Library Hi Tech

Number 1

Linked open data of bibliometric networks: analytics research for personalized library services
Guest Editors: Miltiadis D. Lytras, Saeed-Ul Hassan and Naif Radi Aljohani

1 Editorial advisory board
2 Guest editorial
8 Correlational analysis of topic specificity and citations count of publication venues
Ali Daud, Tehmina Anjaj, Muazzam Ahmed Siddiqui, Naif Radi Aljohani, Rabeeh Ayaz Abbasi and Muhammad Antisham Aslam
19 Scientific collaboration networks in Pakistan and their impact on institutional research performance: a case study based on Scopus publications
Fahad Sabah, Saeed-Ul Hassan, Amina Muazzam, Sehrish Iqbal, Saira Hanif Soroya and Raheem Sarwar
30 Blockchain and OECD data repositories: opportunities and policymaking implications
Miguel-Angel Sicilia and Anna Visvizi
43 Trend analysis of co-authorship network in Library Hi Tech
Fei-Fei Cheng, Yu-Wen Huang, Der-Chian Tsaih and Chin-Shan Wu
57 Tolerance analysis in scale-free social networks with varying degree exponents
Kwok Tai Chui and Chien-chen Shen
72 More descriptive norms, fewer diversions: boosting Chinese researcher performance through social media
Wenfeng Zheng, Yenchun Jim Wu and Yue Lv
88 Bibliometric networks and analytics on gerontology research
Chien-chen Shen, Duong Tuan Nguyen and Po-Yu Hsu
101 Indexing Arabic texts using association rule data mining
Ramzi A. Haraty and Rouba Nasralah
118 Supply chain management scholar’s research impact: moderated mediation analysis
Yucheng Zhang, Yenchun Jim Wu, Mark Goh and Xinhong Liu

ISBN 978-1-78973-005-0
www.emeraldinsight.com/loi/lht
EDITORIAL ADVISORY BOARD

Kathrin Ambrozic
Humboldt-Universität zu Berlin, Germany

Kenning Arlitsch
Montana State University, USA

Dom Bortruex
American University, USA

May Chang
University of California, USA

Timothy W. Cole
University of Illinois, USA

Justin de la Cruz
Atlanta University Center, USA

Dr Bradford Lee Eden
Valparaiso University, USA

Professor Ina Fourie
University of Pretoria, South Africa

Amelia N. Gibson
University of North Carolina, USA

Jane Greenberg
Drexel University, USA

Joe Matthews
JRM Consulting, Inc., USA

Eleanor Mitchell
Dickinson College, Carlisle, USA

Dr Hether Moulaison Sandy
University of Missouri, USA

Steve O’Connor
Information Exponentials Pty Ltd, Australia

Sanghee Oh
Chungnam National University, South Korea

Ed Roberts
RealNetworks Inc., USA

Dr Eva Méndez Rodriguez
Universidad Carlos III de Madrid, Spain

Professor Jennifer Rowley
Manchester Metropolitan University, UK

Joe Sanchez
City University of New York, USA

Jeffrey Trzeciak
Washington University, USA

Sarah Watstein
University of North Carolina Wilmington, USA

EMERITUS BOARD MEMBERS

Morell D. Boone
Eastern Michigan University, USA

Michael Buckland
University of California, USA

Susan Cleyle
Memorial University of Newfoundland, Canada

Charles Hildreth
Long Island University, USA

Larry A. Kroah
Trenton Free Public Library, USA

Karen Markey
University of Michigan, USA

James Rush
Consultant, USA

Dr Judith Wusteman
University College Dublin, Ireland

Sandra Yee
Wayne State University, USA
Introduction

The integration of smart library services in the urban context is a bold move toward sustainable and smart cities. The capacity of emerging technologies, such as data mining, advanced analytics, scientometrics, cognitive computing and immersive interfaces, to set up flexible, personalized spaces for content, context and community integration enhances the traditional value-adding models of libraries. In our perception, we are entering a new era of collective human wisdom, facilitated by intelligent library services promoting the role of libraries as hubs of collaboration and community enhancement.

The objective of the special issue is to communicate and disseminate recent computer engineering and library and information management research and application developments that demonstrate the capacity of emerging technologies to radically change the library experience. The thematic areas for the special issue focus are linked to the following areas of computational intelligence:

1. Smart cities research for sustainable development.
2. Community projects for smart library services in urban areas and cities.
3. Personalized library services and advanced analytics including:
   - social network analysis in publications;
   - automatic semantic annotation of bibliographical resources;
   - citation networks;
   - rising stars analysis; and
   - research matching.
4. Integration of IoT, cognitive computing and cloud services.
5. Design of context-aware, ubiquitous library interfaces.
6. Promotion of international collaboration through knowledge management and know-how transfer.

The next generation library systems will face tremendous changes due to their integration with complementary domains and applications. From solutions related to unique learning experiences to fully functional mobile marketplaces of micro-content exchange and push content applications, a new era of library experience that will offer the potential to promote library services as transparent value-adding integrators has begun. In this context of smart cities-enabled library services, there are requirements for content development, new standards and innovative strategies and models for personalization, unique value proposition and innovations in use cases for new library services in large context.

Current applications of smart library services worldwide present a very interesting picture. Most of the dominant providers provide advanced content management services and context-aware solutions with limited integration to other domains. In parallel, innovative business models, methodologies and frameworks for smart cities promote the critical role of library services as boosters for the use and integration of scientific knowledge to innovation and training.
Personalized library services as an effective content management channel, initially converging with but then perhaps moving beyond social networks and mobile and ubiquitous technologies, offer added value to the library experience.

Several large-scale systems already provide a range of services to different stakeholders in the publishing and library domain, including publishers, editors, students, academia, training institutions and innovation centers. A key challenge for smart library systems to become more mainstream in this area is for traditional urban context to be critically enriched by technology-enabled components offering smart library experiences in the context of a greater vision for an International Global, Sustainable Smart Planet. The potential impact of analytics research for “Personalized library services: a smart cities primer” is reflected in surveys that indicate that the integration of data mining, IoT, cloud and cognitive computing research can radically change library business models, suggesting the capacity of the Library Hi Tech to redefine how knowledge, people and social infrastructures are integrated into the context of smart cities.

The purpose of the special issue is to present state-of-the-art approaches to, and examples of, advanced library systems and components for smart cities and urban applications. Manuscripts have been sought that address these areas. Novel approaches and sound technological solutions will be expected.

2. Unfolding the challenges for smart libraries
This edition presents key future dimensions that tap emerging tools and technologies including advance data mining and information retrieval techniques, scientometrics, social media analytics, cognitive computing for big data and deep learning, to set up flexible, personalized spaces for content, context and community integration to enhance the traditional value-adding models of libraries. In our perception, we are entering a new era of collective human wisdom, facilitated by intelligent library services promoting the role of libraries as hubs of collaboration and community enhancement.

The web is emerging as a preferred platform to publish open data and to interlink it together. This has result in the Web of Documents to be emerged as Web of Data. The emergence of the Web of Data is termed as linked open data (LOD), which consists of interlinked open data sets that are being published in different domains such as government, news, health and geography. Also, the trend of publishing LOD overcomes the limitations of publishing data from a bounded group of domain-specific data access boundaries to global data space. Even though the trend of publishing LOD has got potential in several domains such as government, music and health, but data from various domains such as digital libraries (and digital documents in general), scientific publications, archives and scholarly communications still need to be processed to make it part of global data space. If we consider only scientific publications/documents, the data about these millions of scientific documents that are published by different publishers are either untouched or a minor fraction of it is added in LOD cloud. The main reason is that data about scientific documents are published as a bounded group of publisher-specific templates. The same is true for data contained in digital documents, archives and scholarly contributions. This opens doors for new methodologies, processes, tools and framework to be developed which can deal with challenges of data extraction, processing, interlinking and publication of data from human to machine understandable format.

Given the better accessibilities of LOD and our abilities, as an information scientist, to process heterogeneous large-scale data sets more effectively, the field of library and information sciences (LIS) is currently at ripe for exploration. We believe that the future of LIS would stand on the following research agendas (see Figure 1): better understanding of social usage indices under the umbrella of Altmetrics to disseminate research outcomes; bibliometric network to mine complex networks including social media and LOD; information retrieval model by extracting textual and non-textual meta-data from full-text
digital archives; improved methods of co-citation analysis, collaboration patterns, semantic analysis, etc. to support traditional bibliometrics; and last but not the least, exploiting the role of block chain-related technologies to ensure better data security and integrity of data used for evidence-based policy making.

First, under the umbrella of Altmetrics (Priem et al., 2010), the scientific community would continue to explore the potential of social media contents and to understand better the power of evolving social media platforms. The studies have shown that arts and humanities researchers publish more diverse, non-article outputs than their peers in the sciences: books, book chapters, exhibition catalogs, textual data sets, book reviews, images, videos, audio files and many more. When it comes to research evaluation, humanities researchers are often not concerned with citations counts for their work. Instead, measures of impact in the humanities might include: prestige of publisher for one’s book, peer-review status of work, prizes and awards received, mass media contributions or the number of national or international presentations completed.

The second key future direction would be the mining of complex networks generated by bibliography data set, social media connections or LOD (Jett et al., 2017). Recent statistics reveal that bibliometric big data consist of over 20,000 journals along with over 0.9bn citations and 35m publications in Scopus. With the efforts of scientific community, a number of open source tools are now available to mine large-scale networks at different level of details such as citation network of publications/authors/journals; co-authorship network of authors/organizations; co-citation network of publications/authors/journals; bibliographic coupling network of publications/authors/journals or co-occurrence network of keywords/terms. For this analysis, the following open source tools are among the best choices and continuously improving: VOSviewer and CitNetExplorer (van Eck and Waltman, 2017), Gephi (Heymann, 2017) and Map Equation (www.mapequation.org). These tools have advanced visualization capabilities such as smart labeling algorithms, overlay visualization, built-in support for popular bibliography databases, text mining functionality, layout and clustering techniques. Overall, the advancements in mining large-scale networks would continue to increase for the better understanding of complex networks for informed decision making.

The third key direction of LIS would seek to provide better searching capabilities to users by coupling textual and non-textual meta-data from full-text digital archives (Cabanac et al., 2018). Retrieval evaluations have shown that simple text-based retrieval methods scale up well but do
not progress. Traditional retrieval has reached a high level in terms of measures like precision and recall, but scientists and scholars still face challenges present since the early days of digital libraries. Thus, there is a sheer need in this direction, i.e., to identify relevant features from textual and non-textual meta-data items such as figures, tables, algorithms, etc., using unsupervised and supersized machine learning models, to build improved information retrieval techniques for digital libraries where simple text-based retrieval methods fail. Furthermore, another yet to be explored area of research is mining large-scale data sets such as digital libraries and books to create semantic abstractions for the better understanding of human-created knowledge.

Despite the fact that LIS community is leaping forward to explore new kind of data sources coupled with new tools and technologies, the traditional bibliometric analysis would keep playing a vital role to quantitatively assess the academic prestige of publication venues or author. We also expect to see more advanced bibliometric indices that utilize the advancements in context-based citation analysis, replacing traditional absolute citations-based indices such as impact factor or h-index. Moreover, we also foresee improved methods of co-citation analysis that taps the power of full-text for improved clustering accuracy of scientific publications.

Last but not the least, in the context of addressing basic contemporary societal concerns, such as transparency, accountability and trust in the policy-making process, LIS community would seek to exploit the role of blockchain technology in the near future to ensure better data security and integrity of data used for evidence-based policy making (Sicilia and Visvizi, 2019).

3. Overview of the published research in the special issue

This special issue publishes high-quality, original research contributions that address challenges in extraction, processing, interlinking and publishing data from different sources, specifically focusing on the data in the digital libraries, scientific publications, citation indexes, archives and scholarly communications that are jointly called bibliometrics networks. Undoubtedly, the role of a social media platform in relation to boost academic performance is vital with the emergence of Web 2.0 technology. However, social media platforms would vary across countries. Cheng et al. (2019) and Zheng et al. (2019) argue that in contrast to global acceptance of social media platforms, Chinese social media platforms are actually insufficient to support academic research, thus, make it difficult for the scholars to enhance the outreach of their scientific achievements.

Over decades, bibliometric studies that seek to investigate collaborative patterns have always been a point of great interest due to an implicit knowledge exchange between the collaboration. Cheng et al. (2019) and Zheng et al. (2019) show that more than half of the papers in the field of LIS, more specifically in Library Hi Tech journal between 2006 to 2017, have been written by more a single author. Similarly, Sabah et al. (2019) measure the effectiveness of scientific collaboration in terms of knowledge transfer from technologically advanced countries to developing counties. They present comprehensive use cases of top 50 research active Pakistani institutions in relation to their research outcomes with that of their international collaboration linkages with European and North American institutions. They show that a careful selection of collaborative partners can not only increase research output but it can significantly increase the quality of research.

Keeping in view that around 420m people in the world speak Arabic language, efforts need to be done in the context of building automated modeling techniques to improve digital access of scientific literature in under resource language like Arabic. An interesting work by Haraty and Nasrallah (2019), in this lieu, presents an auto-indexing modeling approach that employs association rule-based data mining techniques. Such models are extremely helpful to enhance the searching capabilities of search engines to facilitate user search needs.

Chui and Shen (2019) present tolerance analysis in scale-free social networks with varying degree exponents. Using the five state-of-the-art network indices, the authors...
analyze conditions where the degree exponents may lie outside the range, i.e. [2, 3], in a scale-free network, named, average density, average clustering coefficient, average path length, average diameter and average node degree.

