




# Domain expertise extraction for finding rising stars

Lin Zhu<sup>1</sup> · Junjie Zhang<sup>2</sup> · Scott W. Cunningham<sup>3</sup> 

Received: 5 April 2021 / Accepted: 2 August 2022 / Published online: 22 August 2022  
© The Author(s) 2022

## Abstract

The field of expertise extraction utilizes published research enabling communities to highlight and identify the skills of researchers within specific scientific domains. This can be useful for evaluating research performance, and in the case of rising stars, in identifying top scientific talent. Previous research has harvested a range of publication indicators in an effort to identify expertise and talent. These include content indicators, citation metrics, and also the position of a researcher within a full collaboration network of scientists. The existing mechanism of expertise extraction utilizes all papers attributed to a scientific author, thereby potentially neglecting their specific or specialized expertise. Here we show that a tensor decomposition technique when applied to the problem addresses a number of useful problems. This includes better identification of individual expertise, as well as an integrated appraisal of an author's role in an extended scientific network. The technique will afford new analyses of knowledge production which consider specialisation and diversity as core elements for further analysis. More generally the tensor decomposition techniques presented in this paper can be applied to a range of scientometric problems where multi-modal data is encountered.

**Keywords** Rising stars · Expertise extraction · Tensor modelling · Individual performance

## Introduction

There are three traditions for measuring expertise. Expertise may be demonstrated through the knowledge of a domain and the use of appropriate scientific or technical language, available knowledge of a social network and one's role within that network, and the technical skills one can marshal to accomplish a task. The first tradition involves the study of co-authorship. Effective scientific collaboration draws upon multiple sources of expertise; therefore measuring collaboration is essential to understanding expertise. The second tradition involves the use of scientific indicators such as publications, and consequent derived indicators such as citation measures. Since publications are a demonstrable indicator of

---

✉ Scott W. Cunningham  
scott.cunningham@strath.ac.uk

<sup>1</sup> School of Humanities and Social Sciences, Qingdao Agricultural University, Qingdao, China

<sup>2</sup> School of Economics and Management, China University of Geosciences, Wuhan, China

<sup>3</sup> School of Government and Public Policy, University of Strathclyde, Glasgow, UK

scientific performance, there has been great interest in (and concern with) the measurement of expertise using scientific outputs such as publication. The third tradition involves the use of semantic indicators of expertise such the words or phrases present in authored texts. These elements of scientific text may provide a deeper indicator of the knowledge and disciplinary commitments of the scientists who use such language. These three traditions are reviewed in capsule form below.

These three traditions are reviewed in capsule form below, and represent a brief introduction to the concept of scientific expertise as it has been conceptualised and measured across different traditions. One tradition of measuring expertise is based on published scientific articles by making use of several performance indicators, including the quantity and quality of the document (Hammarfelt & Rushforth, 2017; Lopez-Herrera et al., 2010). Bordea argues that expertise is closely related to the notion of experience; the assumption is that the more a person works on an expertise topic, the more knowledgeable he or she is (Bordea, 2013). These researchers estimate the expertise of a person based on the number and scientific impact of the articles. Scholars have also explored many methods that might be used to evaluate research performance. For example, bibliometric indicators, like the number of published papers, citation counts, H-indexes, and journal impact factors (Basu et al., 2016; Gulbrandsen & Smeby, 2005; Kademani et al., 2007; Kotsemir & Shashnov, 2017; Lee, 2019; Panaretos & Malesios, 2009). These performance indicators are similar to the frequency indicators (Trausan-Matu & Niculescu, 2008).

Another tradition of measuring expertise is based on observed collaboration, as well as the position of the author in an extended scientific network. Quantifying the expertise of a person is not an easy task, but usually expertise is analysed in the context of an organization or community (Yeung et al., 2011). Expertise can also be measured in extrinsic ways, through the judgement of peers. This can be gathered either directly through interviews or indirectly through citations (Kavitha et al., 2014). Related works in mining expertise in social networks are as follow. For example, Ofek and Shabtai mined experts' expertise by analyzing the social network's activity information (Ofek & Shabtai, 2014). Schall developed an expertise ranking model for expertise mining, and used this model for estimating the relative importance of persons based on reputation in collaboration networks (Schall, 2012). Lappas and his colleagues explored expert's expertise by using information available in social networks, and find experts who can collaborate effectively to complete a task (Lappas et al., 2009). Serdyukov et al. acknowledge that expertise resides in knowledge networks of authorship and thereby derive propagation and assignment metrics (Serdyukov et al., 2008). Steele and Min proposed an expertise measurement based upon online professional social networks (Steele & Min, 2013). Vrabic proposed a scientific network for the structuring of the community's expertise (Vrabic et al., 2018). Other relevant work incorporates social network indicators, such as the number of co-authors, co-author citations (Ding et al., 2018; Panagopoulos et al., 2017; Silva et al., 2019; Zhang et al., 2018).

A third tradition of measuring expertise is based on the semantics of articles published by that author. In order to mine expertise content, existing studies extract expertise topics from document corpora based on latent semantics methods such as LSA, pLSI, LDA to extract information about the content of expertise (Campos et al., 2021; Lee & Kim, 2020). The information about expertise topics is further used to construct expert profiles and to find the experts (Momtazi & Naumann, 2013). In the previous works, the proposed methods for identifying expertise by semantics can be divided into two graph-based (Zhang et al., 2008) and topic modeling-based categories (Balog et al., 2009; Kichou et al., 2020; Tang et al., 2011). Gong proposed a probabilistic graphical model that estimates human expertise, and model human expertise on different topics

(Gong et al., 2018). Liu estimated user's expertise on individual topics with LDA model by document-topic relevance and user-document association (Liu et al., 2014). Liang utilized generative language model for finding knowledgeable groups that have expertise on a given a query topic (Liang et al., 2013). Li applied the biterm topic model (BTM) to model questions and fields of expertise (Campos et al., 2021). Other researchers have created rolling windows of text within documents thereby enriching the analysis of expertise using a full-text document (Petkova & Croft, 2007).

These traditions of the analysis of expertise can and should be fully applied to the field of rising stars. Expertise refers to the mechanisms underlying the superior achievement of an expert, i.e., "one who has acquired special skill in or knowledge of subjects through professional training and practical experience" (Ericsson et al., 1993). Several broad categories of expertise can be identified including cognitive expertise, and social expertise (Farrington-Darby & Wilson, 2006). Cognitive expertise refers to knowledge of a domain, while social expertise is knowledge of a social network. Cognitive expertise can be derived from documents produced by individuals, while social expertise can be mined from social networks, including co-authorship networks or citation networks (Zhang et al., 2008). Currently, rising stars are identified based on their individual performance (Daud et al., 2015; Zhang et al., 2017; Zhu et al., 2019). Current methods can be of help in identifying rising stars, although multiple problems still need addressing. One of these problem involves encoding the native expertise of rising stars. A second problem involves integrating information about co-author networks thereby deonstrating integrated measures of expertise and networked production of knowledge.

Rising stars are scholars that have achieved a high reputation and thereby are on their way to becoming experts in their respective fields in the future (Zhang et al., 2016). The discovery of rising stars is an emerging research direction which may enable research communities to better highlight the achievements of potential researchers (Daud et al., 2017). This is essential in organizations, enterprises, and academic communities. Relevant scenarios include when educational administrators hire early-career researchers, they look to build a future faculty that reflects excellence in a field. In addition, enterprises seek employees with talent that can cultivate vitality, knowledge innovation, and performance within a team.

