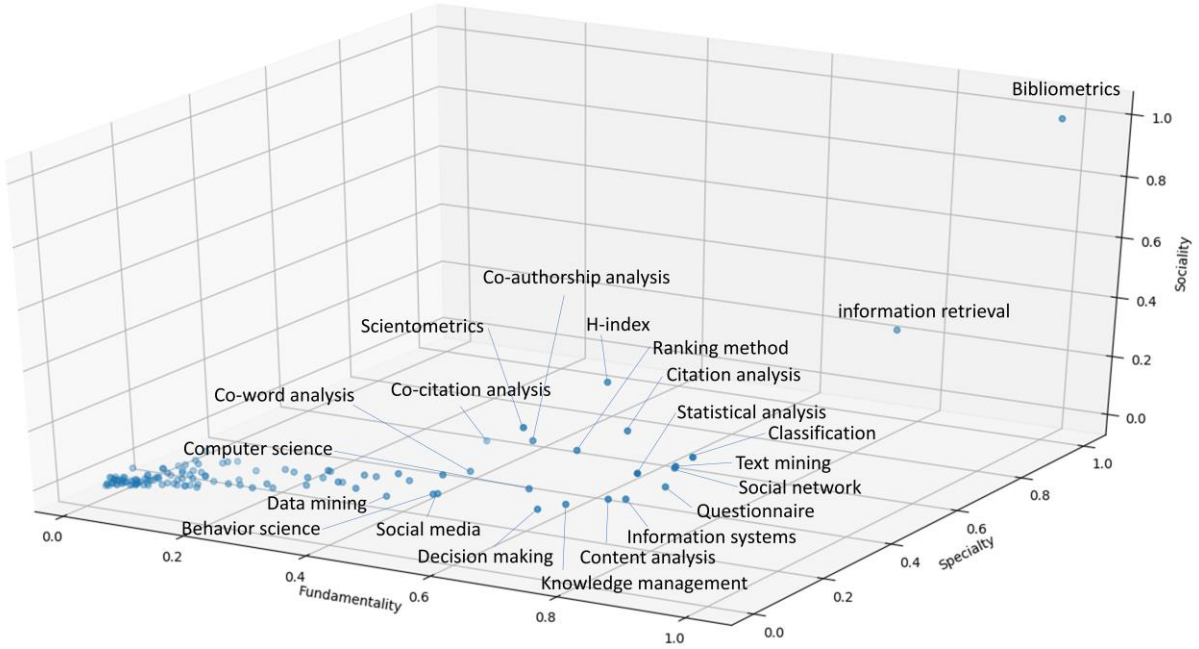
Table 4. Descriptive statistics for the technological characteristics of the current versus the predicted bi-layer networks (normalized)

		Current network		Predicted network	
Characteristic	*Sub-characteristic*	*Mean*	*Std dev*	*Mean*	*Std dev*
Fundamentality	Degree centrality	0.018	0.055	0.034	0.076
	Closeness centrality	0.164	0.111	0.196	0.207
	Between centrality	0.039	0.092	0.073	0.173
	Fundamentality (entropy weighting)	0.122	0.099	0.141	0.173
Speciality	N/A	0.055	0.073	0.054	0.071
Sociality	N/A	0.036	0.060	0.042	0.065

Narrowing our analysis to the 128 selected terms noted earlier, we generated a 3D map of each network, as shown in Figure 5. Initial observations show obvious gaps between the majority of terms and those few with a high value in at least one indicator. The terms ‘bibliometrics’ and ‘information retrieval’ retain their high ranks in both networks, while terms like ‘social network’, ‘text mining’, ‘classification’, ‘information systems’, and ‘content analysis’ further fortify their positions as integral technologies.



(a) Current network



(b) Predicted network

Figure 5. 3D map of the fundamentality, speciality, and sociality indicators for 128 selected terms.

Following the protocol for shortlisting EGPTs, we applied each of the three criteria to all 1000 terms (with top frequency), but paid particular attention to charting changes in the 128 selected terms. The stepwise process and accompanying results follow:

Step 1 Two ranking lists were generated for the 1000 terms: List A from the current network; List B from the predicted network. The ranks of the 128 selected terms in each list were noted.