Finally, Sicilia and Visvizi (2019) propose the first blueprint of a form of sharing that complements open data practices with the decentralized approach of blockchain and decentralized file systems. They argue that the native decentralized (peer to peer) blockchain model offers encryption and validation that ensures the data not to be altered. Since the data are decentralized, cross-checked by the whole network and encrypted, it makes virtually impossible to hack data shared among the organizations to ensure better data security and integrity of data used for evidence-based policy making.

4. Conclusions
The evolution of smart services in modern libraries will continue to evolve. The technological enablers will be linked to the emerging technologies of artificial intelligence, cognitive computing, internet of things and analytics. Within this context a lot of work must be done in the policy making strategic level including:

- Design and implementation of smart cities and smart villages services to promote open access to knowledge and content (Visvizi and Lytras, 2018a, b).
- Integration of added to information systems in terms of perceived usefulness from real users (Lytras and Visvizi, 2018).
- Promotion of sophisticated technologies including cognitive computing and artificial intelligence, as carriers of advanced security, trust and ease of use (Lytras et al., 2017).
- Understanding of advanced publication analytics in social impact terms. The quest of high-index values in current approaches, e.g., impact factor, h-index, must be rational and beyond manipulation strategies to promote the real social value of publications for sustainable socioeconomic growth.

We are happy to deliver this edition to the Library High Tech academic communities. We want to thank the editor in chief of LHT, Professor Michelle M. Kazmer, Florida State University School of Information: Florida’s iSchool, USA, for the opportunity she offered us to serve the LHT community as well as the Emerald people for their continuous support and commitment in this project. Finally we are obliged to our distinguished contributors and reviewers for their excellent and high intellectual work. We do believe that the readers of LHT will value this special issue.

Miltiadis D. Lytras
The American College of Greece, Athens, Greece

Saeed-Ul Hassan
Department of Computer Science, Information Technology University, Lahore, Pakistan, and

Naif Radi Aljohani
Faculty of Computing and Information Technology, Jeddah, Saudi Arabia

References


Further reading


Correlational analysis of topic specificity and citations count of publication venues

Ali Daud and Tehmina Amjad  
IIUI, Islamabad, Pakistan
Muazzam Ahmed Siddiqui  
King Abdulaziz University, Jeddah, Saudi Arabia
Naif Radi Aljohani  
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
Rabeeh Ayaz Abbasi  
Department of Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia and Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan, and
Muhammad Ahtisham Aslam  
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

Purpose – Citation analysis is an important measure for the assessment of quality and impact of academic entities (authors, papers and publication venues) used for ranking of research articles, authors and publication venues. It is a common observation that high-level publication venues, with few exceptions (Nature, Science and PLOS ONE), are usually topic specific. The purpose of this paper is to investigate the claim correlation analysis between topic specificity and citation count of different types of publication venues (journals, conferences and workshops).

Design/methodology/approach – The topic specificity was calculated using the information theoretic measure of entropy (which tells us about the disorder of the system). The authors computed the entropy of the titles of the papers published in each venue type to investigate their topic specificity.

Findings – It was observed that venues usually with higher citations (high-level publication venues) have low entropy and venues with lesser citations (not-high-level publication venues) have high entropy. Low entropy means less disorder and more specific to topic and vice versa. The input data considered here were DBLP-V7 data set for the last 10 years. Experimental analysis shows that topic specificity and citation count of publication venues are negatively correlated to each other.

Originality/value – This paper is the first attempt to discover correlation between topic sensitivity and citation counts of publication venues. It also used topic specificity as a feature to rank academic entities.

Keywords Academic libraries, Entropy, Citation analysis, Indexing, Bibliometric networks, Publication venues

Paper type Research paper

1. Introduction

Citation analysis has been vastly studied to identify the impact of quality of research; specifically, its impact is analyzed for ranking academic entities. Number of citations is the most viable source to identify the importance of a research paper. If paper receives more citations, then it is usually considered more important and useful than its counterparts. But from citation count one cannot determine that it is praising, discouraging or have neutral opinion about others work. Citation count is being used as major factor by many researchers to rank papers and authors. PageRank algorithm (Page et al., 1999) put forth a
base for ranking webpages with the help of in-links or citations. It not only considers the citation links but also the quality of the papers doing these citations in an iterative process of convergence of algorithm. PageRank has been extensively used in bibliometric networks for performing different tasks, such as ranking papers, authors and publication venues, etc.

With time-to-time changes in PageRank, research publications, authors, venues (journals, conference and workshops) and institutes are additionally ranked on the basis of considering variables like publication count (i.e. number of papers published by an institution), citations (i.e. number of citations received) (Page et al., 1999; Rahm and Thor, 2005; Xing and Ghorbani, 2004; Yan et al., 2011), temporal dimension of citations (i.e. earlier published papers usually have numerous citations than that of recently published papers, so considering number of years is also found important) (Fiala, 2012; Shen, 1993), link popularity (Liu et al., 2012; Daud et al., 2017), number of visits of links (i.e. number of times a link is visited) (Agarwal and Agarwal, 2013; Tuteja, 2013; Tyagi and Sharma, 2012) and semantically similar cue words of uncertainty (Chen et al., 2018). Citations are considered to be a popular source for measuring the prestige of webpages. Previously, citation count and co-authorship links were the criteria used for ranking publications, authors, venues exploited in different ways, such as citation networks, co-author networks, etc.

Some existing methods use homogeneous network (e.g. only used publications or only used authors) and others use heterogeneous network (e.g. they used papers, authors and venues simultaneously). Most important factor of these methods is number of citations for a long time. Citation is a vote from a publication to another publication that depicts the quality and usefulness of work, the more the better is usual criteria. Citation provides a quantitative way to measure scientific influence, quality and impact of research. Citations can be used to measure the popularity and prestige of authors. Impact factor (IF) is a quantitative measure to evaluate the level of venue. It is a quantitative evidence for a publisher for ranking journals in different domains. But in case of missing information about citations, ranking cannot be done or in case of missing links, information ranking may be not much precise. Citations are the most used measure. Along with its popularity to get citations from large data sets is a challenging task. More importantly, research articles need some time to get cited. That is why, recently published papers have zero citations in most of the cases. And most papers while ranking academic entities have ignored the papers with no in-links.

In this paper, we investigate the importance of topic specificity of publication venues to overcome the situations where citation information is missing or incomplete and to find some text-based factor to rank academic entities. Intuitively, topic specificity can be one factor that can be used to evaluate the level of venues. Topic-specific venues are usually considered to be of high quality, while the publication venues that publish from a larger academic domain are considered to be of not high quality, although there are exceptions like Nature, Science, Cell and PLOS ONE. All of these measures have their pros and cons. Topic specificity is a text-based qualitative measure of venues that depends on titles of the papers published in a specific venue over a specific year. It is totally independent from factor of citations. But to see how important this factor can be, we will exploit the correlation between topic specificity and the citations of publication venues.

Topic specificity can be computed using entropy, which measures the distribution of a random variable. A uniformly distributed random variable gets high entropy, as all the values are equally likely. On the other hand, low entropy results from an unequal distribution. For publication venues, a uniform distribution means all the topics are, more or less, equally likely, resulting in high entropy. Topic-specific venues tend to publish more on one specific topic and its sub-topics, resulting in unequal topic distribution, and hence low entropy.

We hypothesize that citation count of venues has a statistically significant relationship with topic specificity. If our hypothesis turned out to be true, we would propose topic specificity as an additional or alternative measure to be used for ranking entities. So, in case
we get reasonable amount of correlation between citation count and topic specificity of publication venues, then topic specificity will be a text-based factor usable in addition to citation counts to rank academic entities.

The contributions in this work are as follows:

- discovered a weak negative correlation between topic sensitivity and citation counts of publication venues; and
- usage of topic specificity as a feature to rank academic entities.

2. Literature review
Publications data-oriented research is mostly focused on using the citations count to rank the academic entities. In case there are no citations for some academic entity due to paper is published recently or the citation graph data for the publication are missing or the publication is relatively very new and so far has not received many citations, there is no method which can tell us about the rank of the academic entities.

Use of citation worth has been explored in different ways in the past to identify the importance of academic entities. IF is considered to be the best indexing or most used technique for assessing the research journals. According to Eugene Garfield (2001) “I first mentioned the idea of an impact factor in Science in 1955.” IF is the average number of citations received by an article published in a given journal within a given database. Journals with higher IF are thought as more creative and more vital than that of those which have lower IF. IF totally depends on citations obtained by papers published in that journal over a specific time. It is important to note that IF is metric used to evaluate or rank the journals and should not be directly used to evaluate or rank individual authors or institutes. This is due to the reason some papers are cited a lot and some are not cited a lot over a specific time when IF is calculated. IF was based on citations of journals to provide their indexing or ranking, while few papers have dealt with finding rising entities, e.g. authors who are recently new researchers but are expected to be experts or highly ranked authors in near future (Daud et al., 2013; Zhang et al., 2007; Valenzuela et al., 2015). In 2006, Eugene Garfield (Archambault and Larivière, 2009) talked about the history of IF and its utilization as notwithstanding helping libraries choose which journals to buy. He emphasized on the need and importance of IF in academic research. According to Nazri and Halif (bin Abdul Halif, 2007) “Impact Factor is not a perfect tool to measure the quality of articles but there is nothing better and it has the advantage of already being in existence and is therefore a good technique for scientific evaluation. Experience has shown that in each especially the best journals are those in which it is most difficult to have an article accepted, and these are the journals that have a high impact factor. Most of these journals existed long before the impact factor was devised. The use of impact factor as a measure of quality is wide spread because it fits well with the opinion we have in each field of the best journals in our specialty.” IF totally depends on citations in a specific time.

Author ranking using citation count was extended by including the mutual influence (MuInf) rank was calculated by combined effect of papers and citations with respect to the MuInf of authors. Popularity rank and prestige rank are considered two different things for authors and were calculated by using publication count and citation count, respectively (Ding, 2011). Scientific Collection Evaluator by utilizing Advance Scoring was worked by Antonis Sidiropoulos and Yannis Manolopoulos (2005) to rank venues (conferences and journals) on citation bases to diminish the deficiencies created by CiteSeer and SCI. But the conferences organized in recent years that have zero citations did not show reliable results, so they ignored those.

Composite StarRank (Daud et al., 2013) method finds authors in academic social network, who will be prominent contributors in future. Composite StarRank finds such rising stars
according to author contribution order in paper and dynamic publication venue-based scores. StarRank score was calculated using MuInf of citations of authors and topic specificity in terms of entropy. The title words of papers published in a venue are used for calculating the entropy of venues. The lower entropy value is considered better as it shows that the venue has lesser disorder due to accepting papers specifically matching its areas of interest. Same technique was also used for finding experts (Zhang et al., 2007) on the bases of venues entropy. Use of entropy proved that dynamic publication venue scores are more useful in finding rising stars, where citations graph information is incomplete or missing. Marco Valenzuela (Valenzuela et al., 2015) tackled the important task of identifying important citations in scholarly literature. He identified citations importance by finding their occurrence in different sections of a paper by giving them different weights. Citations used in the method section were considered important citations which depicted that the cited work is used or extended in the citing paper.

Prediction of rising stars in co-author networks is performed by using content and graph features. In total, 11 features are used by applying discriminative and generative models. Venue specificity feature value is calculated by using entropy. The intuition was that mostly higher ranked venues have lesser entropy due to their topic specificity as compared to not higher ranked venues which may accept papers on many different topics. The features are categorized into three types: author-based, co-author-based and venue-based features, and venue-based features are found to be the most effective (Daud et al., 2015). Finding rising stars problem is tackled by proposing collaboration caliber-based method. Entropy is used for calculating the collaboration caliber of authors. They discussed the impact of investigating time interval for predicting the rising stars and found that top ranked rising stars have higher average citation counts as compared to existing methods (Zhang et al., 2016). Venue-influence language models are proposed based on entropy to consider citation links weights without using citations link information. The intuition was based on the idea that an author publishing in topic-specific publication venues will be expert on a topic as compared to an author publishing mostly in multi-topic venues (Al-Barakati and Daud, 2018).

Recent publications play a vital role in the life of academic researchers. Ignoring these publications or their impact is not a suitable solution. Also, in the past, the textual score for ranking academic entities is also not used. This aspect remains unexplored, which we propose to investigate by exploring the topic specificity of publication venues. In this paper, we provide the detailed study of correlational study of topic specificity and citation counts of publication venues.

3. Methodology
In this section, data preprocessing, hypothesis under study and entropy and normalized citation count (NCC) calculations are explained.

3.1 Data set
For implementation, we used the DBLP-V7 data set of 10 years from 2004 to 2013. Numbers, symbols and stop words are removed from titles and all titles are converted to the lower case by following standard practices (Al-Barakati and Daud, 2018) to calculate publication venues entropy. Some publication venues have more than one proceeding’s volume in a same year or variation in name, like DASFAA (1) and DASFAA (2) or Accounts of Chemical Research, Vol. 1 and Accounts of Chemical Research (1). DASFAA (1) and DASFAA (2) become DASFAA and accounts of chemical research, Vol. 1 and accounts of chemical research (1) become accounts of chemical research after preprocessing. We used data of ten different years to also see correlation coefficient trend over different years for all type of publication venues. There are total 758 unique workshops, 2,131 unique conferences and 859 unique journals. The total number of papers used in this study for all publication venues is 528,200.
3.2 Testing of hypothesis

We set forth the null hypothesis:

\[ H_0. \text{ There is no statistically significant relationship between topic specificity and citation count.} \]

To test the hypothesis, we computed the Pearson product moment coefficient between entropy and NCC. As it is mentioned earlier, entropy was our estimate of topic specificity.