This paper argues that new advancements are possible in the measurement of scientific expertise. The literature clearly argues for a complete and contextual understanding of scientific expertise, using a variety of different sources of information. Such information includes the co-authorship community of the scientist, the number or count of publications delivered by the scientist, and the scientific content of the authored publications. Unfortunately many existing approaches to the measurement of expertise take these elements separately, as part of a portfolio of indicators of expertise. In this work we argue that a joint measurement of expertise, jointly conditioned on authorship, co-authorship and scientific content, is needed. To meet the challeges, new approaches are needed which create a comprehensive measurement of specific expertise using all available evidence, and which more clearly distinguish between individual and team expertise. Advancements in these areas will enrich state-of-the-art efforts in developing deterministic and probabilistic knowledge graphs (Petkova & Croft, 2007; Serdyukov et al., 2008). This paper sets requirements for the appropriate representation of individual specific expertise, and proposes the use of a well-established computational method which meets these requirements. The method is relevant for expert finding applications, as well as for the potential prediction of rising stars given their publishing history. Nonetheless the underlying computational method is relevant to a range of scientometric purposes

where there are multiple document attributes which should be analysed in a reduced form.

## Methods

In this section a review of matrix and tensor decomposition methods are presented. The literature reveals that it is mathematically possible and computationally practical to decompose higher-dimensional matrices. Most importantly these techniques reduce in two dimensions to the familiar matrix decomposition techniques which are widely used in bibliometrics and information retrieve. Nonetheless the converse is not true. Generalising the two-dimension decomposition techniques to higher dimensions requires careful consideration of the joint quantities involved.

There is a long history of matrix decomposition methods used in scientometric, bibliometric and informetric application (Cunningham, 1996; Deerwester et al., 1990). The various techniques bear different names, but the underlying mathematics is often based on a linear decomposition of matrices. Relevant matrix methods include latent semantic indexing, a matrix decomposition technique used for information retrieval purposes. Other relevant techniques include principal components analysis, factor analysis and correspondence analysis which are used for visualization or science mapping purposes. These three techniques differ by the similarity metrics used, the representation of uncertainty in the data, and the specific decomposition components which are generated as results. Nonetheless these techniques are all reducible to a fundamental technique of matrix composition known as singular value decomposition (SVD). Decomposed eigenvectors are known alternatively as factors or components depending on the technique and the raw scaling of the data.

Singular value decomposition is suitable for the decomposition of matrices in two dimensions, but is actually a reduced form algorithm for the analysis of higher-dimensional objects. In scientometric application various publication metrics are extremely high dimensional, yet are converted to vector form. Because such metrics are high dimensional they present challenges for analysis, visualisation and validation. Matrix decomposition techniques which reduce the data into a lower dimensional form are often highly usable. A matrix represents a two-dimensional format for the representation of scientometric data. For instance a matrix dimensioned articles by terms may be used to represent the content of a corpus of scientific articles. Or a matrix which is dimensioned articles by authors may be used to represent co-authorship patterns.

Most importantly multiple dimensions of scientific attribution are not necessarily of like kind. For instance a document may be represented by time, authorship, citation relations, and semantic content. Although customarily represented in a series of two-dimensional forms, many scientometric matrices are intimately linked. Such linkages can be represented in three-dimensional or even higher dimensional matrix form. A third or higher dimensional matrix is known as a tensor. For instance a tensor which is dimensioned publication by author by term enables a richer representation of scientific collaboration and scientific expertise than could otherwise be obtained. This is because there are intrinsic dependencies between publications which can only be revealed by a close examination of the authorship and term matrix. Similar insights can thereby be gained about authorship and also scientific terms.

Tensors require expanded methods for analysis. Fortunately there are a family of tensor decomposition techniques closely related to the familiar matrix decomposition techniques.

The use of these techniques is proven and extensive in the field of image processing as well as psychology (De Lathauwer et al., 2000; Sheehan & Saad, 2007; Tucker, 1966). A notable example of a richer, higher-dimensional analysis is offered by Liu (2011). The authors offer an analysis of scientific content and scientific citation using tensors. This principled form of analysis enables the joint construction of models of content and citation. Content and citation are embedded in a common metric space, and information about one is used to anchor and regularize the other. The technique used in the analysis is multi-linear singular value decomposition (MLSVD). Unfortunately – and despite the analytical opportunities afforded by the use of tensor algorithms – such techniques are not yet widespread in the scientometric literature. Nonetheless these techniques enable new and richer forms of scientometric analysis.

There are multiple related tensor decomposition techniques. Within the field of tensors and their decomposition and analysis there appears to be a lack of standardization and unification across the respective disciplinary literatures. Despite this a remarkable contribution is made by Sheehan and Saad (2007). The generalized problem which underlie all of these methods according to Sheehan and Saad (2007) is the higher-order orthogonal iteration of tensors – or HOOI for short. In this paper the authors demonstrate that a variety of different decomposition techniques including 2-D principal components analysis (2-D PCA), higher-order SVD (HOSVD or MLSVD) and the generalized low-rank approximation of matrices (GLRAM) are all special instances of a more general class of tensor decomposition problems. De Lathauwer and co-authors popularized HOSVD and MLSVD (2017), and the terms appear to be synonymous. Most importantly of all these algorithms, the familiar PCA and SVD algorithms are only special cases.

This work is interested in the applications of one scientometric tensor in particular. The tensor has three dimensions consisting of documents, terms, and authors. This tensor is reducible to three separate matrices – a matrix of documents and terms, a matrix of documents and authors, and a matrix of terms and authors. These matrices and their self-products (including co-authorship, and co-word matrices) have revealed a wealth of insight into bibliometric activity. The underlying data, and various related products are shown in Table 1. This and related tables will be used to describe the analysis throughout the paper. Table 2 (in the data section below) deepens the discussion by describing the dimensions and scope of the analysis.

The joint representation of these matrices in tensor format contains rich structural information, albeit in a high dimensioned format. For this reason tensor decompositions are further pursued. The research question to be investigated is whether the full tensor of documents, words and authors contains additional information regarding the unique expertise of authors that can not be otherwise uncovered from the data analysed in reduced form. In pursuit of this question this paper examines the data reduction of word vectors, but also the data reduction of authorship matrices. Document loadings are described as content vectors.

**Table 1** Data products

Symbol	Dimensions	Explanation
$X_1$	$[d \times w \times a]$	Raw data in tensor format
$X_2$	$[d \times w]$	Derived matrix
$X_3$	$[d \times a]$	Derived matrix
$X_3^T \times X_3 \sim N$	$[a \times a]$	Matrix multiplication, resulting in co-authorship graph