Step 2 45 terms of the 128 terms only appeared in List B and with a rank of higher than 200 (refined Criterion 1).

Step 3 27 of those 45 terms had increased by more than one rank from List A to List B (Criterion 2).

Step 4 5 terms appeared in the top 20 ranks of both lists (Criterion 3).

Step 5 30 terms (the 27 from Step 3 and 5 from Step 4) were selected as the shortlist for further analysis – see Table 5.

The choice of thresholds – 200 for Step 2 and 20 for Step 4 – was purely logistical. These limits resulted in a manageable size of candidate terms for visualization and for our experts to validate. However, a tip to inform future studies is that the more terms included in the candidate list, the higher the chances of identifying an emerging technology with a relatively low term occurrence. The trade-off is noise and a greater burden on experts.

Table 5. The 30 candidates with the potential to be EGPTs

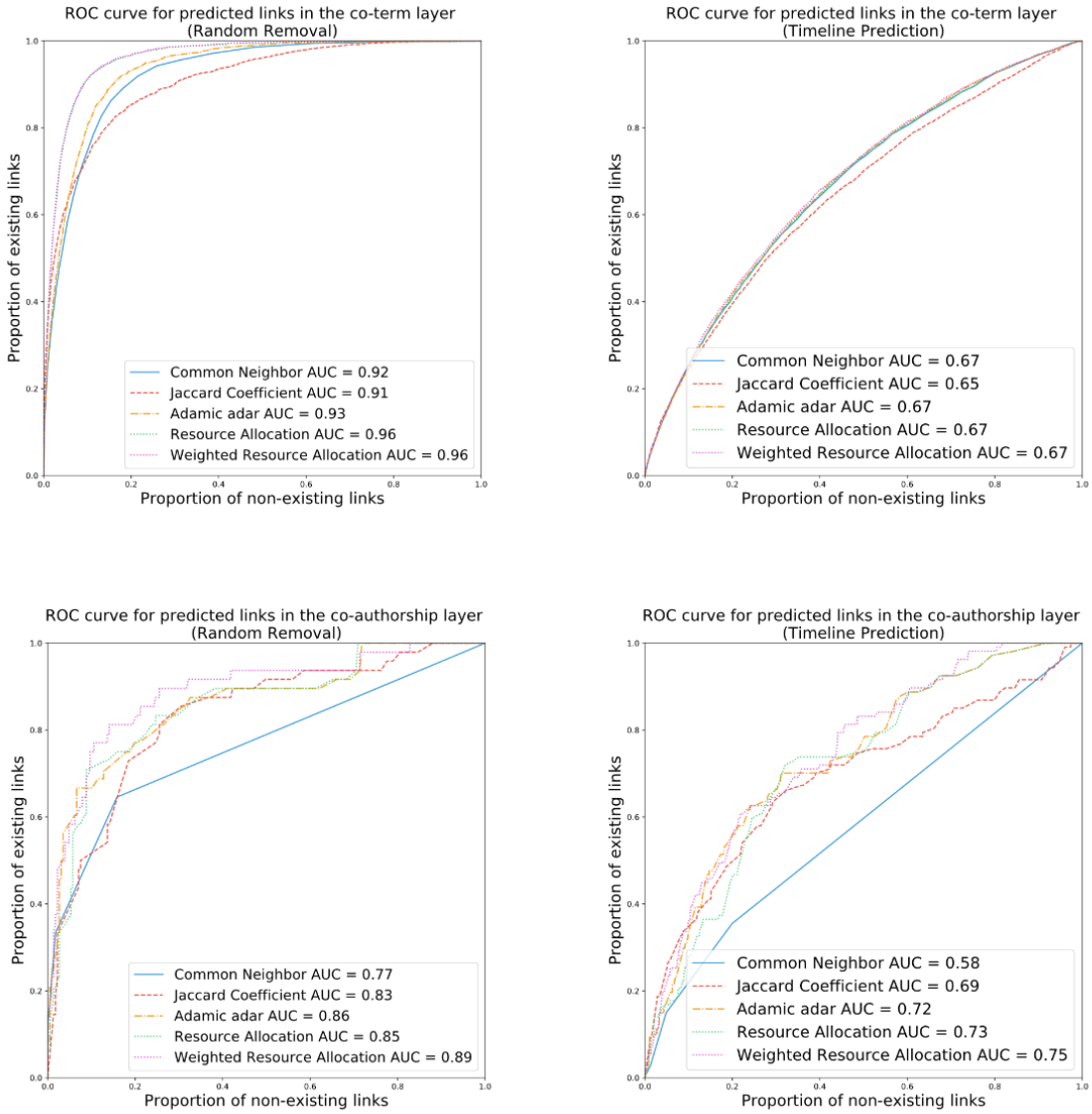
	Terms
The 27 terms in the top 200 that increased their ranks from List A to List B	social media, semi-structured interviews, data mining, probability, data analysis, decision making, co-occurrence analysis, behavioral science, knowledge engineering, information systems, semantic analysis, natural language processing, citation network, knowledge management, full text, content analysis, machine learning, questionnaire, co-word analysis, text mining, metrics, social network, citation analysis, query processing, network analysis, information retrieval, computer science
The 5 terms appearing in the top 20 of both lists	bibliometrics, information retrieval, h-index, classification, citation analysis