3.3 Calculation of normalized citation count and entropy

For the above-described span of years, we calculated NCC and entropy for all the venues classified into journals, conferences and workshops. The title words of papers published in a venue are used for calculating the entropy of venues. Here, it is necessary to mention that the use of entropy enables us to calculate venues rank for the existing and new coming venues even without having the citation count information. Our main intuition is to find whether the entropy may act as a counter part or an additional factor with citations.

Formula for NCC and entropy of venues is given below:

\[
\text{NCC}(v_t) = \frac{1}{P_t} \times \sum_{i=1}^{n} \text{CC}(p_{ti}) ,
\]

(1)

where \( P_t \) is the total papers published in venue \( v \) in year \( t \) and \( \text{CC}(p_{ti}) \) is the citation count of paper \( p_t \) in year \( t \):

\[
\text{Entropy} (v_t) = - \sum_{i=1}^{n} w_i \log_2 w_i ,
\]

(2)

where \( \text{Entropy} (v_t) \) is the entropy of venue \( v \) w.r.t. to papers published in year \( t \) and \( w_i \) is the probability of word, in a venue \( v \).

After calculating both above measures for all venues, we found correlation between entropy and citations count. Then, we computed average correlation using mean, median and trimmed mean to measure the center estimation of the data set.

4. Results and discussion

4.1 Correlation between citation count and entropy

This section discusses the focus of this work which was to find a substitute of citations count or an additional feature to rank academic entities. In Table I, year-wise correlational coefficient between normalized citations count and entropy is shown. In 2004, only 489 conferences have non-zero publication counts. Coefficient of correlation between entropy and citations count is \(-0.28\). There is a weak negative correlation between them.

In 2005, it is \(-0.24\) for 569 conferences which have non-zero publication counts. Till 2012, we can see the similar kind of negative correlation. But for year 2013, one can see it is \(-0.43\), which is quite different from all other years. One can also see that the number of conferences data is only 86 conferences. We analyzed it and found that DBLP updates a few times a year and it is well known that most of the high-level conferences are held in the first half of the year. So, their proceedings were updated in DBLP system only when we crawled the data set. And as the focus of this paper also says that topic specificity is more in the high-level venues, the correlation is higher as compared to other years.

One can also see that, in most cases, if the number of conferences held in a year is high, the correlation is less; we have analyzed the data set and found that many new or multi-topic
conferences held in those years, resulting in less topic specificity of publication venues and thus low correlation.

Coefficient of correlation between entropy and citations count of workshops in 2004 and 2005 is −0.27 and −0.23, respectively. One can see that overall correlation is high in workshops as compared to conferences and journals. It is a well-known fact the workshops mostly focus on specific topics, so that there is more topic specificity which has resulted in higher correlation for workshops. One can see a similar trend like conferences here, i.e. from 2004 to 2011, there are more than 100 workshops and the entropy is mostly below −0.41. One can see that for years 2012 and 2013, there are only 90 and 16 workshops and correlation is −0.50 and −0.53, respectively, and correlation is highest for these two years as compared to other years. We analyzed it and found that DBLP updates a few times a year and it is well know that most of the high-level conferences are held in the first half of the year, so as their workshops are held together, which are also highly topic specific, high correlation is found for years 2012 and 2013. 2012 and 2013 data set is a bit less as compared to other years, as we have crawled the data in first quarter of 2013 from DBLP. DBLP was not updated yet for year 2012 last quarter publication venues and for last three quarters of 2013.

Coefficient of correlation between entropy and citations count of journals in 2004 and 2005 is −0.13 and −0.15, respectively. One can see that overall correlation is less for journals as compared to conferences and workshops. Journals have usually a bit wider scope with more sub-topics on which papers can be submitted, so topic specificity is a bit less.

Relationship found using Pearson correlation came out to be negative in all the years for all venues that show that the entropy of venues can be used in place of citations count or as additional ranking feature for academic entities discussed in this paper.

Figures 1–3 show the average correlation of conferences, workshops and journals, respectively, with NCC. Trend line in all the scatter charts depicts that a negative relationship exists between entropy and NCC. It shows that when average citations increase entropy decreases, which testifies our hypothesis in this paper that the publication venues with high average citations have less entropy.

One can see from Figures 1–3 that the spread of correlation for workshops is more as compared to conferences and journals, which is due to workshops being more topic specific. This situation also matches with the values in Tables I and II, in which the workshops have higher correlation as compared to conferences and journals due to their topic specificity.

<table>
<thead>
<tr>
<th>Year</th>
<th>No. of conferences</th>
<th>Correlation coefficient</th>
<th>No. of workshops</th>
<th>Correlation coefficient</th>
<th>No. of journals</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>489</td>
<td>−0.28</td>
<td>111</td>
<td>−0.27</td>
<td>394</td>
<td>−0.13</td>
</tr>
<tr>
<td>2005</td>
<td>569</td>
<td>−0.24</td>
<td>129</td>
<td>−0.23</td>
<td>453</td>
<td>−0.15</td>
</tr>
<tr>
<td>2006</td>
<td>622</td>
<td>−0.32</td>
<td>125</td>
<td>−0.31</td>
<td>473</td>
<td>−0.09</td>
</tr>
<tr>
<td>2007</td>
<td>711</td>
<td>−0.14</td>
<td>143</td>
<td>−0.28</td>
<td>512</td>
<td>−0.15</td>
</tr>
<tr>
<td>2008</td>
<td>746</td>
<td>−0.19</td>
<td>169</td>
<td>−0.20</td>
<td>530</td>
<td>−0.12</td>
</tr>
<tr>
<td>2009</td>
<td>841</td>
<td>−0.15</td>
<td>171</td>
<td>−0.27</td>
<td>569</td>
<td>−0.18</td>
</tr>
<tr>
<td>2010</td>
<td>835</td>
<td>−0.25</td>
<td>176</td>
<td>−0.41</td>
<td>606</td>
<td>−0.17</td>
</tr>
<tr>
<td>2011</td>
<td>824</td>
<td>−0.33</td>
<td>164</td>
<td>−0.35</td>
<td>637</td>
<td>−0.17</td>
</tr>
<tr>
<td>2012</td>
<td>552</td>
<td>−0.26</td>
<td>90</td>
<td>−0.50</td>
<td>606</td>
<td>−0.18</td>
</tr>
<tr>
<td>2013</td>
<td>86</td>
<td>−0.43</td>
<td>16</td>
<td>−0.53</td>
<td>358</td>
<td>−0.20</td>
</tr>
</tbody>
</table>

Table I. Statistics of correlational coefficient (2004–2013)

4.2 Overlapping between the ranking results
There is no standard yearly author ranking available for publication data sets, so that we may calculate standard ranking evaluation metrics, such as precision, recall, f-measure, NDCG, etc. For manually performing labeling task on whole data set of authors, authors
need a lot of funding and expert human resources, which are not available. That is why we have used OSIM ranking evaluation metric to show quantitative analysis. OSIM (Amjad et al., 2018) is used to find overlapping between two ranking results found using citation count and entropy for year-wise top 10 ranked authors for both methods. The more overlapping between ranking results (larger OSIM value) means the more useful is our proposed topic specificity method. The purpose to find OSIM value is to find usability of our proposed topic specificity feature for ranking academic entities. In workshops 2004, OSIM value is 0.36, in journals 2004, OSIM value is 0.24 and in conferences 2004, OSIM value is 0.33, which gives a clear idea of overlapping between ranking results and shows that topic specificity is useful. The larger the OSIM value, the more useful is the method.
4.3 Topic specificity of conferences and journals

Some conferences and journals publish articles only on specific topics and as they are specialized in a topic or its sup-topics, so they are very strict on accepting papers only related to them. These types of journals and conferences are usually high-level venues though there are few exceptions like *Nature*, *Science*, *PLOS ONE*, etc. High-level venues usually have low entropy as they only accept papers on specific topics and those papers are also reviewed by that specific area experts in that domain. Not-high-level venues have high entropy as papers other than specific topics are also accepted in them or they are multi-disciplinary. The reviewers for these journals are also not mostly that specific area experts, which results in accepting something not very high-quality papers. It is a well-known observation that if a paper is published in high-level venue, it will automatically have more chances to get large number of citations and will be an important paper for research community (Sidiropoulos and Manolopoulos, 2005). Table III provides some high-level and not-high-level conferences. One can see that most of the well-known conferences in specific areas have lower entropy as compared to less known conference provided in another column. And overall average of entropy of high-level conferences is also less as compared to the not-high-level conferences. Similarly, Table IV provides some high-level and not-high-level journals. One can see that most the high-level journals are area specific and have lower entropy as compared to not-high-level journals which are not topic specific but most are focused on the application of computing in different fields. And overall average of entropy of high-level journals is also less as compared to the not-high-level journals.
5. Conclusion
We have found correlation of entropy of venues (topic specificity) with citation count. Citation count is an existing important factor to rank academic entities. We conclude that topic specificity has a weak negative correlation with citations count. Therefore, one can use topic specificity score of publication venues as a ranking factor for academic entities. The application of topic specificity based ranking and reasonable overlapping with citations count based ranking also shows that it is a valuable addition to academic entities rankings in terms of score of text, which is not exploited before. As a future work, it will be interesting to apply the topic specificity of text based on entropy idea on other textual data sets, such as blogs, news, product reviews, etc.

References


**Corresponding author**
Ali Daud can be contacted at: ali_msdb@hotmail.com

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgrouppublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com
Scientific collaboration networks in Pakistan and their impact on institutional research performance

A case study based on Scopus publications

Fahad Sabah
Information Technology University, Lahore, Pakistan and
Superior University, Lahore, Pakistan

Saeed-Ul Hassan
Information Technology University, Lahore, Pakistan

Amina Muazzam
Lahore College for Women University, Lahore, Pakistan

Sehrish Iqbal
Information Technology University, Lahore, Pakistan

Saira Hanif Soroya
University of the Punjab, Lahore, Pakistan, and

Raheem Sarwar
College of Science and Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong

Abstract

Purpose – The purpose of this paper is to analyze the scientific collaboration of institutions and its impact on institutional research performance in terms of productivity and quality. The researchers examined the local and international collaborations that have a great impact on institutional performance.

Design/methodology/approach – Collaboration dependence measure was used to investigate the impact of an institution on external information. Based on this information, the authors used “index of gain in impact through collaboration” to find the impact of collaborated publications in institutional research performance. Bibliographic data between 1996 and 2010 retrieved from Scopus were used to conduct current study. The authors carried out the case study of top institutes of Pakistan in terms of publication count to elaborate the difference between high performing institutions and those who gain disproportionally in terms of perceived quality of their output because of local or international collaboration.

Findings – The results showed that the collaboration of developing countries institutes on international level had a great impact on institutional performance and they gain more benefit than local collaboration. Altogether, the scientific collaboration has a positive impact on institutional performance as measured by the cumulative source normalized impact per paper of their publications. The findings could also help researchers to find out appropriate collaboration partners.

Originality/value – This study has revealed some salient characteristics of collaboration in academic research. It becomes apparent that collaboration intensity is not uniform, but in general, the average quality of scientific production is the variable that most often correlates positively with the collaboration intensity of universities.

Keywords Bibliometrics, Collaboration networks, Institutional research performance, International collaborations, Research performance, Scientific impact

Paper type Research paper

1. Introduction

Scientific collaboration has gained a lot of importance in science policy. The tremendous growth of collaboration between research institutes since last 20 years has been in the interest of nations and institutes, as science policy initiatives. There are several examples of
exactly how under-developing countries are remodeled into developed countries within one or two decades. Once they have given the highest priority to the education, innovation and entrepreneurship, for example, significant growth can be observed in Hong Kong, Japan, Republic of Korea, China and Malaysia (Bakri et al., 2017; Wan et al., 2017), whereas Vietnam is in growth stage of building research capability (Nguyen et al., 2017). International publications of China have been increased rapidly which shows that collaboration among countries, institutions and authors has also increased (Chen et al., 2016).

In recent decades, the government in the global context has increased the policies toward the encouragement of international collaboration. These policies have supported the belief that scientific collaboration ends up in productivity gains for the researchers and ultimately for the country. Some examples of these policies are: Pakistan program for collaborative research, Pak-France PERIDOT research program and social integration outreach program. In each policy, researchers are required to collaborate as a condition to get research funding. Some other examples of policies for the encouragement to collaborate are inter-departmental policies (such as evaluations or rankings and employment or tenure choices that need a minimum quantity of publications) that are not totally discounted articles by the number of authors. Because of these policies, scientific collaboration between authors has been increased over recent decades. Indeed, the data showed that the share of collaborated research increased 66 percent within institutions from 2001 to 2002. Many authors have provided explanations for this increase, together with much higher gains from the specialization and division of labor, falling communication prices and a bigger pressure to publish among others as well as due to the impact of collaborations in terms of productivity (Banal-Estañol et al., 2013). There are many reasons for increase in collaborated publications; a very important one is the growing complexity of science and the increasing specialization of scientists, which suggests that a good solution to a scientific or technological drawback might need inputs with variety of potentials that may solely be derived from complementary information and skills of various scientists (Abrahams et al., 2018); moreover, the internationally co-authored publications receives more citations (Nguyen et al., 2017). The trend has increased for international collaboration in science and a growing awareness of scientific effort in distant regions through publications in common international journals. The arrival of telecommunications and internet played a sweeping role in the formation of contacts between geographically separated scientists and resulting in cooperative work between scientists and students who have not met except on the internet. Most of the knowledge transferred through the exchange of ideas and information among people (Sarwar and Hassan, 2015). Scholars have emphasized the role of collaboration for the advancement of scientific competence and economic process in developing countries (Park et al., 2014).