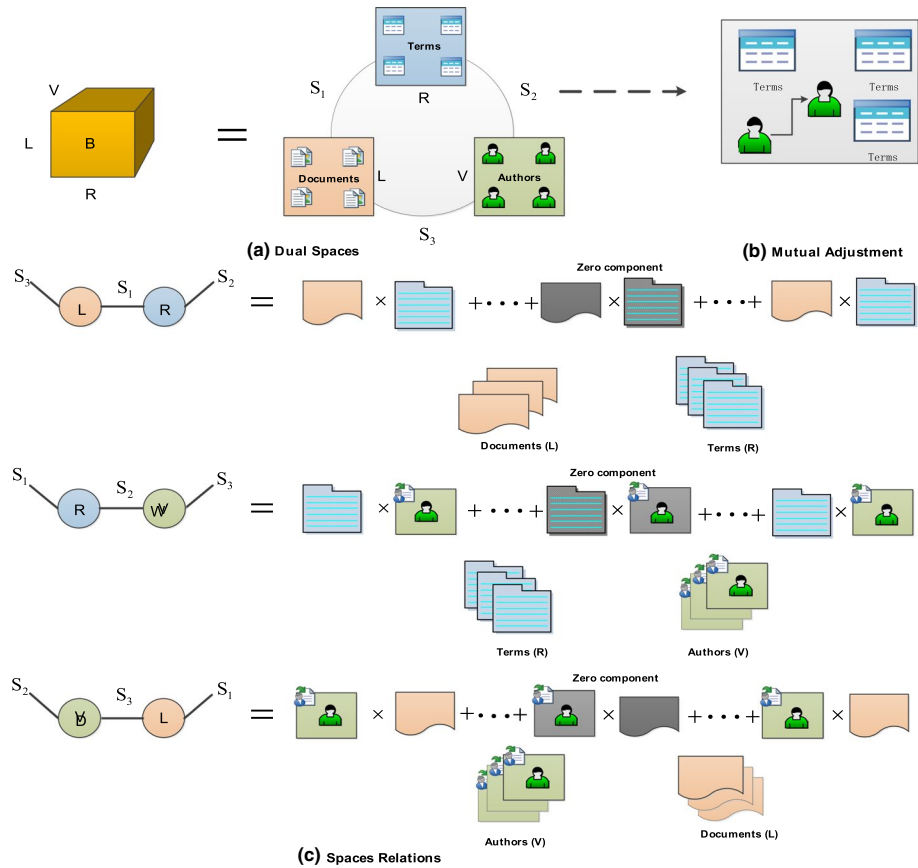
**Table 2** Dimensions of the analysis and the data

Dimension	Explanation	Full Sample	Study protocol
$w$	Number of words	1000	1000
$a$	Number of authors	21,992	100 for main analysis; 599 for visualisation
$d$	Number of documents	23,145	10,000 for main analysis; 394 for visualisation
$f$	Number of eigenvectors; possibly separately dimensioned ( $f_1, f_2, f_3$ ) for HOOI		3
$t$	Number of trials		100

The word vectors are summarily labelled semantic vectors, in keeping with long-standing practice. For conciseness the new measure explored here, a reduction of the authorship vector conditioned on documents and words, is described as expertise. This measure will be compared with a leading indicator of expertise, the aggregate or average content of the papers authored.

A decomposition of this tensor will simultaneously result in a low-ranked approximation of the underlying data (Fig. 1a). The decomposition represents the best available low-ranked approximation of the underlying data, and obeys a set of additivity constraints. A tensor of documents, authors and words can be decomposed into three separate spaces, with the requisite transformations for comparison. The separate spaces are described by the L, R and V tensors, and the core tensor used to translate between the spaces is represented as B. Each document vector is an approximate representation of the contents of the document, and each term vector is an approximate representation of underlying meaning and semantics. Each vector assigned an author may thereby be interpreted as the expertise of the author. Most notably this representation is conditioned on the measurement of expertise of all other authors in the network. Thus the assignment of expertise is made given the underlying semantics of terms and documents, as well as the assignment of expertise to all other authors and co-authors in the document corpus.

Figure 1 presents a conceptual framework describing the merits of a tensor decomposition perspective. The figure shows how there are three separate spaces resulting from the tensor decomposition pursued in this research. There is a vector space of documents, a vector space of terms, a vector space of authors. The resultant projections of these spaces involve documents  $\times$  terms, terms  $\times$  authors, and authors  $\times$  documents. The three spaces are inter-related to each other. The decomposition represents the best available low-ranked approximation of the underlying data, and obeys a set of additivity constraints. These constraints are that each document vector represents both the sum of the constituent term vectors, and the sum of the constituent author vectors. The assignment is made given the underlying semantics of terms and documents, as well as the assignment of expertise to all other authors and co-authors in the document corpus. This is known as idempotency. Figure 1b shows the adjustment of a single author so that the author sits at the average of all the terms used by this author. Because of the property of idempotency, this adjustment is made mutually between authors, terms and



**Fig. 1** Spaces, Relations and Mutual Adjustment. The figure graphically portrays the dual, idempotent spaces of a decomposed matrix in **a** at top. The spaces are mutually adjusted to the positioning of elements, shown in **b** at top. Each element of each space is assigned accordingly to the required spaces, relations and transformation rules. This is portrayed in **c** at the bottom

documents. Each is projected and scaled so that the entities sit at the average of their respective spaces. This averaging property is the result of the least-squares character of the tensor reduction. Specifically the law of cosines relates vector projections to Euclidean distances in the respective decomposed tensor spaces. These two-dimensional projections are all related to one another, so that the structure of one informs the other two, see Fig. 1c.

In this paper we investigate whether the full tensor representation contains relevant information beyond that conveyed by the reduced form matrix of documents and words. If this is the case then information about which authors work together on which manuscripts will better enable a model that will jointly estimate the expertise of the individual authors. This will further clarify the semantic representation of individual terms, and thereby generate a more appropriate classification of document content.

## Data

To demonstrate and validate our methodology, we conducted an empirical study on scholars in gene editing domain. Although we selected this domain, any other domain would be a suitable subject for our method and analysis. The analysis involves all 11,458 articles indexed in Web of Science until October 2020 under the search strategy  $TS = ((\text{Genome OR Gene OR Genetic OR DNA}) \text{ NEAR}/2 \text{ Editing})$  (Huang et al., 2019).

The dataset is created by collecting information about scholars, the titles, the scholar keywords, the keywords plus, and the full abstract and citation information from the articles. Indexing and analyses are performed in Python 3.3. Indexing is performed using the scikit-learn package `feature_extraction.text`. Stopwords (as provided by the package) are removed. The various forms of copyright used by Elsevier are removed as well. Words are reduced to lower case and punctuation is removed. No lemmatization or stemming is used. As will be demonstrating in the robustness analysis, lemmatization or stemming does not materially change the output. Using plain text eases further interpretation. The top 1000 words by total count are used to index the articles. This number of features is sufficient to produce a rich description of documents content, and to effectively place individual articles within a taxonomy of content. After indexing the documents with the 1000 terms a dense document-term matrix is produced.

Table 2 presents summary statistics of the data set. There are over 23,000 documents. In this sample nearly 22,000 unique authors are identified using the available ORCID. A vector of the top 1000 words, minus any stop words, is used to represent document content. The test is conducted with two separate samples. In the first analysis a full 10,000 of papers are investigated.

In the second analysis a smaller sample is investigated for exploratory purposes. The documents, and the uniquely identified authors of the documents, are linked in a scientific collaboration graph. There are multiple disconnected components to this graph. The largest of these components contains 5154 articles authored by 6800 uniquely identified scientists. The analysis focuses in detail on this giant component. For the purposes of this case it is helpful to take a sharp focus on smaller groups of researchers to better compare and evaluate the effects of measuring expertise using different techniques. Therefore the most centrally located researcher in the giant component is first identified. Then all researchers within three hops of the most central researcher are identified. Including the central researcher this results in a case study sample of 394 articles and 599 researchers. The structure and organisation of the graph ( $N$ ) is discussed further below in the analysis.