4.3 Validation results

This section begins with the results of the link prediction experiments, followed by the empirical evaluation of the candidate EGPTs.

4.3.1 Experimental validation for link prediction

The ROC curves for both the random removal and timeline prediction scenarios appear side-by-side in Figure 6. Vertically, the panels show the various layers of the network. The AUC values for each layer are noted in the legends of each chart, and the averages are provided in Table 6.



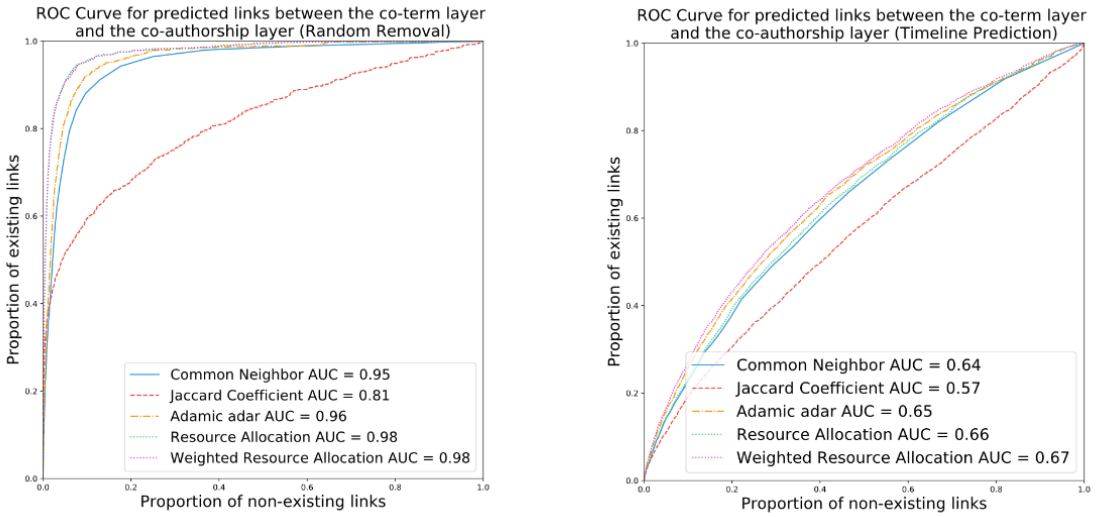


Figure 6. Validation results of the link prediction experiments
Note: The random removal scenario appears on the left, with the time prediction scenario on the right.

Table 6. Average AUC values for the five approaches by experiments

Baseline	*Random removal*	*Timeline prediction*
CN	0.88	0.63
JC	0.85	0.64
AA	0.92	0.68
RA	0.93	0.69
WRA	**0.94**	**0.70**

According to the results, WRA and RA were the most accurate in both experiments. WRA's slight advantage over RA demonstrates the benefits gained from the including the weighted index of resource allocation. An interesting observation is that all the values for the timeline prediction are substantially lower than for random removal. We attribute this to the time lag associated with technological recombination. Technological innovation has its own lifecycle, but our data divisions only left a two-year window to observe potential innovations. Not all technologies will recombine or progress over such a short time. Certainly, measuring technological change by detecting the dynamics of a network's topological structure over time could be an interesting avenue in both bibliometrics and innovation management, and we anticipate future studies in this direction.