Many researchers have explored relative differential stress in collaboration with different countries in the numerous disciplines by using different techniques. Over the last decade, there has been a major increase in the range of international co-authorship (Hassan et al., 2016). We found that the perceived quality (as seen from the cumulated impact) of internationally co-authored papers was considerably more than for entirely native papers. Co-authorship may absolutely have an effect on institution’s productivity through many channels: cooperation could be more productive than operating alone as larger groups with concepts and skills of people. Another advantage is that the reduction of time to dedicate a project, as groups, makes the scientists think in the same direction with different authors at the same time. Observing the dramatic increase of co-authored articles between individual scientists and among institutions, one is inclined to assume that collaboration has become a necessity for contemporary science. As well as the collaborations result in evolution of modern scientific fields (Coccia and Wang, 2016).

Collaboration gain involves a number of statistical indicators to reveal the nature and change in collaborative networks. Traditional bibliometric indicators based on population
average cannot apprehend scaling relationships, e.g. citations per paper cannot encapsulate
the scaling correlation between impacts of collaborative and non-collaborative articles. This
case study tends to get the chance of putting on scientific collaboration as a replacement
methodology of gauging institutional research performance or quality of scientific
publications. The objective of the current study is to discuss the link between collaboration/
co-authorship and the nature of bibliometric information as well as its impact on an
institution’s performance.

This study aimed to discover the collaboration preferences of institutions and then we
examined the scholarly impact of collaborations on institutional research performance.
Furthermore, we examined whether or not the activeness of scientific collaboration between
institutes may be use to measure the research performance other than the old indicators like
the variety of produced articles or citations resulted from the institutes. We focus to examine
whether or not the scientific collaboration between institutes can be used to measure the
research performance. Hence, the method is compared with old indicators like the variety of
produced articles or citations resulted from the institutes. Followings are the main
objectives of this study:

(1) to investigate the collaboration dependency of an institution;
(2) to study the co-authorship preference index across the institutions; and
(3) to investigate the gain in impact through collaboration with international institutes.

This investigation shows that the collaboration of developing countries’ institutes at
international level had a great impact on institutional performance and they gain more
benefit than local collaboration. Altogether, the scientific collaboration has a positive impact
on institutional performance as measured by the cumulative source normalized impact per
paper of their publications. The main objective of this study is to evaluate collaboration
performance at institutional level. Hence, institutional research performance on the basis of
their local and international collaborations is measured as well as their preferences to
collaborate have been analyzed. The aim is to bring institutions of higher performance
closer to others so to enhance collaboration for developing human resource and
strengthening research in Pakistan’s institutions. Findings of this study would help scholars
and decision makers in understanding the current status and likely future trends of the
research, and help them to select the most appropriate collaboration partners.

2. Literature review
International collaboration enhanced scientific impact of almost all countries (Fu et al.,
2018). Many articles have revealed the significance of international collaboration as a
research outcome by using its citation impact. Factors, for example, the quality of
research, impact factor of the journal, authors’ count and international cooperation, are
stronger predictors for citations than other characteristics (Tahamtn et al., 2016). Papers
produced by foreign collaborations are more highly cited than those papers produced by
local collaborations (Hassan et al., 2016; Puuska et al., 2014). Studies have shown that
international collaborations have positive impact (Wagner et al., 2017). Numbers of studies
have explained reasons that the citation impact of international research collaboration is
higher than that of a local collaboration; they claim that international collaboration is
adequate with an advantage in terms of the diffusion of the research results compared to
domestic research collaborations. Researchers have been analyzing scientific
 collaborations as well as it is an established subject in scientometrics, but estimations
of domestic scientific collaborations remained speculative with studies based on a limited
number of fields or using data too insufficient to be illustrative of collaborations at
domestic level (Aref et al., 2018).
In the past, researchers have discovered collaboration patterns by the use of bibliographical data and social network examination. Analysis of the research collaboration network was at the level of regions, countries, organizations, institutions or individual authors (Park et al., 2014). The basic indicators like absolute publication and citations counts, their shares and averages are widely accepted as useful tools in measuring research performance, and their uncritical use can result in incorrect interpretations development. Academic relationships have a substantial impact on the upcoming number of citations (Tahmooresnejad and Beaudry, 2018). The current analysis approach generally supports quantitative figure, for example, publications count on international journals that may imperfectly reflect the research performance of the institution. However, such systems dominate the present ranking systems all over the world.

A scientific system can increase the impact of its knowledge by encouraging collaboration over non-collaboration. Collaboration is a positive strategy for policymakers to foster greater impact of scientific systems (Ronda-Pupo and Katz, 2017). Scientific collaborations are among the main supports of advancement in small national science systems (Aref et al., 2018). As experience with the collaborative development of science and technology, the probable overheads of a sole collaborative relationship might reduce. Thus, the lower costs for collaboration may possibly encourage the formation of partnerships with other diverse research and development bodies. Research and development entities in developing countries that ultimately encounter new techno-scientific challenges when evolving to the phase of autonomous innovation would increase and therefore enjoy the diversity of their partnerships at reduced costs for collaboration.

The less absorptive capability of institutions in developing countries might hamper the potential opportunities and becomes the hindrance for learning from other information sources. Initially, collaborations might facilitate the emerging countries or institutions to catch up with the beginning stages of the scientific and technological development easily but, based on the model, it perhaps induces the emerging nations or institutions as subordinate to the knowledge and technology providers can possibly hamper the growth of a country’s or an institution’s individual scientific inventiveness. Universities recommend their faculty members to indulge in research activities as their primary duty, as research performance plays an important role in recruiting and evaluating the process in the institutions. Top-ranked institutions in most of the developed countries are trying to obtain world-class university taxonomy since such rank and title allocate more respect and suggest robust competitive advantage (Ductor, 2015).

3. Data and methodology
We selected the publications data of Pakistan to investigate Pakistani collaborations. Recently, both international and local collaborations have been raised. The total publication count of Pakistan from 1996 to 2010 is around 40,000 out of which 27 percent are international collaborations which are 11,479 publications and local collaborations of Pakistan are about 35 percent which is 15,319 and publications without collaborations are around 38 percent which is 16,535.

We collected data using the Scopus to apply the typical measures of research performance and to examine the patterns of national and international cooperative research studies conducted by institutes of Pakistan. We gathered published data (1996–2010) for top 50 (with respect to the publications) institutes of Pakistan from the Scopus. The data of each document include author names, title, abstract, date, document type, addresses, cited references, etc. Since this study is focused on scientific research collaboration dynamics, we specifically analyzed the top 50 institutes of Pakistan in terms of publication output during 1996–2010.

The main reason of selecting the data set between 1996 and 2010 is that the international activities of the Pakistani universities expanded in scope, volume and complexity in this era. Since Higher Education Commission Pakistan was constituted in the subjected era and HEC
encouraged and supported the research culture in Pakistan in the form of collaborative research programs, this was the era which was of special attention regarding collaboration among institutions and their research productivity. If we just take an example of contributions of Pakistani authors in foreign library and information science journals, it shows that during the early years of the creation of Pakistan, the contribution of Pakistani authors to foreign LIS journals was in a single digit averaging less than one article per year. Only 13 articles were published between 1947 and 1970, 24 articles between 1971 and 1980 and 77 were published between 1981 and 1995, whereas more than 400 articles were published between 1996 and 2011 which is around 78 percent of all publications (Khurshid, 2013).

Recall that, the main objective of this investigation is to show the importance of dependency on external knowledge. In order to show that, we used the collaboration dependence measure, whereas, as for the impact of collaboration, it is measured on the basis of SNIP in GIFCOL which is simply the difference in the impact of publications with and without collaborations. Another measure used in the study is collaboration preference as institutional collaboration count does not state the preference of any institution to collaborate with another institute but it depends upon the overall collaborations of that institute as well as collaborations done by other institutes.

3.1 Bibliometric indicators used in this study
After selecting above-stated institutions, we extracted collaborations count in between these top institutes to find out the national co-authorship count of each university and drew their collaboration network. Note that any paper with at least one author from Pakistan is taken as the output from Pakistan for that year from any institution. If a paper contains no address originating outside that institution, then it is regarded as indigenous production. For papers involving collaboration between different institutions in Pakistan, full credit has given to each author regardless of the order of author names or relative positions of addresses. Following are the measures we deployed in our study.

3.1.1 Collaboration dependence. One of the indicators that we have used in our study is “Collaboration dependence” which is the dependence upon external knowledge of any institution for creating novel knowledge. Following is the formula for calculating Collaboration dependence:

\[
\text{Collaboration dependence (CD)} = \frac{n}{n-1} \left( 1 - \frac{\sum_{p=1}^{q} \frac{f_p}{N} }{N} \right),
\]

where \( n \) is the number of non-overlapping institutes; \( f_p \) the number of papers in which \( p \) institutes participated; \( N \) the total number of publications; and \( q \) is the maximal number of institutes in any paper.

3.1.2 Collaboration preference. Institutional collaboration count does not state the preference of any institution to collaborate with another institute, but it depends upon the overall collaborations of that institute as well as collaborations done by other institutes. Some researchers used an indicator for alleviating this issue using an affinity index, which later on renamed as collaboration preference.

Co-authorship or collaboration preference of an institution \( i \) toward institution \( j \) is measured by the co-authorship preference index, \( Q(i, j) \):

\[
Q(i, j) = \frac{(X(i, j)/X(i, *))}{(X(\ast, j)/(X(\ast, \ast)−X(i, \ast)))},
\]

where \( X(i, j) \) is the element in the \( i \)th row and \( j \)th column of the co-authorship matrix; \( X(i, \ast) \) and \( X(\ast, j) \) are the row sums of the \( i \)th and \( j \)th column, respectively; and \( X(\ast, \ast) \) is the overall sum of the elements of the matrix.
In other words, the co-authorship preference index is the share of the joint publications of institution \( i \) and institution \( j \) in institution \( i \)'s all co-authorships divided by the share of institution \( j \)'s all co-authorships in country's total co-authorships excluding those of institution \( i \).

### 3.1.3 Index of collaborative gain in impact

Gain in impact due to collaborations with international (GIFCOL) ensures comparability between institutes with varying performance levels and extent of collaboration since the performance characteristics of collaborative papers are being compared with the indigenous output of the same institute. The index of collaborative gain in impact factor is defined as the "gain in impact through collaboration" of an institution, as the differential increase in percentage terms, as follows:

\[
\text{GIFCOL} = \left( \frac{\text{Cumulated } IF_{\text{COL}}}{\text{Cumulated } IF_{\text{ALL}}} \times 100 \right) - \left( \frac{\text{Number of } PAP_{\text{COL}}}{\text{Number of } PAP_{\text{ALL}}} \times 100 \right) .
\] (3)

GIFCOL takes into account the proportion of papers that are collaborative and measures the difference between this and the corresponding proportional gain in impact. Obviously, if there is no difference between the average \( IF \) of internationally co-authored papers and the rest of the papers, this measure will be 0. Thus, the index is designed to vanish if the impact factor of co-authored papers of an institute is not more than the average impact of all other papers of the same institute. A positive value indicates a disproportionate gain in impact through collaboration. The index is a measure of the differential perceived quality (impact) of collaborative and indigenous production. However, the citation-based impact originated by these collaborations may be varied, since citation-based indices not always correlated well with each other (Haddawy and Hassan, 2014; Haddawy et al., 2016).

### 4. Results and discussion

Figure 1 shows the collaboration network among some of the selected top 50 institutions; nodes represent the institutions and edges represent the collaborations between these institutions. In this figure, the weight of edges represents the collaboration count between the two institutions and strong collaboration between University of Sargodha and Bahauddin Zakariya University, University of Punjab and PCSIR Laboratories Complex and Government College University and the University of Sargodha.

Further using Equation (1), on the retrieved data, we got the results as shown in Figure 2, which depicts the increase in dependence of external knowledge, thus, shows that the institutions are getting more dependent on external knowledge for creating novel knowledge. By using collaboration preference formula, it can be found that how much an institution \( i \) prefers to collaborate with institution \( j \). It can also be found that although the collaboration matrix is symmetric, co-authorship preference is not a vice versa or symmetric relationship. Based on co-authorship preference and gain in impact due to collaboration, we may propose some collaboration mechanism for these institutions.

#### 4.1 Analysis of international collaborations of Pakistan

Using Equation (3), we got the gain in impact due to collaborations both at national and international level. The results depict that top institutions, which do more collaboration with bottom-level institutions, get higher and positive impact out of their collaborations; overall, there is a positive impact on collaboration. Figure 3 depicts that institutions of Pakistan mostly collaborate with Asia, Europe and North America. During recent years, the trend has been changed drastically; as Figure 3 shows, collaborations count with Asia and Europe has been raised rapidly, whereas these numbers got higher with North America as well but not increased as fast as in the case of Asia and Europe.
With respect to collaborations count, Europe stands at the first and Asia lies at the second position, whereas North America is standing at the third position. Apparently, with respect to the gain in impact, Europe is at the first position, North America at the second and Asia stands at the third position. Since the number of collaborations with these selected continents is not the same, comparing with respect to gain in impact would have been more meaningful if we consider the average gain per collaboration. The results portray that with respect to the average gain per collaboration, North America is at the first, Europe is at the second and Asia stands at the third position.