The use of the central-most community in this network enables an incisive exploration of expertise and its representation. Other comprehensive studies could also be designed to investigate the operation of expertise in real-world scientific communities. This research however is a feasibility test to investigate whether this new measure of expertise shows promise and should be trialed in detailed case studies. The paper returns to this point in the conclusions. A low-dimensional decomposition is sufficient to conduct this exploratory test—a three-dimensional decomposition is therefore used. Care is taken to investigate the robustness of the resultant network. One hundred trials are conducted and tests are performed to investigate the stability of the content, semantic and expertise network.



**Table 3** Analytical procedures

	Procedure	Decomposition	Explanation
1	HOOI ( $X_1, f, t$ )	$B \times_1 L \times_2 R \times_3 V_1$	Tensor decomposition
2	HOOI ( $X_1, f, t$ ) $\sim L, i$	$X_4: [f \times d \times t]$	Experimental design tensor
3	SVD ( $X_2, f$ )	$U \times W \times V_2^T$	Matrix decomposition

**Table 4** Decomposed components

	Decomposed element	Dimension	Explanation
4	$B$	$[f \times f \times f]$	Core tensor, from HOOI
5	$L$	$[f \times d]$	Tensor decomposition of content
6	$R$	$[f \times w]$	Tensor decomposition of semantics
7	$U$	$[d \times f]$	Matrix decomposition of content
8	$V_1$	$[f \times w]$	Tensor decomposition of expertise
9	$V_2$	$[w \times f]$	Matrix decomposition of semantics

## Analysis

The analysis procedure is elaborate, although the outputs are sparse. Therefore the section breaks down the analysis into a series of procedures, steps and products. These are displayed in a series of tables, and discussed individually. Table 3 displays the core analytical procedures. Table 4 displays the decomposed elements of the analysis. Derived products of the analysis are detailed in Table 5. Then testing procedures are described in Table 6.

The principle analytical procedures are matrix decomposition (SVD) and tensor decomposition (HOOI) (Table 3). Tensor decomposition is implemented using a custom made function in python, following the pseudo-code as described by Sheehan and Saad (2007). This implementation is made available by the authors (Cunningham, 2022). The function utilizes the numpy library, and the linear algebra package, and SVD subpackage (numpy.linalg.svd). The SVD procedure is suitable only for a reduced form of the data, and therefore is conducted on the document by factor matrix. The data has therefore been collapsed on the authorship dimension. The scipy library in python is used for this analysis, specifically the sparse matrix and linear algebra packages, and the svds subpackage (scipy.sparse.linalg.svds). These tables use pseudo-code, whereby analytical procedures are described as function calls with inputs and outputs.

The tensor and matrix decompositions result in multiple decomposed components of scientometric interest. These components and their explanation are discussed (Table 4). The dimensioning and symbolic representation of these components are also provided.

Three derived products are calculated from these analysis (Table 5). The first of these products is a leading procedure for evaluating the expertise of individual authors. This involves determining the topics of all the articles they have published, and then assigning each author the sum total of all papers they have individually authored, or co-authored in a team. Here the content is evaluated using the SVD procedure. This derived calculation is presented as element 10. Products 5 and 6 are projections of one matrix upon another. These are interesting elements for analysis and testing since they demonstrate potential

**Table 5** Derived products

	Procedure	Dimensionality	Explanation
10	$X_3^T \times U = V_2$	$[a \times d][d \times f] = [a \times f]$	Matrix multiplication, interpretable as the sum of published content for each author
11	$L \times U = D_1$	$[f \times d][d \times f] = [f \times f]$	Matrix multiplication, interpretable as the projection of one space of content onto another
12	$V_1 \times V_2 = D_2$	$[f \times a][a \times f] = [f \times f]$	Matrix multiplication, interpretable as the projection of summed content onto expertise

**Table 6** Testing procedures

	Procedure	Quantities	Test
13	CENT( $N$ ) $\sim c$	[ $a \times 1$ ]	Eigen- vector central- ity
14	SVD ( $D_1, f$ )	$W$ , the eigenvalues of the document embed- ding	Span
15	SVD ( $D_2, f$ )	$W$ , the eigenvalues of the expertise embed- ding	Span
16	HOOI ( $X_4, f$ )	$B$ , the core tensor	Span in the space of model runs

structural similarities or differences between the models, whether representing content, semantics, or expertise.

A number of different analytical and visual techniques are used for evaluating the results (Table 6). The first derived test involves the calculation of the network eigenvector centrality. This is also a test using SVD, in another guise. This is implemented using the network centrality measure available in the networkx package in python (networkx.algorithms.centrality.eigenvector\_centrality). Two more derived quantities are motivated by a comparison of the representation of documents and expertise generated by SVD and by HOOI. These comparisons require joint projections for further analysis, as shown previously lines 11 and 12.

Since the outputs are inherently multidimensional in character, it is also convenient to use data decomposition techniques. Thus the SVD and HOOI procedures are used both in analysis and testing, but the rationale and use of the procedures is very different. The general applicability of the techniques for a range of different testing and analysis procedures is not in this cause tautological, and does not compromise the validity of the subsequent testing. The procedures are to evaluate the span of  $D$ , the joint projections in documents and in expertise. Span is a technical property of matrices, and it represents the degree of overlap of the two matrices. The calculated eigenvectors provide an additional information, since small eigenvalues indicate the degree of shared span. Similarly the HOOI procedure is useful to evaluate whether the repeated experimental design revealed any systematic variation across testing runs.

Three hypotheses and tests are formulated to investigate this question. In the first test the latent semantic dimensions generated by SVD and HOOI are compared. The hypothesis is that there is only partial overlap between the resultant dimensions. This is significant because SVD is at best an approximate decomposition of any higher-dimensioned dataset. A lack of overlap indicates a potential threat to validity for using SVD in place of more appropriate higher dimensioned techniques. In the second test an aggregate content vector is assembled out of the results of the matrix SVD, and compared with the latent expertise vector extracted directly from HOOI. Here again the hypothesis is that the two representations of expertise partially but not completely overlap. A partial overlap is explained because the expertise vectors of SVD and HOOI are only partially related to one another.

More significantly the expertise vectors are only partially related because the assignment of expertise used by SVD and HOOI actually differ. HOOI will utilize the proven expertise of authors, demonstrated on other papers and collaborators in the database, to assign vectors of expertise. In the third test the rated expertise of authors is displayed on a collaboration graph, and contrasted with the resultant indicators of network centrality and summed content. The hypothesis is that an truly effective measure of expertise will reflect local epistemic communities.

Regardless of the outcomes of these three experiments, the contributions offered by fuller consideration of tensor decomposition techniques are of general methodological value. Tensor decomposition is of wide applicability to the field of scientometrics as a whole, and therefore there are additional contributions to be offered beyond the specifics of the case. In particular the applicability of tensor decomposition for extended scientometric analysis, regardless of the specific formulation of the tensor, has been rarely examined. The exceptional work of Liu et al. (2011) is noted above. Thus the results of this paper help advance and enrich current bibliometric practice wherever SVD and related decomposition matrices are already being used. The aforementioned analyses are conducted, the derived products calculated, and the described tests and validation procedures performed. The results are described in the following section.

## Results

The result of comparing the document embedding (described previously as test 14) is shown below. First a few words about the SVD procedure are offered. The SVD procedure addresses both the rotation as well as the scaling, of the projected matrices, before any comparison is made. Each eigenvector is scaled to 1.0. If one document vector is completely projected onto the other the resultant score would be one. If there were no overlap and the two document vectors were orthogonal from one another then the resultant vector would be zero.