4.3.2 Empirical evaluation

As mentioned, the empirical evaluation comprised two components: one with a panel of experts and the other based on evidence from the literature.

(1) Expert assessment

Five domain experts that are actively involved in information science and/or its related disciplines (e.g., ST&I management) were invited to form a panel to evaluate the results. Individually, they represent the Georgia Institute of Technology, Nanjing University, the South China University of Technology, the Chengdu Branch of National Science Library of Chinese Academy of Science, and the Beijing Institute of Technology. Prior discussions with the panel members confirmed our suspicions that the qualities of an EGPT are too subjective (hence, the need for this study). By contrast, there is much more consensus surrounding the constructs of emergent and generality. Therefore, we chose to separate the two. A list of the 30 candidates (ordered alphabetically with the ranks removed) was sent to each expert for independent assessment, along with the following instructions:

- Each term should be as marked as generality and emergence.
- Based on your expertise, score each term in the interval [0, 1] as generality and emergence, where 0 means strongly disagree, 1 means strongly agree, [0, 0.5] indicates a negative sentiment, and [0.5, 1] indicates a positive sentiment.

The descriptive statistics and intraclass correlation coefficients (ICC) for the expert assessment follow in Table 7. Recall that each expert was asked to score each of the 30 terms for its qualities of emergence and generality.

Table 7. Descriptive statistics and intraclass correlation coefficients among expert marks.

	Max	*Min*	*Avg.*	*Std dev*	*ICC*
Generality	0.717	0.507	0.584	0.080	0.824
Emergence	0.605	0.337	0.493	0.104	0.766

According to Table 7, the ICC values for both features indicate great agreement among the five experts. The average mark of 0.584 for generality, with a maximum mark at 0.717 and a standard deviation of 0.08, reflects acceptance that the terms do represent GPTs in information science. The scores for emergence are not as appealing as we had hoped for, but they do confirm our contention in Section 2.3 that historical data is very useful for identifying GPTs while discovering emerging technologies is likely to require further future-oriented analysis. It was intriguing to see how many and which terms received high marks from experts in both features. Thus, we plotted the averages on a quadrant, as shown in Figure 7. The division point for each feature sits at 0.5.

From the chart, we find:

- The 6 terms in the upper right quadrant (Q1) could be EGPTs, e.g., content analysis, semantic analysis, text mining, and natural language processing. These terms received marks higher than 0.5 for both qualities, which might be strong evidence of the bibliometric community's increasing interest in uncovering complicated semantics from bibliometric texts.
- The 9 terms in the upper left quadrant (Q2) could be novel technologies that have not yet sufficiently infiltrated information science, e.g., machine learning, social networks, full text, social media, and behavior science. These terms received relatively high marks for emergence but not for generality. So far, they are not influential enough to be considered GPTs.
- The 13 terms in the bottom right quadrant (Q3) could be GPTs but not EGPTs, e.g., co-word analysis, co-occurrence analysis, and classification. These terms are well-recognized in information science and received relatively high marks for generality but not for emergence.
- The 2 terms in the bottom left quadrant (Q4), i.e., probability and h-index, are too broad and non-specific to represent technologies. We classified this quadrant as incorrect characterizations.

Quantifying results like these is simplistic. However, from this rudimentary perspective, it is reasonable to estimate that: 1) 20% of the results (Q1) were strongly endorsed by our experts; 2) our error rate was 6.67% (Q4); and 3) around 77% of the terms sat somewhere on the continuum between emergent and general purpose.