4.2 International collaborations of top-selected institutes of Pakistan (1996–2010)
We analyzed the following selected institutes of Pakistan: Karachi University, Quaid-i-Azam University, Punjab University and COMSAT for the detailed analysis of collaborations and their impact on institutional performance during 1996–2010. Karachi University has done 42 percent of its publications with international collaborations, 26 percent with local collaborations and the rest are without any collaboration. Quaid-i-Azam University has
done 38 percent of its publications with international collaborations, 34 percent with local collaborations and 28 percent are without any collaboration. Punjab University has done 37 percent of its publications with international collaborations, 23 percent with local collaborations and 40 percent are without any collaboration, while COMSAT has highest international collaborations rate that is 49 percent of its total publications, 31 percent with local collaborations and 20 percent of the publications are without any collaborations.

In Figure 4 we analyzed international collaborations of top-selected institutes of Pakistan. The top influenced collaborator of the University of Karachi is Turkey because out of top five institutes, two collaborator institutions are from Turkey, which depicts that collaborations of Karachi University with Turkey resulted in increased research performance. Virginia Commonwealth University, USA, is at the first place then there is the University of Sussex, UK, and Bielefeld University of Germany is at the last. The top impacted collaborator of Quaid-i-Azam University is the USA because three institutions are from the USA. Columbia University, USA, is at the first place then there is Texas A and M University, USA, and the University of Delaware, USA, then Imperial College of London, UK, and Jiangnan University, China.

Figure 3. Yearly collaboration and gain in impact of collaborations of Pakistan

Figure 4. Top influenced international collaborations of selected Pakistani universities (1996–2010)
For Punjab University out of top five, two institutions are from the USA, which depicts that collaborations of Punjab University with the USA resulted in increased research performance. National Institute of Health, USA, is at the first place then there is Karadeniz Technical University, Turkey, and Johns Hopkins University School of Medicine, USA, Nihon University, Japan, and the University of Cologne, Germany. We then analyzed international collaborations of COMSAT and the results in Figure 4 depict that China is the top impacted collaborator of COMSAT. Nanjing University, China, is at the first place then there is University of Birmingham, UK, and Texas A and M University, USA, then Max Planck Institute, Germany, and Tianjin Polytechnic University, China.

4.3 Gain in impact due to collaborations with international institutes (GIFCOL)
As stated earlier that instead of focusing on counts we are interested in gain and institutional research performance so we also considered GIFCOL in our results. Figure 5 depicts that the University of Karachi while collaborating with the USA gains more impact out of its collaboration than all other countries. The gain in the impact of collaborations with China is negative which shows that the University of Karachi instead of getting better its institutional research performance, out of collaborations with China the overall research performance is decreased. Quaid-i-Azam University while collaborating with the USA gains more impact while collaborations with Germany and Turkey are negative.

Punjab University while collaborating internationally performs well although impact in case of collaborations of PU and USA is once again at the top and then there are China, Turkey, Germany, Japan and then the UK, respectively. COMSAT while collaborating with Germany gains more impact out of its collaboration than all other countries. The gain in impact of collaborations with the UK, the USA, China and Japan is at the second, third, fourth and fifth place, respectively, but the gain with collaborations of Turkey is negative.

5. Concluding remarks
The study shows that Pakistan has improved both relative research effort and an absolute number of overall publications. Optimistically, the changes that have occurred in the recent years in Pakistan are having noticeable effects on world’s research output profile.
Nevertheless, university managers and policymakers aim to stimulate interdisciplinary research. For them, the results of this study have the following three general implications: first, attracting staff from outside academia may increase the propensity of interdisciplinary collaborations. University managers should consider relieving the current financial barriers for industrial researchers to return to academia in order to enhance the diversity of university staff. Second, various instruments can be used to stimulate interdisciplinary research not all of which are of financial nature; the availability of funding for cross-disciplinary projects or programs seems a necessary but not sufficient condition for interdisciplinary research collaborations. Third, the ranking systems should consider collaboration as a quality measure.

References


Corresponding author
Raheem Sarwar can be contacted at: rsarwar2-c@my.cityu.edu.hk
Blockchain and OECD data repositories: opportunities and policymaking implications

Miguel-Angel Sicilia
Department of Computer Science, University of Alcalá, Alcalá de Henares, Spain, and
Anna Visvizi
School of Business and Economics, Deree College – The American College of Greece, Athens, Greece and College of Business, Effat University, Jeddah, Saudi Arabia

Abstract

Purpose – The purpose of this paper is to employ the case of Organization for Economic Cooperation and Development (OECD) data repositories to examine the potential of blockchain technology in the context of addressing basic contemporary societal concerns, such as transparency, accountability and trust in the policymaking process. Current approaches to sharing data employ standardized metadata, in which the provider of the service is assumed to be a trusted party. However, derived data, analytic processes or links from policies, are in many cases not shared in the same form, thus breaking the provenance trace and making the repetition of analysis conducted in the past difficult. Similarly, it becomes tricky to test whether certain conditions justifying policies implemented still apply. A higher level of reuse would require a decentralized approach to sharing both data and analytic scripts and software. This could be supported by a combination of blockchain and decentralized file system technology.

Design/methodology/approach – The findings presented in this paper have been derived from an analysis of a case study, i.e., analytics using data made available by the OECD. The set of data the OECD provides is vast and is used broadly. The argument is structured as follows. First, current issues and topics shaping the debate on blockchain are outlined. Then, a redefinition of the main artifacts on which some simple or convoluted analytic results are based is revised for some concrete purposes. The requirements on provenance, trust and repeatability are discussed with regards to the architecture proposed, and a proof of concept using smart contracts is used for reasoning on relevant scenarios.

Findings – A combination of decentralized file systems and an open blockchain such as Ethereum supporting smart contracts can ascertain that the set of artifacts used for the analytics is shared. This enables the sequence underlying the successive stages of research and/or policymaking to be preserved. This suggests that, in turn, and ex post, it becomes possible to test whether evidence supporting certain findings and/or policy decisions still hold. Moreover, unlike traditional databases, blockchain technology makes it possible that immutable records can be stored. This means that the artifacts can be used for further exploitation or repetition of results. In practical terms, the use of blockchain technology creates the opportunity to enhance the evidence-based approach to policy design and policy recommendations that the OECD fosters. That is, it might enable the stakeholders not only to use the data available in the OECD repositories but also to assess corrections to a given policy strategy or modify its scope.

Research limitations/implications – Blockchains and related technologies are still maturing, and several questions related to their use and potential remain underexplored. Several issues require particular consideration in future research, including anonymity, scalability and stability of the data repository. This research took as example OECD data repositories, precisely to make the point that more research and more dialogue between the research and policymaking community is needed to embrace the challenges and opportunities blockchain technology generates. Several questions that this research prompts have not been addressed. For instance, the question of how the sharing economy concept for the specifics of the case could be employed in the context of blockchain has not been dealt with.

Practical implications – The practical implications of the research presented here can be summarized in two ways. On the one hand, by suggesting how a combination of decentralized file systems and an open blockchain, such as Ethereum supporting smart contracts, can ascertain that artifacts are shared, this paper paves the way toward a discussion on how to make this approach and solution reality. The approach and architecture proposed in this paper would provide a way to increase the scope of the reuse of statistical data and results and thus would improve the effectiveness of decision making as well as the transparency of the evidence supporting policy.
Social implications – Decentralizing analytic artifacts will add to existing open data practices an additional layer of benefits for different actors, including but not limited to policymakers, journalists, analysts and/or researchers without the need to establish centrally managed institutions. Moreover, due to the degree of decentralization and absence of a single-entry point, the vulnerability of data repositories to cyberthreats might be reduced. Simultaneously, by ensuring that artifacts derived from data based in those distributed depositories are made immutable therein, full reproducibility of conclusions concerning the data is possible. In the field of data-driven policymaking processes, it might allow policymakers to devise more accurate ways of addressing pressing issues and challenges.

Originality/value – This paper offers the first blueprint of a form of sharing that complements open data practices with the decentralized approach of blockchain and decentralized file systems. The case of OECD data repositories is used to highlight that while data storing is important, the real added value of blockchain technology rests in the possible change on how we use the data and data sets in the repositories. It would eventually enable a more transparent and actionable approach to linking policy up with the supporting evidence. From a different angle, throughout the paper the case is made that rather than simply data, artifacts from conducted analyses should be made persistent in a blockchain. What is at stake is the full reproducibility of conclusions based on a given set of data, coupled with the possibility of *ex post* testing the validity of the assumptions and evidence underlying those conclusions.

Keywords OECD, Decentralization, Blockchain, Ethereum, IPFS, Open archives

Paper type Technical paper

1. Introduction

Due to advances in sophisticated information and telecommunications technology (ICT), the potential inherent in data repositories and the specter of potential misuse of these repositories have increased substantially in recent years. The ultimate test of the value of a data repository rests in whether the artifacts are made persistent in those repositories and to what extent the content of the repositories is of a certain quality and is truthfully described (Rousidis *et al.*, 2014). Clearly, apart from the opportunities, advances in ICT also produce risks and challenges. Access to data and new forms of data usage suggest that a new breed of cybersecurity threats might be at stake (Li *et al.*, 2017). From a different angle, as our societies evolve and the notion of the global village (McLuhan, 1962) increasingly is an undeniable reality, the argument of a more transparent governance system gains momentum. Along with the argument of more transparency come corresponding claims of more accessibility, accountability and legitimacy in the decision-making process (Lytras *et al.*, 2017). And indeed, advances in ICT make the otherwise nearly utopian claims of cosmopolitan democracy (Archibugi and Held, 1995) very feasible today. It is in this context that blockchains seem to offer a way of reconciling the need for data-driven decision making and claims for more transparency, accountability and verifiability of decisions made and actions taken. Clearly, blockchain technology revives the perennial questions of the use and production of knowledge and the nature of science (Visvizi, 2015; Hossain *et al.*, 2016).

Blockchain technology emerged as a supporting infrastructure for public, massively redundant immutable shared ledgers. While the first blockchain application, Bitcoin (Antonopoulos, 2014), had the purpose of providing an internet-based store of value, i.e., a digital currency, the underlying technology is applicable to any situation in which there is a need to have tamper-proof stores of histories of transactions. This makes blockchain technology a good candidate to serve as a supporting infrastructure for sharing data and analytic artifacts that need to be immutable for later scrutiny, as required by the above-discussed motivating needs. Current blockchain uses are limited by some scalability constraints that may prevent the storage and processing of large binary objects, as some statistical data sets may represent. However, the combination of blockchains with decentralized file systems such as InterPlanetary File System (IPFS) has been proposed as a combination that overcomes those limitations by separating concerns, so that cases of digital archival can be handled, as proposed elsewhere (Garcia-Barriocanal *et al.*, 2017).

The added value of blockchain technology is that, unlike traditional databases, it stores the full record of every piece of data and its modifications. Another feature of blockchain is
that security is enforced by a network of computers that do the processing redundantly and work toward the same end under a system of incentives. This redundancy brings security against tampering but at a high cost of processing. This is why for large files it is preferable nowadays to use an external system such as IPFS. Some blockchains are public, i.e., open to anybody and not owned by anybody, e.g., Ethereum[1]. There are other intended for (semi-)private uses, e.g., those based on Hyperledger[2] (Linux Foundation) software. In the context of blockchain, smart contracts allow not only the transfer of value but also general purpose processing, for cases that require the unique characteristics of the blockchain. This means that if one assumes that data once entered in a given data repository is persistent and immutable, then the artifact thus created becomes a part of an ever-larger body of data and artifacts. Thereafter, building on the distinct quality of the blockchain, diverse ways of processing data and artifacts is possible, not only discovery processes are boosted but also knowledge generation processes can be substantially enhanced.

This perceived promise inherent in blockchain notwithstanding, blockchains remain an emerging technology whose potential and limitations are still under-explored. Against this backdrop, the objective of this paper is to employ the case of Organization for Economic Cooperation and Development (OECD) data repositories to test, prove and suggest how blockchains could be employed more effectively, for instance, enabling citizens to get access to open, tamper-proof data and analytics artifacts, foster research or to enable epistemic communities, e.g., diplomats or members of highly specialized bodies operating under the aegis of international organizations, to communicate and devise policy recommendations more effectively.

Current actions developed and tracked by governments, agencies and other policymakers are the result not only of a set of values directed by political agendas but also of evidence and data, that is, systematically collected, stored and shared. Specifically, socio-demographic and economic data, when combined with administrative and geographical data, provide the basis for a wide range of analytic results, which in turn represents key evidence supporting policies and other decision-making processes. The concept of “evidence-based policy,” while controversial in its practical feasibility or the degree of mix of evidence and values (Marmot, 2004), represents a trend toward such approaches to policy and program development that consider the link to supporting data as a key element in justifying and later testing the effectiveness of concrete actions. For managers whose decisions result in tangible outcomes in the public sphere, the opportunity is apparent for continuous improvement in policy settings and program performance, based on a well-informed debate of options, but it is also multi-dimensional in the types of knowledge used (Head, 2008).