The results for content suggest that there are two leading factors of content that are largely comparable, and one which is incompatible. The sum of squares of the eigenvectors indicates the share of variance preserved in the projection. Since the total variance in the two content matrices is 3.0, it is possible to calculate the percent of variance held in common, and uniquely by each of the measures. The results indicate that 61% of the variance is shared in the projected matrix, while another 39% is unique to the individual space. On the one hand these results suggest a high face validity for both measures of content, since they corroborate one another. On the other hand the techniques do not corroborate across the third factor of content (Table 7).

The evaluation is more complex for the expertise projection matrix (line 15). The summed content matrix results in varying amounts of summed content by author, and therefore is not unit scaled. The jointly projected matrix of expertise and summed content has a variance of 33.97. This variance is poorly predicted by the summed content vectors, since only 40% can be explained solely by means of the left eigenvectors representing the summed content. On the other hand the right eigenvectors representing the expertise vectors explain fully 90% of the variance in the projected matrix. These results suggest that the space of expertise can be fully expressed using transformations of content. Nonetheless summed content cannot be fully expressed using these derived measures of expertise.

This section presents a short case study drawing upon a subset of the authors and papers contained in the larger sample. The query used for the analysis is intrinsically

**Table 7** Test results

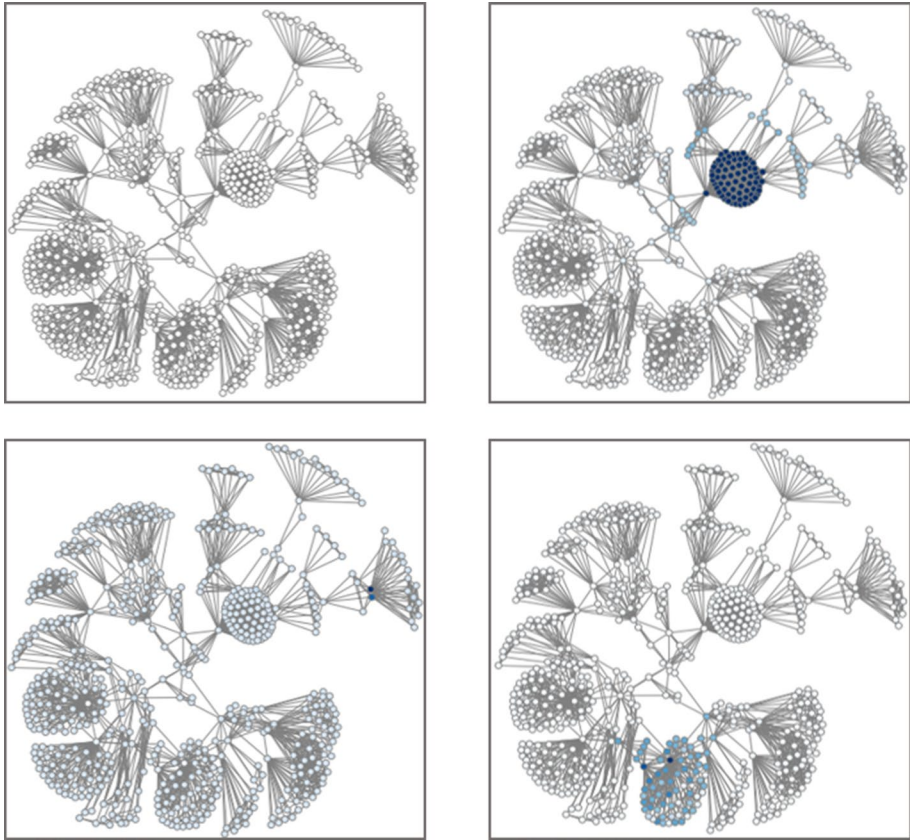
Eigenvalue source	Eigenvalues	Unique and common variance
Document projection matrix	[0.983, 0.926, 0.074]	Content: 61% of the variance is held in common and 39% is unique
Expertise projection matrix	[5.738, 0.866, 0.546]	Summed content: 40% of the variance is held in common, and 60% is unique Expertise: 90% of the variance is held in common, and 10% is unique

interesting since it has been previously and extensively studied (Huang et al., 2019; Zhu et al., 2019). The case serves an expository role by illustrating fundamental questions of the measurement of expertise. Therefore rather than characterise the entirety of the semantic and author space, a smaller case study is drawn from the data.

The study of expertise is necessarily very high dimensional. This researcher therefore uses graph overlays to communicate the relationship between expertise and scientific networks. The base graph is the collaboration network of scientists surrounding the most connected author in the network. Then three separate indicators are plotted for comparative purposes, starting with the base social network (Fig. 2a). As expected the network centrality measure (Fig. 2b) is centred around the most central actor. As a measure it rapidly diminishes at one or two hops away from the central author. As a measurement of expertise it appears inadequate overall, although it has some attractive features in demonstrating the presence of coherent research groups within the graph.

An authorship summary measure is also shown for comparison (Fig. 2c). This measure summarises the expertise of individual authors in terms of the content of all the papers which they have authored. A simple three factor decomposition is used for illustrative purposes. The first factor is displayed in the graph using a contrastive color scheme indicating the degree of expertise signalled by the particular author. This measure also appears inadequate since expertise is averaged across the network, with very few indicators of local knowledge or expertise. This runs counter to theories of the sociology of knowledge. (Note however the hotspot of unique expertise in the upper right of the Fig. 2c, demonstrating one or two authors with a sharply differing average knowledge profile than the others.)

This needs a fuller discussion and interpretation. Figure 2d shows expertise fully decomposed by paper, author and content using the HOOI technique. A simple three factor decomposition is used for illustrative purposes, and for suitable comparison with the other cases above. This measure demonstrates local communities of knowledge, as expected by theory. This group is seen at the bottom of the graph. The community indicated by the first factor is not solely the most central community, and therefore is not artefactual of network positioning. Furthermore even this community shows considerable heterogeneity in its expertise, with some members sharply focused on the specific measured attribute of expertise and other co-authors demonstrating very different sources of expertise. This case is only one part of the necessary, larger effort needed to validate this (and indeed all other) measures of scientific expertise. These efforts are further discussed in the validation and discussion sections below.



**Fig. 2** Network Positioning and Expertise. Upper left, **a** This represents the co-authorship network of the core of the graph. Upper right, **b** This represents an overlay of network centrality on the social network. Bottom left, **c** This represents an overlay of the summed content on the social network. Bottom right, **d** This represents the overlay of expertise on the social network

## Application

The concept of expertise refers to the knowledge content of specific topic of scholars, also refers to the experience of the acquirement of special skill in or knowledge of specific topic. What makes rising stars different from stars is that their experience on expertise has been on the rise in recent years. The experience on expertise can be measured by bibliometric indicators, such as the number of published papers, citation counts. We use the trend index algorithm for measuring the upward trend of expertise experience (Zhu et al., 2019). The trend indexes are constructed following the definition of rising stars: including active trends index and recent trends index. Active trends index measures the relatively continuous trends in ongoing multi-year activity of expertise experience, whereas recent trends index watches for relatively short-period, abruptly increasing activity of the expertise experience. Both trend indexes can be assessed at the productivity level and impact level. A high trend index value means the rising stars have

a significant upward trend in their research productivity and impact, therefore, they have more potential become brilliant stars in the future. The trend indexes are given by:

*Active-Trend Index*  $_{Productivity/Impact}$ : Ratio of the change in the number of publications/ citations in the last 5 years to that in first 5 years.