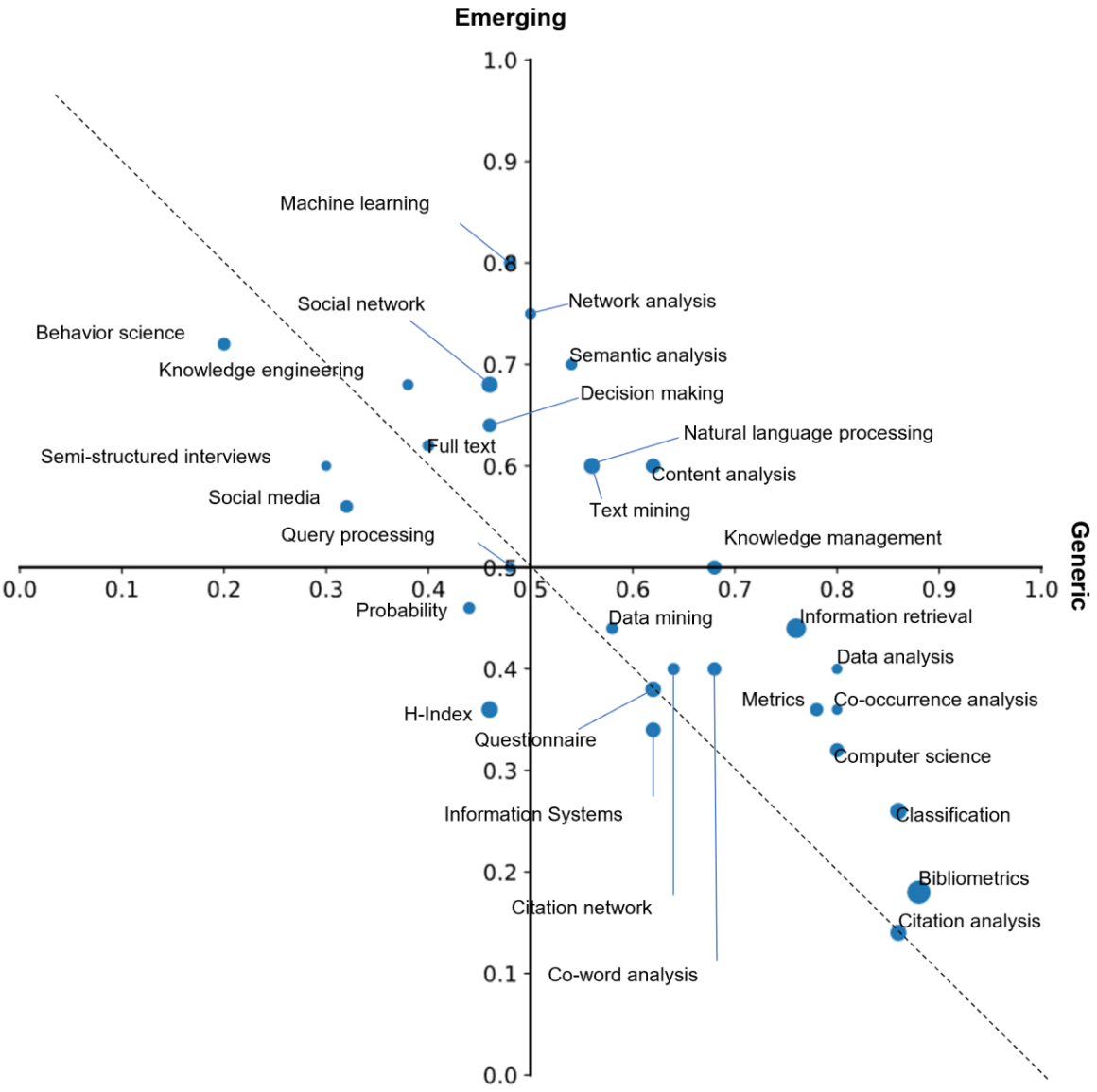


Figure 7. A quadrant map of the 30 candidate technologies.

Note that nodes "Text mining" and "Natural language processing" overlap with each other in the map because they are at the same placement in the quadrant.

Based on our own knowledge of information science and on discussions with the expert panel, we find these results to be an acceptable starting point. More importantly, the insights uncovered from this evaluation endorse our claims of both the challenges and opportunities associated with characterizing EGPTs from a bibliometric perspective. Both give us pause to continue working on the methodology.

(2) Evidence from the literature

Given the diverse understandings of generality and emergence held by our experts, it was critical that we triangulate our findings with an empirical assessment of evidence in the literature. For this, we chose the results of two bibliometric studies and compared our predictions to the conclusions drawn from these published analyses. The details of each follow.