The evidence supporting programs and policies is nowadays typically cited in reports or in the regulatory or motivating texts of those actions. However, in some cases it is not clear if that evidence comes from the raw, unprocessed data cited or if it is the outcome of the process of some analytic result done on that data. In other cases, the evidence is a published study that in turn offers conclusions based on some analytic process. As the action is backed and justified by the data, it is critical that all the steps in the process, back to the original data source, are traceable so they can be reproduced as a guarantee of trust as well as a way of re-evaluating them with new data as to whether the conditions that motivated the action still hold in light of the new evidence. As a consequence, an ideally well-documented action should provide traces to both the data and the analytic procedures (which may, e.g., be described in statistical programming languages) applied to derive the conclusions from that data. It also requires that these artifacts are permanently available, immutable and tamper-proof. From a different angle, the possibility of retrospectively evaluating actions already taken is hampered, thus reducing their transparency and accountability. A significant effort to address this problem took the form of e-government initiatives in the early 2000s and the Open Government Data Initiatives (OECD, 2016c). This paper explores the question of how...
and to what extent blockchain technology may be of use to address the challenges and constrains related to data repositories and their efficient use. It also investigates the question of how effective application of blockchain can assist us in addressing salient societal concerns related to data accessibility, transparency in decision making, accountability and legitimacy. To this end, the problem of tracing the evidence used in policy and program actions is addressed by focusing on the provision of public infrastructure that increases transparency and evaluation. A combination of blockchain and decentralized file systems are proposed as the core infrastructure elements for a solution, following previous work on the topic (García-Barriocanal et al., 2017). Then, a proof of concept case is described based on the OECD framework, illustrating the benefits and pitfalls of that solution.

The rest of this paper is structured as follows: Section 2 provides background information on the context of the research done and the target technology used; Section 3 describes the main elements of the model and architecture for the solution presented; Section 4 reports on the specifics of a concrete proof of concept using OECD data; then, in Section 5, conclusions and an outlook are provided.

2. Background

2.1 Policies, analytics and data sharing

A key component of e-government and open government initiatives (Ubaldi, 2013) is the ability of multiple government and nongovernmental organizations to share and integrate information across their traditional boundaries. Several e-government Interoperability Frameworks (e-GIFs) have been proposed to date, sharing similar principles and comprising groups of documents that specify a set of common elements, such as vocabularies, concepts, principles, policies, guidelines, recommendations, standards and practices (Lisboa and Soares, 2014; Ubaldi, 2013; OECD, 2015). However, we have not found a detailed metadata schema that can model policies or programs directly and contain elements to link to supporting evidence in an interoperable way, albeit some ontologies have been proposed (Loukis, 2007) that describe policies. In any case, this could be achieved as an initial solution via general purpose metadata schemas such as Dublin Core, widely used in e-government initiatives (Alasem, 2009). For example, the relation element in Dublin Core could be used to make a general statement of the relation of a program to a specific data set, and additional qualifiers could be developed as a tailored vocabulary for making that statement more precise. On the other side of the traceability chain, the development of data description frameworks is more mature and well developed. Here, we take the case of the SDMX standard as the reference framework, as it is backed by main actors in the provision of statistical data, including Eurostat and the OECD, and has been used as a case study for open data (Capadisli et al., 2015).

2.2 Data sharing practices and problems of linking policy and analytics to data

For the sake of simplicity, we will abstract the authors from our problem setting into three categories: data providers, who expose and share the base statistics and data of interest, data analytic services, which exploit and combine that data to produce derived information, findings or summaries and action makers, who use the results of the former as supporting evidence for mandates, initiatives or programs. It should be noted that these are roles, so that the same agent may play several of them.

At the level of data providers, frameworks for publishing and curating data are commonplace. The abovementioned SDMX provides an adequate and well-founded model for the sharing of statistical data, and other archival models as the Open Archival Information System could complement it in a more traditional archival direction. Further, the sharing and
linking of data can also be done using technology based on the paradigm of internet publishing and supporting semantics also (Capadisli et al., 2015). Multiple software tools can be used for that purpose (Barbosa et al., 2017). Consequently, the framework proposed here comes to complement the data storage layer of these frameworks (which is essentially based on the traditional client-server model of the web) with the benefits of shared public ledgers and file systems based on replication. The level of analytics is more difficult to approach, as there are still not widely used conventions for it. In recent years, the ideas of reproducible research (Peng, 2011) or scientific workflows have contributed to advance in ways of sharing the processing of data, but this is still unclear. Finally, at the level of action makers, current e-GIFs do not address the specificities of referencing data as supportive evidence for policy, so that a more ad hoc approach should be used. However, recent research (Tang et al., 2017) points out that analytics is important in internal audits, and that perception is arguably expanding also to external scrutiny of data published in the public.

2.3 Decentralized file systems and blockchains

The term blockchain refers to the core data structure of a category of decentralized database architectures that rely on cryptographic techniques and distributed consensus (based originally on peer-to-peer network architectures) to provide tamper-proof distributed ledgers, leveraging previous research on algorithms but considering a trustless environment (Wattenhofer, 2016). The first widespread blockchain application was that of Bitcoin. It used that technology to implement the possibility of interchanging a token (a digital currency) among non-trusted parties without the need of a central authority and preventing double spending.

The application of blockchains have since Bitcoin inception be extended to the notion of “smart contracts” – first proposed by Nick Szabo (1997), in which non-trusted parties can interact with the blockchain for different kinds of transactions including some logic that can be implemented in Turing-complete programming languages. This supports the implementation of token interchange systems similar to Bitcoin, but also many other applications as voting, registries or futures’ markets or even accounting (Jun and Vasarhelyi, 2017; Lemieux, 2017). The main current exponent of such technology is Ethereum (Wood, 2014) that we consider here as the foundation for the analysis.

As blockchains today carry out highly redundant computation to achieve high levels of security, they currently have limitations or are expensive when storing large digital files. In consequence, there is a need to use associated technology that complements this aspect, as decentralized file systems. According to its proponents, the IPFS is a peer-to-peer distributed file system that seeks to connect all computing devices with the same system of files (Benet, 2014). IPFS and other similar frameworks bring a disruptive approach to the archival of digital resources that is based essentially on independence of location and decentralized storage by networks of untrusted peers (swarms). This technology features important implications from the technical perspective (as built-in de-duplication), but also from the view of the governance of open data and the current reliance on trusted data providers.

3. Decentralizing data systems and associated analytics

3.1 Sharing data artifacts in decentralized file systems

The fundamental data sharing commitment of the system considered here is based on the assumption that all documents and data, including metadata records, are stored in a decentralized file system. This entails two important implications. The first is that documents are hashed based on their byte content, and content hashes become identifiers for these documents. This entails the need to deposit canonical versions, i.e., versions of the same document in different formats should be avoided (as they will result in different identifiers).
or if deposited, additional metadata needs to be used to assert they are the same document and cross-reference them. The second is that data and resources in these file systems are constrained to be directed acyclic graphs (DAGs), so that there are no referencing cycles. That sharing provides immutable, permanent storage to all the actors in the solution, however, the authenticity of the deposits requires additionally the cryptographic guarantees of a blockchain as described below.

3.2 Smart contract-based architecture

Figure 1 shows the overall high-level architecture for a solution to provide links to analytic results that in turn refer to publicly available data where full traceability is guaranteed. The different actors, as said above, deposit the resources in the distributed file system, obtaining their hashes. For example, a data provider, or an ingestion module retrieving periodically data from a data provider, may regularly deposit statistical data and metadata, the latter referencing the former. Then, an analytic provider may reference that data or metadata and deposit analytic objects, e.g., scripts, diagrams or sets of resources for complete reproducibility. Finally, policy documents, for example, may reference either data or analytic objects, forming DAGs with their references.

The blockchain uses smart contracts to allow the different actors to register via transactions “statements” about the deposited objects. A statement is a digital claim, in which a cryptographic identity asserts something about a permanent digital object. For example, a policy maker may deposit a policy document and an associated metadata record, and that metadata record may reference other objects deposited, for example, a script that does some statistics, which in turn may reference a data set. The transaction gets digitally signed in the blockchain, so that users can verify it comes from the right provider and can also trust that the resources referenced have not tampered or modified. Eventually and in cases needed, that digital signature may be combined with the use of digital certificates so that known physical world entities as persons, organizations or institutions can be the regarded as the issuers of particular transactions. It should be noted that these are similar to “verifiable claims” as described in W3C’s drafts[3] that models them as triples: “entity, subject, claim.” The key point is that the claims are made immutable by the blockchain, and in our setting, they are also made public and openly accessible, providing the transparency needed for the evidence supporting policy. Further, arbitrary data about which claims can be made are stored in a decentralized file system, so that they do not rely on particular trusted parties but on an open network of actors that contribute to the maintenance of

![Figure 1. Overall architecture](image-url)
contents linked in a DAG and that cannot be tampered thanks to content hashing (this in IPFS is achieved by its internal data structures, Merkel DAGs).

This structure creates in consequence a referencing dependency from the blockchain to the items deposited as depicted in Figure 1. The figure shows on the left the role of a decentralized system for file storage. Given that such storage is based on content hashes, they serve as immutable, replicated repositories of files, supporting the need for permanent identifiers of files and consequently, links to permanent versions of the file. The ingestion module can be used for maintaining a catalogue of files deposited for a particular purpose or application. Data providers can use that module to deposit files, but full provenance requires a digitally signed transaction in a blockchain that references the file, in case needed, associated to a digital certificate if traceability to an individual or organization is wanted. Then, analytic providers and policymakers can use the catalogues to access data, and if required, check that the provenance is correct. Metadata at different levels can be included in this schema, chaining references in the decentralized system, as described by García-Barriocanal et al. (2017). This simple approach enables full traceability to permanent resources (raw data, clean data, analytic artifacts and so on), leaving reproducibility to software for specific purposes and the metadata conventions and analytic languages of choice for a particular area of concern. The choice of metadata schema and conventions is key for interoperability, but it is not considered further here as it belongs to the discussion on standards and community practices in the field of e-government. The architecture proposed is agnostic to particular forms of metadata, which may eventually emerge as de facto or formal standards.

The processing involved in the smart contracts in the blockchain should cover the basics for creating (and eventually revoking by their original issues, but never deleting) and maintaining claims or statements. Additional services may be conceived in the future, but the base system is covering essentially archival functions, and the rest of the processing is left to external end-user applications that exploit that archived history of claims.

The resulting architecture is completely open and allows for organizations as the OECD to deposit data with verifiable provenance via digital certificates, and other organizations to deposit artifacts of various kinds referencing those. The added benefits are immutability of data deposits and complete transparency. This results in a chain of artifacts and claims that can be traced back to the original raw data, following a relatively simple set of publishing conventions that can be implemented on current smart contract-based blockchains.

4. Proof of concept and discussion
4.1 Provisioning, storage and provenance statement of data and structural metadata
The initial step in using the model described above would be that of regularly updating the decentralized infrastructure with the corresponding updates in the existing one, i.e., in the regular repositories of the publishing organization. If we consider the case of SDMX data, the acquisition of data for the decentralized infrastructure can be done, for example, using the SDMX-ML protocol provided by the OECD (2016a). Essentially, data structure definitions can be retrieved using the GetDataStructure method and associated data using the GetData method. Provisioning the data can be done by extracting the frequency of provision that can be obtained from the CL_XX_FREQUENCY code list as an approximation, or via other common means such as RSS notifications. The ingestion module would then push the new data and the metadata into the decentralized file system, then data providers are able to register it though a DataRegistry smart contract as the one sketched below. That contract is limited to recording statements, i.e., claims made by a particular digital signer that some piece of data follows some description. The following snippet in the Solidity 0.4 contract language (one of the languages in the Ethereum public blockchain[4]) shows the basic transaction structure (Figure 2).
The contract essentially allows any publisher (represented by its address) to claim that a given metadata resource pair that is stored in a distributed file system (referenced by the rhash and mhash content addresses, relative, respectively, to the rservice and mservice services, which in our case would be “IPFS”) is stored in the blockchain. The mmeta and rmeta strings are for supplementary meta-metadata, i.e., meta-descriptions of the claim itself, for example, contact information for the creator, dates or other references. The claim is then stored and indexed by publisher first and secondary by document hash. Additional functions in the smart contract can be used to search or retrieve the list of claims from a publisher or document, requiring additional data structures for efficiency that we do not discuss here.

The use of some “oracle,” i.e., a service that checks external facts, is needed to guarantee that the resources exist in the decentralized file system prior to storing the claim (this avoids “dangling pointers” to non-existing pieces of information). However, the only way of knowing that the resources are the right ones is by trusting the signer of the transaction. This should not be considered a limitation, as there is a basic assumption of trust in the issuer of the claim.

It should be noted that the generic pattern above can be used to “retract” claims or update them, so that the full history is recorded, and clients can check that history to assess the different parts of a traceable data provenance chain. This is accomplished by using different verbs: add, retract, etc.

It should be noted that the simple approach just described covers the archival function, but the linking that enables the trace is left to conventions in metadata, as will be discussed in what follows.

4.2 Statements for analytic procedures
An organization that has conducted an analytic process on some data operates over a set of references to data and metadata that has been stored previously. In this case, the ProcessRegistry smart contract operates on a similar means as above, but the external checks are extended to include also the software components or libraries required for the processing. Ideally, there should be a full check of the results of the analytic process that is repeatable. The problem is that the results of analytic processes are varied and may go from a simple statistical test of concrete hypotheses to a complex analysis in which a visual interpretation is the key insight obtained.