$$\begin{aligned}
 \text{Active - Trend}_{Prod} &= \left( \frac{pub(x,t_n)}{\sqrt{pub(x,t_n)}} + \frac{pub(x,t_{n-1})}{\sqrt{pub(x,t_{n-1})}} + \dots + \frac{pub(x,t_{n-4})}{\sqrt{pub(x,t_{n-4})}} \right) \\
 &\quad - \left( \frac{pub(x,t_{n-5})}{\sqrt{pub(x,t_{n-5})}} + \dots + \frac{pub(x,t_{n-9})}{\sqrt{pub(x,t_{n-9})}} \right) \\
 \text{Active - Trend}_{Imp} &= \left( \frac{Cit(p_{x_{t_n}}, t_n)}{\sqrt{Cit(p_{x_{t_n}}, t_n)}} + \frac{Cit(p_{x_{t_{n-1}}}, t_{n-1})}{\sqrt{Cit(p_{x_{t_{n-1}}}, t_{n-1})}} + \dots + \frac{Cit(p_{x_{t_{n-4}}}, t_{n-4})}{\sqrt{Cit(p_{x_{t_{n-4}}}, t_{n-4})}} \right) \\
 &\quad - \left( \frac{Cit(p_{x_{t_{n-5}}}, t_{n-5})}{\sqrt{Cit(p_{x_{t_{n-5}}}, t_{n-5})}} + \dots + \frac{Cit(p_{x_{t_{n-9}}}, t_{n-9})}{\sqrt{Cit(p_{x_{t_{n-9}}}, t_{n-9})}} \right)
 \end{aligned}$$

*Recent-Trend Index*  $_{Productivity/Impact}$ : Ratio of the change in the number of publications/ citations between the most recent 2 years and the prior 2 years.

$$\begin{aligned}
 \text{Recent - Trend}_{Prod} &= \left( \frac{pub(x, t_n)}{\sqrt{pub(x, t_n)}} + \frac{pub(x, t_{n-1})}{\sqrt{pub(x, t_{n-1})}} \right) - \left( \frac{pub(x, t_{n-2})}{\sqrt{pub(x, t_{n-2})}} + \dots + \frac{pub(x, t_{n-3})}{\sqrt{pub(x, t_{n-3})}} \right) \\
 \text{Recent - Trend}_{Imp} &= \left( \frac{Cit(p_{x_{t_n}}, t_n)}{\sqrt{Cit(p_{x_{t_n}}, t_n)}} + \frac{Cit(p_{x_{t_{n-1}}}, t_{n-1})}{\sqrt{Cit(p_{x_{t_{n-1}}}, t_{n-1})}} \right) - \left( \frac{Cit(p_{x_{t_{n-2}}}, t_{n-2})}{\sqrt{Cit(p_{x_{t_{n-2}}}, t_{n-2})}} + \dots + \frac{Cit(p_{x_{t_{n-3}}}, t_{n-3})}{\sqrt{Cit(p_{x_{t_{n-3}}}, t_{n-3})}} \right)
 \end{aligned}$$

where  $p_{x_{t_n}}$  is the set of publications for author  $x$  in the year  $t_n$ ,  $pub(x, t_n)$  is the number of publications to author  $x$  in the year  $t_n$  and  $Cit(p_{x_{t_n}}, t_n)$  is the number of citations to author  $x$  in the year  $t_n$ .

Top 15 rising stars are identified based on the uptrends of their expertise experience. The results presented in Table 8. It provides some insight into the rising stars with upward productivity and impact trends over the last two and five years. The active-trend indexes show significant uptrends for productivity and impact of rising stars over the recent five years, they maintain these uptrends on their expertise experience over successive years. There are some rising stars, such as Anstee, Quentin M.; kim, Jin-Soo; Varshney, Rajeev K. and Specchio, Nicola, whose productivity or impact have surged in the last two years.

**Validation**

A basic robustness and testing concern for this study involves demonstrating that the data has been fully and completely indexed. Bibliometric matrices tend to be highly sparse; this is even more the case with tensor representations of the data. Table 9 displays

**Table 8** Top 15 rising stars with an upward trend on their expertise experience

Rising stars	<i>Active – Trend<sub>Prod</sub></i>	<i>Active – Trend<sub>Imp</sub></i>	<i>Recent – Trend<sub>Prod</sub></i>	<i>Recent – Trend<sub>Imp</sub></i>
Anstee, Quentin M	3.317	16.914	0.372	5.663
Kim, Jin-Soo	8.749	13.898	5.302	0.969
Dardiotis, Efthimios	6.982	10.214	0.255	3.027
Varshney, Rajeev K	5.142	9.186	2.243	2.380
Specchio, Nicola	2.650	7.705	1.822	1.410
Qasim, Waseem	2.414	7.080	0.904	1.300
Voytas, Daniel F	1.617	6.553	0.802	1.442
Guerrini, Renzo	0.732	5.637	0.586	1.688
Bush, Stephen J	3.529	4.631	0.822	1.325
Mussolino, Claudio	2.968	4.830	0.586	1.200
Scala, Marcello	4.474	3.284	0.768	1.879
Has, Cristina	2.132	3.790	0.268	0.793
Striano, Pasquale	1.915	1.973	1.332	2.335
Kullmann, Dimitri M	1.453	1.613	0.588	1.064
Opriessnig, Tanja	2.025	1.067	1.817	0.467

some of the key characteristics of the data including words, ORCID, and identified and unidentified authors. The table reveals an adequate coverage of the data (Table 9).

One potential validity concern for the analysis is the fact that many authors are unascrbed in the analysis since they lack an ORCID. This choice is perhaps less distorting than the alternative, which is to impute unique authorship from some combination of first name, last name, and perhaps city or institution. This is because some imputations of authorship also use expertise, which would be unacceptable for the purpose of this paper. Nonetheless the consequences of such missing authorship data are potentially consequential. Concerning cases could occur if authors are concentrated in particular areas of expertise, or in one or more critical communities of collaboration.

The principal validation concern for this study lies in the strategy used for indexing words and phrases. The concern may be that the semantic factors underlying the documents are not robust. Thus the introduction of a handful of terms may cause the solution to pivot to other solutions, or otherwise introduce noise into the indexing. In order to test this concern, one hundred separate SVDs are performed on the baseline document-term matrix. On each of these runs, one hundred of the terms are dropped at random from the thousand terms used to index matrix. The corresponding counts are then zeroed out of the matrix before decomposition. In each run a three factor solution

**Table 9** Descriptive statistics

Numbers	5%	50%	average	95%
Indexed words per document	6	78	74.0	131
ORCID per document	0	1	1.5	14
Unascrbed authors per document	0	3	4.6	12
Documents per ORCID	1	1	1.6	4



is extracted from the data, and the resultant tensor of documents, loadings, and model runs is recorded and analysed.

The resultant solution suggests that any SVD or HOOI decomposition of this data will be highly robust. There were no systematic variances in document loadings not otherwise explainable by factors flipped or rotated in the shared semantic space. The ratio of the first two eigenvectors is 1000 to 1, suggesting the vast majority of the variance in the data is explainable by a shared, low dimensional solution. The robustness of the solution to perturbations of the data is initially surprising, but it is well-known that the matrix decomposition method is an effective means of data and noise reduction.