- Hou et al. (2018), Emerging trends and new developments in information science: A document co-citation analysis, published in *Scientometrics*.

The corpus for this study was drawn from the exact search strategy proposed by Hou et al. (2018) but with an extended sample period from 1996-2016 up to 1996-2018. Since their goal was also to identify emerging trends and new developments in information science, it is intriguing to compare our results with theirs in terms of which technologies should be emerging. We made two main observations from analyzing their text:

- In their "Emerging trends…" section, Hou and his colleagues identified scientific evaluation (e.g., h-indexes), bibliometrics, citation analysis, scientific collaboration, knowledge mapping and visualizations, and altmetrics (with a focus on social networks) as the next research frontiers in information science. We share four of these six predictions (see Table 5), and the two differences are very explainable. 1) Scientific collaboration may be a research frontier, but it is a practical application, not a technology, so it did not make our list of 128 selected terms. In fact, we explained h-indexes away in Q4 as an error. 2) The specific term "knowledge mapping and visualization" does not appear in our list but is a summary of "co-occurrence analysis", "co-word analysis", "co-citation analysis", and "bibliographic coupling". We also suggest that these technologies may be more general than emergent.
- Both their study and our predicted network identified techniques in computer science as a mainstay of information science now and in the future, but with slightly different terminologies. Their study cites the term "system, computing, and computer", and, in our list, there are terms like "information systems", "data mining", "machine learning", "text mining", and "natural language processing".

- Zhang, Lu, et al. (2018), Does deep learning help topic extraction? A kernel k-means clustering method with word embedding, published in the *Journal of Informetrics.*

  This paper is a pilot study involving the first author of this paper on using word embedding techniques to improve topic extraction focused on articles published in the three top journals in bibliometrics: the *Journal of the Association for Information Science and Technology*, the *Journal of Informetrics*, and *Scientometrics*. The evaluation case was to identify the most fundamental topics in bibliometrics. By our definition, this may coincide with generality. Eight integral technologies were identified in the article. Comparing these to the results from this methodology, we find:

- Five of the eight topics are common: information behavior (i.e., behavior science), bibliometric analysis, citation analysis, information retrieval, and h-index. Two of the remaining three topics – "research performance" and "scientific collaboration" – are practical applications of bibliometrics and so were not included in our focused analysis as previously discussed. The final topic, "search engine", does not appear on our list but "query processing", which evokes similar concepts, does.

Thus, both the generality and emergence of the identified technologies do accord with previous knowledge and results in the literature. Therefore, it is reasonable to consider the predictions of EGPTs in information science as reliable and, in turn, the methodology that produced those predictions to also be reliable.

5 Discussion and conclusions

In this paper, we presented a methodology for characterizing EGPTs based on bi-layer network analytics. We defined three indicators to quantify the impact of EGPTs and applied a refined link prediction approach based on weighted resource allocation to reveal emergence. A comparison between the ranked terms in the current network and the predicted network reveals candidate EGPTs for further analysis by experts and/or an empirical review of the literature. We incorporated both types of analyses into a case study on information science. The results of each support the feasibility and reliability of the proposed methodology.

5.1 Technical implications for practical use

In terms of the methodology's robustness and adaptability, we found three sensitivity issues during our analysis. Therefore, we have several recommendations for researchers and analysts wishing to apply this methodology in their own work.

- With the exception of the community detection algorithm for calculating speciality, the three impact indicators are non-parametric. We were able to use the default setting of the small local moving algorithm within

VOSViewer, but it was clear to us that the number of detected communities could become a sensitive factor in calculating speciality values.

- We gave equal weight to each indicator because, in this exploratory study, there was no need to prioritize one characteristic over another. However, for most practical purposes, a weighting strategy that could optimize the combined indicator scores to emphasize certain characteristics as needed would probably be preferable. One way to accomplish this would be to calculate the combined indicator as a multi-objective decision-making problem.
- Narrowing the top 1000 terms down to a selection of the 128 required human intervention. However, information science is a research discipline, not a piece of technology. And, as yet, there is no fully automatic way to distinguish a technology from an application or a context or mode of practice given a simple list of terms. Such intervention may be considered as a specific process for this case and would not be required for pure technological areas, e.g., nanotechnologies, since there is no need to distinguish technology- and non-technology-related terms.