Eventually, analytic processes may store new derived data sets using the same procedures described above, which in turn can be used in subsequent analytic processes. Following the steps devised, the full trace would always be available transparently. Here, we

```solidity
contract DataRegistry {
    // ...
    mapping(address => mapping(bytes => Claim[])) statements;

    function claim(bytes rhash, string rservice, string rmeta,
        bytes mhash, string mservice, string mmeta){
        var resource = Handler(rhash, rservice, rmeta);
        var meta = Handler(mhash, mservice, mmeta);
        oraclize(resource); oraclize(meta);
        var claim = Claim(resource, meta, Verb.Add, block.timestamp);
        statements[msg.sender][resource.hash].push(claim);
    }

    function retract(...){
        // ...
    }

    function check(address curator, bytes rhash)
        constant public returns (bool){
        // ...
    }
}
```
delineate the process that could be used for the case of statistical processing using the open-source framework of the R statistical language as a concrete case, a computing environment for which different domain-specific practices, tools and ideas for reproducible research have emerged in recent years (Stodden et al., 2014). We start from considering a reproducible, R package aiming at providing high-level, robust, machine- and OS-independent tools for making deeply reproducible and reusable content in R. It essentially allows the analyst to store selected artifacts as binary files together with their metadata and relations, identifying them via MD5 hashes. Redirecting R objects to IPFS rather than to a local file results in a full trace of the objects required to reproduce the analysis. This must be extended also to the R package itself and all the libraries used. The missing element is the piece of metadata that is produced as a result of the analysis and used to support the action or policy, which should be incorporated into the metadata of that action.

4.3 Linking to actions and supporting assessment

The final step in the link chain requires a description of policies in some form. As there are no widely accepted formats for encoding policies, here we use an ad hoc markup for the representation, using in our case a fragment expressed in W3C Resource Description Framework (RDF) language for the description, but ideally, interoperable schemas for policy descriptions would be used instead. There are emerging interoperability standards for linking data and literature that may be relevant for this function (Burton et al., 2017), even though they are nowadays still not widespread that may also be applicable.

We take as an example the policies regarding foreign aid that are common in many European governments. Typically, foreign aid is grouped in programs that comprise a number of concrete actions. The following is an example of a fragment of metadata for this kind of action, in which general information on a policy is combined with references to analytic results. Note that this action description should be deposited in the data registry as a claim ideally by the agency or organization responsible of the policy. Figure 3 shows an example fragment using the Turtle RDF syntax illustrating the syntax.

The references term[5] uses a pointer in the source qualifier, an IPFS path to a supporting document for the action. IPFS has an associated specification of links called IPLD[6] for hash-links that could be adapted for this use instead. This reference could be used as input to some function in the DataRegistry discussed above to retrieve associated claims, for example. It can also be used to retrieve the actual resources from the P2P network of IPFS. Note that references are not qualified, i.e., they are not typed and no additional information on the “kind” of reference is used. All these should be dealt with by schemas and ontologies describing policies, as the one described by Loukis (2007) that incorporates programs, alternatives, issues and criterions that may serve as a point of departure for semantically processing the chains of artifacts resting in the architecture presented.

The utility of the terms in the aims namespace are those recording the policy objectives of the action and are in this case expressed as a query, since the data may still not be available. This example has only ideas for two of the key elements in action description: references to some supporting evidence document, and then a statement of the objectives of the action. Of course, this is just a very simple pattern and many other references and descriptions could be added, but they would be represented similarly.

The type of conventions and approach presented so far albeit simple can be used for a number of applications and use cases. The most basic are those of tracing resources and their provenance, together with providing permanent and immutable storage for an analytic artifact. Those functions are traditionally provided by repositories, while semantics in archival should be applied as usual (Llanes-Padrón and Pastor-Sánchez, 2017), and decentralization brings the benefit of not depending on the sustainability and security of a particular provider organization. However, the important benefits come from the use case of
fully reproducing the analytic procedures end-to-end, including references to people and organizations that are accountable, using similar models as those in IR/CRIS systems (Rybinski et al., 2017). This may start from a particular policy action and then go back to the analytical artifacts that support or provide the motivation or rationale for the action, finally being able to retrieve the original data providing the supporting base evidence. If the descriptions are rich enough to support automation of the analytic process and the full set of artifacts has been deposited, this can be used not only for checking actions but also to test variations of analytical procedures on the same data. This brings a new level of reuse and alternative testing to assessing policies and actions.

5. Conclusion and future work
Data repositories are a source of wealth, the value and utility of which until recently have remained under-explored. Recent advances in sophisticated technology, particularly analytics and blockchain, suggest that our ability to use data repositories might increase efficiently. As the debate on blockchain applications to analytics has only begun, more focused research on the specificities of blockchain, including its strengths, weaknesses, promise and limitations in this area, is required. To address this, we employed the case of OECD data repositories to examine the potential and limitations of blockchain technology with regard to addressing basic societal concerns, such as transparency, accountability and efficiency in a policymaking process today. It was argued that blockchain not only revolutionizes the conceptualization of data repositories but also the very use of data repositories and their potential for the purposes of research and decision making. The point was made that it is particularly valid in the context of official socio-demographic and economic statistics, which are unique and used widely across fields and domains. The findings of the discussion highlighted that at the end of the day, the question of blockchain and its utilization is about the use of knowledge and the very quality of contemporary systems of governance.
The practical implication of the research presented in this paper is twofold. On the one hand, by suggesting how a combination of decentralized file systems and an open blockchain, such as Ethereum supporting smart contracts, can ascertain that artifacts are shared, this paper paves the way toward a discussion on how to make this approach and solution reality. On the other hand, by bringing together the debate on blockchain and the policymaking process, light was shed on the prospect of bypassing some of the most topical concerns contemporary societies express, including those related to transparency, data accessibility and accountability. Indeed, the approach and architecture proposed in this paper would provide a way to increase the scope of the reuse of statistical data and results, and thus improve the effectiveness of decision making as well as the transparency of the evidence supporting policy.

The paper offers the first blueprint to a form of sharing that complements open data practices with the decentralized approach of blockchain and decentralized file systems. The case of OECD data repositories was used to highlight that while data storing is important, the real added value of blockchain technology rests in the possible change on how we use and reuse data and data sets stored in our repositories. Sharing data and making artifacts persistent in data repositories, as well as developing the corresponding architecture to support it, could extend current data sharing practice well beyond its current scope, provided that schemas and semantic models are adopted by the different stakeholders. Therefore, it would enable a more transparent and actionable approach to linking policy up to the supporting evidence.

Blockchain and related technology are evolving rapidly, and more research is needed to explore their potential in full. As this study indicated, several issues require particular consideration in future research, including security, scalability, validity and stability of data repository, and, certainly, ethical questions. With this paper serving solely as an introduction to the debate, in our future research the question of how the sharing economy concept could be employed in the context of blockchain will be addressed.

Throughout the paper, frequently implicitly, a case was made that, rather than simply data, it is the artifacts of analysis already conducted that should be made persistent in blockchain. What is at stake is the full reproducibility of conclusions based on a given set of data, coupled with the possibility of ex post testing of the validity of the assumptions and evidence underlying those conclusions. Arguably, the latter point inserts itself into the broader movement toward “reproducible science.”

Notes
1. www.ethereum.org/
2. www.hyperledger.org/
3. www.w3.org/TR/verifiable-claims-data-model/
5. The actual content identifier, schemas and queries presented in Figure 3 are just examples for the purpose of illustration of the syntax, not references to actual data or resources that are meaningful to the case.

References


Further reading


Corresponding author

Miguel-Angel Sicilia can be contacted at: msicilia@uah.es

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com
Abstract
Purpose – The purpose of this paper is to examine the evolution of collaboration among researchers in Library Hi Tech based on the co-authorship network analysis.
Design/methodology/approach – The Library Hi Tech publications were retrieved from Web of Science database between 2006 and 2017. Social network analysis based on co-authorship was analyzed by using BibExcel software and a visual knowledge map was generated by Pajek. Three important social capital indicators: degree centrality, closeness centrality and betweenness centrality were calculated to indicate the co-authorship. Cohesive subgroup analysis which includes components and k-core was then applied to show the connectivity of co-authorship network of Library Hi Tech.
Findings – The results indicated that around 42 percent of the articles were written by single author, while an increasing trend of multi-authored articles suggesting the collaboration among researchers in librarian research field becomes popular. Furthermore, the social network analysis identified authorship network with three core authors – Markey, K., Fourie, I. and Li, X. Finally, six core subgroups each included six or seven tightly connected researchers were also identified.
Originality/value – This study contributed to the existing literature by revealing the co-authorship network in librarian research field. Key researchers in the major subgroup were identified. This is one of the limited studies that describe the collaboration network among authors from different perspectives showing a more comprehensive co-authorship network.
Keywords Co-authorship, Social network analysis, Analysis, Library Hi Tech, BibExcel, Pajek
Paper type Research paper

1. Introduction
Researchers from different fields are facing a rapidly changed environment because of the development of scientific techniques, tools and methods. Thus, the collaboration among researchers with different specialties has increased (Cronin et al., 2003, 2004) in order to improve the scientific research quality, efficiency and visibility (Beaver, 2001). Co-authorship is one of the most important ways to demonstrate that the social network exists among researchers in one specific research field and is a popular issue in bibliometrics research, which includes author co-citation, co-authorship and co-word analysis (Leydesdorff and Vaughan, 2006; Otte and Rousseau, 2002; Owen-Smith et al., 2002). The co-authorship analysis will be able to illustrate the evolution of collaboration and social networks among researchers. Furthermore, key actors (either individuals or groups) and relationships among these actors will be identified (Fatt et al., 2010; Lu et al., 2010; Perianes-Rodriguez et al., 2010; Said et al., 2011).

A network is made up of relational ties that are used to allocate resources in a social structure. Social network is defined as social structure which is made up of a set of social
actors and the social interactions between actors (Wellman and Berkowitz, 1988). Major components in social network are actor and relational tie. Actor in social network corresponds to a node, which can be individual or a group. The relational tie is the relationships among actors, which includes collaboration, friendship, exchange, etc. Network ties are the stable, relational patterns that represent an individual’s position within a network (Stevenson and Greenberg, 2000).

The collaboration among authors can be identified in co-authorship network, which occurs when two or more authors co-publish a study (Lu and Wolfram, 2012). In social network analysis research, there are three important social capital indicators, including degree centrality, closeness centrality and betweenness centrality (Li et al., 2013). Those indicators will be used in this study to demonstrate the authors’ importance in co-authorship network of Library Hi Tech community.

Prior studies on co-authorship social network analysis were reviewed as follows. First, Newman (2004) illustrated the co-authorship networks in three fields: Biology, Physics and Mathematics. The results revealed that most researchers have only a few coauthors, while a few have many. Biological scientists tend to have more coauthors than mathematicians or physicists. Leydesdorff and Wagner (2008) explored the network of global collaborations as measured by co-authorship relations on refereed papers from Science Citation Index and the results suggested that the global network of science is increasing. Furthermore, Badar et al. (2013) examined the co-authorship network of Pakistani researchers publishing in Chemistry and its sub-fields retrieved from Institute for Scientific Information (ISI) Web of Science (WOS). The results suggested that the network centrality positively predicts research performance. Koseoglu (2016) addressed the author collaborations from articles published in the Strategic Management Journal. The authorship patterns, author productivity, ranking of authors and visualization of the co-authorship network were demonstrated and core authors were identified.

In summary, the research questions that current study aims to address include:

RQ1. What is the co-authorship network of Library Hi Tech?

RQ2. Who are the key actors in the librarian research field?

RQ3. What is the major trend of publication of Library Hi Tech in terms of the collaboration among researchers?

The contribution of this study is illustrated as follows:

- the co-authorship network of Library Hi Tech identified in this study can contribute to researchers who need to identify the key researchers and related social network in different research areas;
- this study can provide the collaboration patterns of scientific communities in librarian research field; and
- the research methods applied in this study can provide academic contributions for future research by using similar methodology to explore academic social networks in different research fields.

2. Data and methodology

The co-author network structure and collaboration among authors in Library Hi Tech journal was demonstrated in this study. Library Hi Tech was first published by Emerald in 1983. With more than 30 years of history, the journal published up-to-date articles regarding technology-assisted information systems that support libraries and cultural memory, education and the academy, health and medicine and government and citizenship issues and
is the pioneer in the library domain. Articles published in *Library Hi Tech* provide new insights for library information generated by the development of information technologies.

Data set to be analyzed in this paper was retrieved from WOS, which provides a comprehensive citation search and supports reference cross-disciplinary research. WOS as a scientific citation indexing service was originally produced by the ISI and covered a variety of citation databases including Conference Proceedings Citation Index, Science Citation Index, Social Sciences Citation Index, Arts & Humanities Citation Index, Book Citation Index, etc.

The data were downloaded from the WOS website on November 1, 2017. The data set of this paper was generated as follows. First, “*Library Hi Tech*” was used as the search term of publication name and the document type was restricted to only the “article” type. The publication time ranged from 2006 to 2017. The *Library Hi Tech* was first published in 1984, while the data collected in study was restricted to 2006–2017 which covered all the *Library Hi Tech* articles included in the WOS database.

In total, 522 records were found. The BibExcel program was used to conduct the co-authorship analysis and generate the degree centrality, closeness centrality and betweenness centrality. The authorship and co-authorship analysis via network visualizations and analyses were performed using Pajek, network analyses software package.

3. Data analysis

The data analysis result was described as follows. First, the distribution of number of articles across years was examined. Productivity in terms of author appearance, country and year was also analyzed in the first section. Second, the analysis on co-authorship was performed, followed by social network analysis. Finally, the fourth section revealed the cohesive subgroup analysis.

3.1 The distribution of articles across years

A descriptive analysis on 522 articles retrieved from WOS database was conducted and the distribution of number of articles in past decade was illustrated in Figure 1. The result indicated that *Library Hi Tech* published the highest number (50) of articles in 2011, while the number of articles was consistently around 45 in general.

Table I described the most productive authors according to the top 3 author appearance. According to Table I, in the past 12 years Seadle is the most productive author in *Library Hi Tech* who published 18 articles, all of which in the first author appearance. Professor Seadle is Director of the School and Deputy Dean of the Faculty of Humanities of Humboldt University in Germany. His research focuses on digital libraries, evaluating digital libraries, user studies, long-term preservation, open archives initiative, copyright and information science history. He also served as editor of a variety of journals in library science fields.