There are of course other validity concerns. One concerns the appropriateness of a low dimensional representation of the space in terms of terms, or factors. This important concern is shared across the field of bibliometrics. Another concern is the nature of the exploratory case study used here, where the giant component was isolated and analysed and the rest of the author graph is set aside. This decision, made for computational and case study reasons, does not sacrifice the validity of the resulting model. Each of the authorship graph components is separately decomposed with their own eigenvalues and eigenvectors. This means that while one component of the graph cannot fully inform the other, it also does not alter or distort the findings of each of the components. Note also that all other components of the co-authorship graph are much smaller than the giant component.

In addition to statistical conclusion validity, any validation test should also be concerned with concept validity. That is to ask whether or not the concept being tested is suitable for purpose. Any algorithm designed to attribute expertise to authors should obey several desirable properties. The algorithm should be *efficient*, *symmetric*, *linear*, and subject to the *null player* constraints. An efficient algorithm given evidence of scientific expertise allocates the entirety of that credit to the authors on the paper. A symmetric algorithm treats two authors with an identical publication record identically. A linear algorithm is both additive and multiplicative in its credit. That is to say authors of a manuscript with twice the evidence of expertise should receive twice the credit; likewise there should be no difference in credit received if the same evidenced content were distributed across multiple manuscripts. The null player constraint indicates that an author should not receive credit when there is greater evidence of substantive contribution by other authors. These four properties are familiar to the Shapley value for fair allocation of resources.

These four properties are familiar to the Shapley value for fair allocation of resources (Shapley, 1952). The Shapley value is a Nobel-prize winning mediation mechanism for fairly distributing the proceeds of joint effort. The analogy applies here if a publication is considered a joint effort, subject to the contributing efforts of each of its constituent authors. Effort cannot be demonstrated, but as previously discussed, the proven publication histories of the authorship team can be.

The Shapley value is the only available allocation procedure satisfying these four properties. Subdivision of expertise according to a HOOI decomposition satisfies the Shapley value. Like the Shapley value it effectively averages through all expertise of all publications of which the author has been involved. Perhaps just as importantly the Shapley value does not credit the author for any expertise not evidenced by their own work. On the contrary at least one of the available alternatives, the total weighting of papers, will fall awry of this null player constraint. This method will apportion full credit to all authors even if there is substantial prior publication evidence of sole expertise by one of the co-authors.

## Discussion and conclusion

The testing procedures reveal a number of insights about content and expertise. It is clear from the results that in this study a distinct factor is captured from the data which is not available from SVD. This quantity, which we choose to call expertise, maps closely to established measures of document content and word semantics. Its distribution in the scientific network under investigation meets sociological expectations, while the aggregate measure of content does not. The aggregate content measure introduces additional variance not otherwise explainable as expertise. The algorithm used to derive expertise displays valuable properties of fair assignment.

This paper applied new techniques for identifying expertise to the practice of identifying rising stars. The method effectively partitioned a research community into separate groups, demonstrating heterogeneity in scientific content and co-authorship patterns. We believe it is important to use expertise rather than a naïve rating of publication output since different communities will have different practices of output and co-authorship.

Furthermore expertise measures may help unethical practices of co-authorship, including the awarding of ghost, gift or honorary authorship, which may distort rising star measures. This is the primary effect of incorporating better measures of expertise in the identification of rising stars. Nonetheless a range of secondary models may be considered in future research. It is possible to create tensor models which rate changes in author expertise over time. Such models could be built on an extended tensor of publication by author by content by year. Such models may help model the propensity of publication as well as the growth and diminishment of expertise. The resultant expertise score might be used more directly in the reporting of rising stars.

It will be attractive to consider neural architectures for this and related problems. Neural networks are procedures for implementing algorithms, rather than a distinct class of algorithm in their own right. Neural networks have the advantages of being expressed in a highly parallelizable format, which can therefore be tasked to multiple cores or graphical processing units. Tensors (as used in this paper) are the common data structure underlying many popular neural network libraries, such as TensorFlow. The HOOI algorithm as used here is a simple extension of existing architectures within a neural network framework. The HOOI algorithm is most closely related to a class of neural network architectures known as transformers (Vaswani et al., 2017). Transformers have been credited with remarkable advances in transfer learning in recent years. Transfer learning involves applying the knowledge discovered in one domain to another very different domain. In this scientometric application the HOOI algorithm creates a shared representation between authors, keywords, and papers. The HOOI algorithm in three dimensions entails one input layer, one multidimensional hidden layer (representing the core tensor), and three output layers (representing each of the three transformers).

We foresee four areas for continued research in the space of tensor applications in scientometrics. The first is to examine parallel or neural architectures as described above. The second is to explore probabilistic algorithms and to deepen the understanding of conditional and unconditional independence of entities in text (Petkova & Croft, 2007). A third area for investigation is to examine other higher-order representations, including those using time or citation information. A fourth direction is to advance the science of science. This requires a deeper analysis of specific case studies of scientific communities in action in order to better disentangle relationships between scientific communities, knowledge and organization.

A potential drawback of the HOOI procedure is the fact that it is so memory intensive. Conducting the procedures described in this paper for instance, required storing not one matrix but the equivalent of one matrix of content for every author considered in the sample. Nonetheless such procedures can be moved to the cloud where more memory is readily available. Furthermore the iterated least squared procedures of Sheehan and Saad (2007) can be adapted to online variants which utilizes a window in the dataset and operates through successive passes of the data. Sparse matrix formats and algorithms are another potential solution to this problem. These algorithms may also be comparatively easy to implement using the successive least-squared approximation procedure described by Sheehan and Saad (2007).

The contributions of this paper are as follows. First we advance and enrich current practice where SVD and related algorithms on matrices are already being used. This paper advances and enriches current scientometric practice where SVD, and where multivariate extensions such as HOOI may be applied. This paper, as well as the methodologically related work of Liu et al. (2011) presents one of the few papers in the scientometric literature to apply tensor decomposition. We argue that a vector decomposition of the data results in manuscripts being placed at the average estimated expertise of all the contributing authors. This is a contribution to separating individual and team expertise. Further instruments are hereby created using the methods of this paper for the better understanding of scientific collaboration about expertise. The fact that the co-authorship dimension of the tensor should meaningful has been hypothesized by many, yet a joint appraisal of authorship and expertise has been challenging.

The results presented in this paper have important implications for expert identification and retrieval, and for the finding of new and rising scientific stars. The method applied in this paper will also be useful for analysing the structure and content of scientific fields, and may help in the development of new theories of scientific knowledge production. Future work may examine the role of scientific generalists, whose knowledge spans multiple fields, as well as scientific specialists, who contribute deeply to a single field. Generalists and specialists may play very different roles within teams and across organisations, and they may experience very different career trajectories.

**Acknowledgements** The authors appreciate the commons of four anonymous reviewers and those of editors whose efforts helped us to greatly improve the manuscript.