5.2 Limitations and future directions

Several future directions of research would address the limitations of this study. First, integrating the three impact indicators could be converted into a multi-objective decision-making task to be solved through computational intelligence. Beyond allowing a more flexible and complex weighting scheme, this may improve the efficiency of the process and allow for analyses in real-world environments with limited human intervention. In this iteration of the methodology, the bi-layer network only reflects co-term and co-authorship statistics. However, it may be interesting to add more layers reflecting information, such as citation and co-citation statistics. Examining and testing the methodology in more technological areas would help to further validate its adaptability and generate new thoughts for possible improvements.

6 Acknowledgments

An early version of this work was published in the *Proceedings of the 2019 International Conference of the International Society for Scientometrics and Informetrics* (Zhang, Zhu, et al., 2019).

This work was supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994 and the National Nature Science Foundation of China under Grant 71774013.

7 References

Aarstad, J., Ness, H., & Haugland, S. A. (2015). Network position and tourism firms' co-branding practice. *Journal of Business Research*, *68*(8), 1667-1677.

Allan, J. (2002). *Topic detection and tracking: Event-based information organization*. Springer.

Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of the literature. *Research Policy*, *16*(2), 131-141.

Basu, S., & Fernald, J. (2007). Information and communications technology as a general‐purpose technology: Evidence from US industry data. *German Economic Review*, *8*(2), 146-173.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892-895.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., & Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS One*, *7*(7), e39464.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of econometrics*, *65*(1), 83-108.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics*, *115*(1), 35-49.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015). On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, *58*(9), 82-90.

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change*, *158*, 120146.

David, P. A. (1989). *Computer and dynamo: The modern productivity paradox in a not-too distant mirror*.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C‐word, and cocitation methods. *Journal of the Association for Information Science and Technology*, *65*(10), 2084-2097.
Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187-203.
Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network [journal article]. *Scientometrics*, *95*(1), 225-242. https://doi.org/10.1007/s11192-012-0796-4
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.
Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215-239.
Furukawa, T., Mori, K., Arino, K., Hayashi, K., & Shirakawa, N. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. *Technological Forecasting and Social Change*, *91*, 280-294.
Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821-7826.
Glänzel, W., & Thijs, B. (2012). Using "core documents" for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399-416. https://doi.org/10.1007/s11192-011-0591-7
Graham, S. J., & Iacopetta, M. (2009). Nanotechnology and the emergence of a general purpose technology. *Annals of Economics and Statistics*, *115*, 116.
Grupp, H. (1990). The concept of entropy in scientometrics and innovation research: an indicator for institutional involvement in scientific and technological developments. *Scientometrics*, *18*(3-4), 219-239.
Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change*, *105*, 27-40.
Guo, Y., Huang, L., & Porter, A. L. (2010). The research profiling method applied to nano‐enhanced, thin‐film solar cells. *R&D Management*, *40*(2), 195-208.
Hall, B. H., & Trajtenberg, M. (2004). *Uncovering GPTs with patent data* (0898-2937).
Hicks, D., Martin, B. R., & Irvine, J. (1986). Bibliometric techniques for monitoring performance in technologically oriented research: The case of integrated optics. *R&D Management*, *16*(3), 211-223.
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., Jia, X., Zhang, Y., Zhou, X., & Zhu, Y. (2018). Detecting Hotspots in Interdisciplinary Research Based on Overlapping Community Detection. 2018 Portland International Conference on Management of Engineering and Technology (PICMET),
Huang, L., Zhu, Y., Zhang, Y., Zhou, X., & Jia, X. (2018). A Link Prediction-Based Method for Identifying Potential Cooperation Partners: A Case Study on Four Journals of Informetrics. 2018 Portland International Conference on Management of Engineering and Technology (PICMET),
Jovanovic, B., & Rousseau, P. L. (2005). General purpose technologies. In *Handbook of economic growth* (Vol. 1, pp. 1181-1224). Elsevier.
King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*(5), 261-276.
Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984-998.
Kostoff, R. N., & Schaller, R. R. (2001). Science and technology roadmaps. *IEEE Transactions on Engineering Management*, *48*(2), 132-143.
Lipsey, R. G., Bekar, C., & Carlaw, K. (1998). What requires explanation. *General purpose technologies and economic growth*, *2*, 15-54.
Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, *41*(6), 1462-1480.
Lü, L., & Zhou, T. (2010). Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, *89*(1), 18001.
Moser, P., & Nicholas, T. (2004). Was electricity a general purpose technology? Evidence from historical patent citations. *American Economic Review*, *94*(2), 388-394.
Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social networks*, *27*(1), 39-54.
Noyons, E., & Van Raan, A. (1998). Advanced mapping of science and technology. *Scientometrics*, *41*(1-2), 61-67.
Ohniwa, R., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. *Scientometrics*, *85*(1), 111-127.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, *32*(3), 245-251.
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.
Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *nature*, *435*(7043), 814-818.
Park, I., & Yoon, B. (2018). Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network. *Journal of Informetrics*, *12*(4), 1199-1222.
Petralia, S. (2020). Mapping general purpose technologies with patent data. *Research Policy*, *49*(7), 104013.
Porter, A. L., & Cunningham, S. W. (2004). *Tech mining: Exploiting new technologies for competitive advantage* (Vol. 29). John Wiley & Sons.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change*, *49*(3), 237-255.
Ravikumar, S., Agrahari, A., & Singh, S. (2015). Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics (2005–2010). *Scientometrics*, *102*(1), 929-955.
Ristuccia, C. A., & Solomou, S. (2014). Can general purpose technology theory explain economic growth? Electrical power as a case study. *European Review of Economic History*, *18*(3), 227-247.
Rosenberg, N., & Trajtenberg, M. (2004). A general-purpose technology at work: The Corliss steam engine in the late-nineteenth-century United States. *The Journal of Economic History*, *64*(1), 61-99.
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827-1843.
Schultz, L., & Joutz, F. (2010). Methods for identifying emerging General Purpose Technologies: a case study of nanotechnologies. *Scientometrics*, *85*(1), 155-170.
Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2009). Detecting emerging research fronts in regenerative medicine by citation network analysis of scientific publications. PICMET'09-2009 Portland International Conference on Management of Engineering & Technology,
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*(8), 1450-1467.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(19), 2464–2476.
Takeda, Y., & Kajikawa, Y. (2009). Optics: A bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics*, *78*(3), 543-558.
Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471.
Yan, E. (2015). Research dynamics, impact, and dissemination: A topic‐level analysis. *Journal of the Association for Information Science and Technology*, *66*(11), 2357-2372.
Yan, E., Ding, Y., & Zhu, Q. (2009). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics*, *83*(1), 115-131.
Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, *8*(2), 295-309.
Yang, C., Park, H., & Heo, J. (2010). A network analysis of interdisciplinary research relationships: The Korean government's R&D grant program. *Scientometrics*, *83*(1), 77-92.
Youtie, J., Iacopetta, M., & Graham, S. (2008). Assessing the nature of nanotechnology: can we uncover an emerging general purpose technology? *The Journal of Technology Transfer*, *33*(3), 315-329.
Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z., & Lu, Z. J. P. o. (2012). Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis. *PLoS One*, *7*(4), e34497.
Zhang, Y., Huang, Y., Porter, A. L., Zhang, G., & Lu, J. (2019). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change*, *146*, 795-807.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, *12*(4), 1099-1117.
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*, *to appear*.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.

Zhang, Y., Qian, Y., Huang, Y., Guo, Y., Zhang, G., & Lu, J. (2017). An entropy-based indicator system for measuring the potential of patents in technological innovation: rejecting moderation. *Scientometrics*, *111*(3), 1925-1946.
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, Vancouver, Canada.
Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179-191.
Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology*, *68*(8), 1925-1939.
Zhang, Y., Zhu, Y., Huang, L., Zhang, G., & Lu, J. (2019). Characterizing the potential of being emerging generic technologies: A methodology of bi-layer network analytics. International Conference of the International Society for Scientometrics and Informetrics, Rome, Italy.
Zhou, T., Ren, J., Medo, M., & Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, *76*(4), 046115.
Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., & Zhu, D. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics, 100*(3), 705-721.