**Funding** The authors declare no research funding.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, 45(1), 1–19.
- Basu, A., Banshal, S. K., Singhal, K., & Singh, V. K. (2016). Designing a Composite Index for research performance evaluation at the national or regional level: Ranking Central Universities in India. *Scientometrics*, 108(3), 1695–1697.
- Bordea, G. (2013). *Domain adaptive extraction of topical hierarchies for Expertise Mining*.
- Campos, L. M., Fernandez-Luna, J. M., Huete, J. F., & Redondo-Exposito, L. (2021). LDA-based term profiles for expert finding in a political setting. *Journal of Intelligent Information Systems*, 56(3), 529–559.
- Cunningham, S. W. (1996). *The content analysis of British scientific research*. University of Sussex.
- Cunningham, S. W. (2022). *Python implementation of the HOOI Algorithm*. Retrieved from <https://github.com/cunninghamsw/HOOI/tree/main>
- Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. R. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, 102(2), 1687–1711.
- Daud, A., Aljohani, N. R., Abbasi, R. A., Rafique, Z., Amjad, T., Dawood, H., & Alyoubi, K. H. (2017). Finding rising stars in co-author networks via weighted mutual influence. *International Conference on World Wide Web Companion* (pp. 33–41).
- De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications*, 21, 1253–1278.
- Deerwester, S., Dumais, S. T., Fernas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Ding, F., Liu, Y. Q., Chen, X., & Chen, F. (2018). Rising star evaluation in heterogeneous social network. *IEEE Access* (pp. 29436–29443).
- Ericsson, A. K., Krampe, R. T., Tesch-Romer, C., Ashworth, C., & Schneider, V. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406.
- Farrington-Darby, T., & Wilson, J. R. (2006). The nature of expertise: A review. *Applied Ergonomics*, 37(1), 17–32.
- Gong, S.-S., Hu, W., Ge, W.-Y., & Qu, Y.-Z. (2018). Modeling topic-based human expertise for crowd entity resolution. *Journal of Computer Science and Technology*, 33(6), 1204–1218.
- Gulbrandsen, M., & Smeby, J. C. (2005). Industry funding and university professors' research performance. *Research Policy*, 34(6), 932–950.
- Hammarfelt, B., & Rushforth, A. D. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation*, 26(3), 169–180.
- Huang, Y., Porter, A., Zhang, Y., & Barrangou, R. (2019). Collaborative networks in gene editing. *Nature Biotechnology*, 37(10), 1107–1109.
- Kadmani, B. S., Kumar, V., Surwase, G., Sagar, A., Mohan, L., Kumar, A., & Gaderao, C. R. (2007). Research and citation impact of publications by the Chemistry Division at Bhabha Atomic Research Centre. *Scientometrics*, 71(1), 25–57.
- Kavitha, V., Manju, G., Geetha, T. V., & Ieee. (2014). Learning to rank experts using combination of multiple features of expertise. *3rd International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1053–1058).
- Kichou, S., Boussaid, O., & Meziane, A. (2020). Tag's depth-based expert profiling using a topic modeling technique. *International Journal on Semantic Web and Information Systems*, 16(4), 81–99.
- Kotsemir, M., & Shashnov, S. (2017). Measuring, analysis and visualization of research capacity of university at the level of departments and staff members. *Scientometrics*, 112(3), 1659–1689.
- Lappas, T., Liu, K., Terzi, E., & Acm. (2009). Finding a team of experts in social networks. *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 467–475).
- Lee, D. H. (2019). Predicting the research performance of early career scientists. *Scientometrics*, 121(3), 1481–1504.
- Lee, H. I., & Kim, J. W. (2020). An intelligence support system research on KTX rolling stock failure using case-based reasoning and text mining. *Journal of Intelligence and Information Systems*, 26(1), 42–73.
- Liang, S. S., de Rijke, M., & Assoc Comp, M. (2013). Finding Knowledgeable Groups in Enterprise Corpora. *36th ACM SIGIR Annual International Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 1005–1008).
- Liu, X., Glanzel, W., & De Moor, B. (2011). Hybrid clustering of multi-view data via Tucker-2 model and its application. *Scientometrics*, 88(3), 819–839.

- Liu, X., Wang, G. A., Johri, A., Zhou, M., & Fan, W. (2014). Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Information Systems Frontiers*, 16(4), 715–727.
- Lopez-Herrera, A. G., Cobo, M. J., Herrera-Viedma, E., & Herrera, F. (2010). A bibliometric study about the research based on hybridating the fuzzy logic field and the other computational intelligent techniques: A visual approach. *International Journal of Hybrid Intelligent Systems*, 7(1), 17–32.
- Momtazi, S., & Naumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, 3(5), 346–353.
- Ofek, N., & Shabtai, A. (2014). Dynamic latent expertise mining in social networks. *Ieee Internet Computing*, 18(5), 20–27.
- Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics*, 11(1), 198–222.
- Panaretos, J., & Malesios, C. C. (2009). *Influential Mathematicians: Where do they come and where do they go?* Berlin: International Statistical Institute.
- Petkova, S., & Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. *CIKM '07: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (pp. 731–740).
- Schall, D. (2012). Expertise ranking using activity and contextual link measures. *Data & Knowledge Engineering*, 71(1), 92–113.
- Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. *CIKM 08: Proceedings of the 17th ACM Conference on Information and Knowledge Mining*.
- Shapley, L. S. (1952). *Notes on the n-person game: Value of an n-person game*. RAND Corporation.
- Sheehan, S., & Saad, Y. (2007). Higher order orthogonal iteration of tensors (HOOT) and its relation to PCA and GLRAM. *Proceedings of the 2007 SIAM International Conference on Data Mining* (pp. 355–365).
- Silva, F. S. V., Schulz, P. A., & Noyons, E. C. M. (2019). Co-authorship networks and research impact in large research facilities: Benchmarking internal reports and bibliometric databases. *Scientometrics*, 118(1), 93–108.
- Steele, R., & Min, K.-H. (2013). Towards capturing population-wide expertise via online professional social network systems. *2nd International Conference on Information Technology and Management Innovation (ICITMI 2013)* (pp. 115–124).
- Tang, J., Zhang, J., Jin, R. M., Yang, Z., Cai, K. K., Zhang, L., & Su, Z. (2011). Topic level expertise search over heterogeneous networks. *Machine Learning*, 82(2), 211–237.
- Trausan-Matu, S., & Niculescu, C. (2008). *A framework for an ontology-based information system for competence management*. economyinformatics.ase.ro.
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279–311.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.
- Vrabic, R., Kozjek, D., Ozturk, E., Tunc, L. T., Malus, A., & Butala, P. (2018). Identification of the CIRP expertise network based on public data. In *51st CIRP Conference on Manufacturing Systems (CIRP CMS)* (pp. 165–168).
- Yeung, C. M. A., Noll, M. G., Gibbins, N., Meinel, C., & Shadbolt, N. (2011). SPEAR: Spamming-resistant expertise analysis and ranking in collaborative tagging systems. *Computational Intelligence*, 27(3), 458–488.
- Zhang, C. X., Liu, C., Yu, L., Zhang, Z. K., & Zhou, T. (2017). *Identifying the academic rising stars via pairwise citation increment ranking*, *Web and Big Data* (pp. 475–483).
- Zhang, J., Ackerman, M. S., & Adamic, L. (2008) WWW 2007 / Track: E<sup>2</sup>-applications session: E-communities expertise networks in online communities: Structure and algorithms.
- Zhang, J., Ning, Z. L., Bai, X. M., Wang, W., Yu, S., & Xia, F. (2016). *Who are the rising stars in academia?* (pp. 211–212). Digital Libraries.
- Zhang, J., Xu, B., Liu, J. Y., Tolba, A., Al-Makhadmeh, Z., & Xia, F. (2018). PePSI: Personalized Prediction of Scholars' impact in heterogeneous temporal academic networks. *IEEE Access*, 6, 55661–55672.
- Zhu, L., Zhu, D. H., Wang, X. F., Cunningham, S. W., & Wang, Z. N. (2019). An integrated solution for detecting rising technology stars in co-inventor networks. *Scientometrics*, 121(1), 137–172.