![Figure 1. The number of published papers each year](image-url)
Professor Fourie also published 18 articles in *Library Hi Tech*. Fourie is Professor in the Department of Information Science at the University of Pretoria. Her research concerns information literacy, information retrieval, information organization, information behavior and information seeking. The second productive author is Professor Noh from Department of Library and Information Science, Konkuk University, who published six articles in *Library Hi Tech*. His major research interest is information retrieval technique.

The country from which the authors came was also analyzed and the result is shown in Table II. The results indicated that the USA is the most productive country, with 233 articles (44.64 percent) published in *Library Hi Tech*, followed by Germany that contributed 46 articles (8.81 percent) published. Asian countries account for limited proportions of articles, of which the top 3 Asian countries are People’s Republic of China (7.09 percent), Taiwan (3.64 percent) and South Korea (3.26 percent). This suggested that authors from Asian countries still have room for more publication in librarian research field.

Table III captures the number of authors, number of articles and authors per article. The results showed that authors per article increased from 1.67 in 2006 to 2.31 in 2017. Dividing the period from 2006 to 2017 into two subperiods will be able to show more fully significant changes and trends in authorship analysis. The number of authors per article in the second period (2013–2017) is doubled than that in the first period (2006–2012). This result showed the trend that more and more researchers in recent years conduct research in collaboration with others.

3.2 Co-authorship analysis

Figure 2 illustrated the co-authorship analysis result. In the period between 2006 and 2008, the number of articles written by a single author was greater than the number of authors per article. Dividing the period from 2006 to 2017 into two subperiods will be able to show more fully significant changes and trends in authorship analysis. The number of authors per article in the second period (2013–2017) is doubled than that in the first period (2006–2012). This result showed the trend that more and more researchers in recent years conduct research in collaboration with others.

<table>
<thead>
<tr>
<th>Author</th>
<th>1st author</th>
<th>2nd author</th>
<th>3rd author</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seadle, M.</td>
<td>18</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Fourie, I.</td>
<td>15</td>
<td>3</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Noh, Y.</td>
<td>6</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Hendricks, A.</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Zha, X.J.</td>
<td>3</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Chen, J.</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Lu, Q.</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Yan, Y.L.</td>
<td>1</td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total count</td>
<td>51</td>
<td>5</td>
<td>6</td>
<td>62</td>
</tr>
</tbody>
</table>

Table I. The most productive authors in terms of author appearance

<table>
<thead>
<tr>
<th>No.</th>
<th>Countries</th>
<th>Counts</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>233</td>
<td>44.64</td>
</tr>
<tr>
<td>2</td>
<td>Germany</td>
<td>46</td>
<td>8.81</td>
</tr>
<tr>
<td>3</td>
<td>People’s Republic of China</td>
<td>37</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>Canada</td>
<td>31</td>
<td>5.94</td>
</tr>
<tr>
<td>5</td>
<td>South Africa</td>
<td>24</td>
<td>4.60</td>
</tr>
<tr>
<td>6</td>
<td>England</td>
<td>20</td>
<td>3.83</td>
</tr>
<tr>
<td>7</td>
<td>Taiwan</td>
<td>19</td>
<td>3.64</td>
</tr>
<tr>
<td>8</td>
<td>South Korea</td>
<td>17</td>
<td>3.26</td>
</tr>
<tr>
<td>9</td>
<td>India</td>
<td>14</td>
<td>2.68</td>
</tr>
<tr>
<td>10</td>
<td>Spain</td>
<td>12</td>
<td>2.290</td>
</tr>
</tbody>
</table>

Table II. Top 10 highly productive countries
multi-authored articles. After this period, the number of multiple-authored articles increased, the highest number of multi-authored articles included two authors, followed by three authors.

Table IV summarized the collaboration rates in *Library Hi Tech* each year. Findings suggested that the collaboration rate increased from 47.92 in 2006 to 58.62 percent in 2017, with the average of 57.09 percent. The collaboration rate in recent five years is more than 60 percent, indicating an increasing trend of collaboration among researchers in librarian research field.

3.3 Social network analysis

This section examined the social network analysis results in terms of three important social capital indicators: degree centrality, closeness centrality and betweenness centrality (Li *et al.*, 2013).

Degree centrality identified how many collaborators a researcher works with (Koseoglu, 2016). More collaborators indicate higher degrees of influence, information flow and information
exchange activities in a scientific community (Yan and Ding, 2009). In other words, degree centrality provides the information regarding the communication activity and popularity of actors (Abbasi et al., 2011). Degree centrality denotes the number of links connected to a node in the network. Degree centrality was calculated according to the following equation:

\[ CD_{ni} = \sum_{j=1}^{g} z_{ij} \]

where \( d(n_i) \) is the degree of node \( n_i \), and \( g \) indicates the total number of actors in the network. Nodes with higher degree centrality are more central to the structure and tend to have a greater capacity to influence others (Yan and Ding, 2009).

While \( g \) will influence degree centrality, it might be difficult to make comparison among different networks. Thus, the normalized value of degree centrality of actors in the network is a better indicator and can be calculated by the following equation:

\[ CD_{0i} = \frac{d(n_i)}{g-1} \]

Degree centrality in current study indicates the number of different researchers that are directly connected with, and is obtained by identifying and subsequently quantifying relationships of co-authorship. It is a measure that reflects the extent of collaboration maintained by authors.

Closeness centrality was used to measure the extent of author influence in social network and it explains how long information flows from an author to others (Yan and Ding, 2009). Closeness is a metric of “how long it will take information to spread from a given vertex to others in the network” (Yin et al., 2006, p. 1603). Closeness centrality was calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. The more central a node is, the closer it is to all other nodes. The equation of closeness centrality is as follows:

\[ C_c(n_i) = \frac{1}{\sum_{j=1}^{g} d(n_i, n_j)} \]

where \( d(n_i, n_j) \) is the length of the shortest path between two nodes.
The normalized value of closeness centrality is calculated as follows:

\[ C'_c(n_i) = \frac{(g-1)}{\sum_{j=1}^{g} d(n_i, n_j)} \]

“Closeness is a surrogate measure for the independence and efficiency for communicating with other nodes in the network” (Abbasi et al., 2011, p. 597). Thus, authors with higher closeness centrality show a larger extent of influence over the entire co-authorship network.

Betweenness centrality describes an author’s capacity to connect other authors within the network (Acodo et al., 2006). High betweenness centrality of an author indicates a crucial role to connect different groups in the community (Yin et al., 2006). The betweenness centrality can be calculated by the following equation:

\[ C_B(n_i) = \sum_{j < k} g_{jk}(n_i)/g_{jk} \]

where \( g_{jk}(n_i) \) is all geodesics linking node \( j \) and node \( k \) which pass through node \( i \) and \( g_{jk} \) is the geodesic distance between the vertices of \( j \) and \( k \).

The normalized value of the betweenness centrality of a node in a network is calculated by the following formula:

\[ C_B = \frac{\sum_{i=1}^{g} [C_B(n_i^*) - C_B(n_i)]}{(g-1)} \]

In co-authorship networks, authors with high betweenness are “broker or gatekeeper” who controls the information flow and is able to make interconnections among authors in different research fields. Thus, the betweenness centrality indicated the authors’ engagement in research of different fields. High betweenness centrality shows the interdisciplinarity of authors.

The descriptions of the above social capital indicators are summarized in Table V.

Table VI listed the top 30 ranking of authors in terms of degree, closeness and betweenness centrality. The result indicated a collaborative network among those three scholars who play an important role in Library Hi Tech to influence others. First, the results showed that Professor Markey, K. holds the highest position in ranking in terms of degree and closeness centrality, and ranked no. 2 in betweenness centrality. Fourie, I. was ranked no. 2 in degree and closeness centrality, but owns the first position in betweenness centrality. Professor Li, X. is one of the top 3 author in all centrality indicators. Those three researchers are popular actors, with larger influence and are able to connect researcher from other research fields in the Library Hi Tech network.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Researcher's name</th>
<th>Degree centrality</th>
<th>Rank</th>
<th>Researcher's name</th>
<th>Closeness centrality</th>
<th>Rank</th>
<th>Researcher's name</th>
<th>Betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Markey, K.</td>
<td>9</td>
<td>1</td>
<td>Markey, K.</td>
<td>0.0145</td>
<td>1</td>
<td>Fourie, I.</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td>Fourie, I.</td>
<td>8</td>
<td>2</td>
<td>Li, X.</td>
<td>0.0119</td>
<td>2</td>
<td>Markey, K.</td>
<td>0.0001</td>
</tr>
<tr>
<td>3</td>
<td>Li, X.</td>
<td>7</td>
<td>3</td>
<td>Li, X.</td>
<td>0.0119</td>
<td>3</td>
<td>Li, X.</td>
<td>0.0001</td>
</tr>
<tr>
<td>4</td>
<td>Zhuang, X.</td>
<td>7</td>
<td>4</td>
<td>Rosenberg, V.</td>
<td>0.0119</td>
<td>4</td>
<td>Chen, J.</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>Rosenberg, V.</td>
<td>7</td>
<td>5</td>
<td>Jenkins, A.</td>
<td>0.0119</td>
<td>5</td>
<td>Lee, S.</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>Jenkins, A.</td>
<td>7</td>
<td>6</td>
<td>St Jean, B.</td>
<td>0.0119</td>
<td>6</td>
<td>Sonnenwald, D.</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>Xing, W.</td>
<td>7</td>
<td>7</td>
<td>Frost, R.</td>
<td>0.019</td>
<td>7</td>
<td>Bothma, T.</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>Si, L.</td>
<td>7</td>
<td>8</td>
<td>Yao, X.</td>
<td>0.019</td>
<td>8</td>
<td>Huang, X.</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>St Jean, B.</td>
<td>7</td>
<td>9</td>
<td>Swanson, F.</td>
<td>0.019</td>
<td>9</td>
<td>Zhou, L.</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>Frost, R.</td>
<td>7</td>
<td>10</td>
<td>Jennings, B.</td>
<td>0.019</td>
<td>10</td>
<td>Liu, G.</td>
<td>0.0000</td>
</tr>
<tr>
<td>11</td>
<td>Yao, X.</td>
<td>7</td>
<td>11</td>
<td>Zhuang, X.</td>
<td>0.019</td>
<td>11</td>
<td>Resnick, T.</td>
<td>0.0000</td>
</tr>
<tr>
<td>12</td>
<td>Swanson, F.</td>
<td>7</td>
<td>12</td>
<td>Xing, W.</td>
<td>0.019</td>
<td>12</td>
<td>Zhang, J.</td>
<td>0.0000</td>
</tr>
<tr>
<td>13</td>
<td>Jennings, B.</td>
<td>7</td>
<td>13</td>
<td>Si, L.</td>
<td>0.019</td>
<td>13</td>
<td>Carrigan, E.</td>
<td>0.0000</td>
</tr>
<tr>
<td>14</td>
<td>Sonnenwald, D.</td>
<td>6</td>
<td>14</td>
<td>Sonnenwald, D.</td>
<td>0.0102</td>
<td>14</td>
<td>Burford, N.</td>
<td>0.0000</td>
</tr>
<tr>
<td>15</td>
<td>Xin, J.</td>
<td>6</td>
<td>15</td>
<td>Xin, J.</td>
<td>0.0102</td>
<td>15</td>
<td>Arlitsch, K.</td>
<td>0.0000</td>
</tr>
<tr>
<td>16</td>
<td>Li, Y.</td>
<td>6</td>
<td>16</td>
<td>Li, Y.</td>
<td>0.0102</td>
<td>16</td>
<td>Hahn, D.</td>
<td>0.0000</td>
</tr>
<tr>
<td>17</td>
<td>Hua, X.</td>
<td>6</td>
<td>17</td>
<td>Lo, P.</td>
<td>0.0102</td>
<td>17</td>
<td>Lu, Q.</td>
<td>0.0000</td>
</tr>
<tr>
<td>18</td>
<td>Chen, J.</td>
<td>6</td>
<td>18</td>
<td>Chen, J.</td>
<td>0.0102</td>
<td>18</td>
<td>Radio, E.</td>
<td>0.0000</td>
</tr>
<tr>
<td>19</td>
<td>Lee, S.</td>
<td>6</td>
<td>19</td>
<td>Lee, S.</td>
<td>0.0102</td>
<td>19</td>
<td>Baker, T.</td>
<td>0.0000</td>
</tr>
<tr>
<td>20</td>
<td>Lo, P.</td>
<td>6</td>
<td>20</td>
<td>Lo, P.</td>
<td>0.0102</td>
<td>20</td>
<td>Monroegulick, A.</td>
<td>0.0000</td>
</tr>
<tr>
<td>21</td>
<td>Chiu, D.</td>
<td>6</td>
<td>21</td>
<td>Chiu, D.</td>
<td>0.0101</td>
<td>21</td>
<td>Badurina, B.</td>
<td>0.0000</td>
</tr>
<tr>
<td>22</td>
<td>Carrigan, E.</td>
<td>6</td>
<td>22</td>
<td>Carrigan, E.</td>
<td>0.0101</td>
<td>22</td>
<td>Sen, B.</td>
<td>0.0000</td>
</tr>
<tr>
<td>23</td>
<td>Burford, N.</td>
<td>6</td>
<td>23</td>
<td>Burford, N.</td>
<td>0.0101</td>
<td>23</td>
<td>Yan, Y.</td>
<td>0.0000</td>
</tr>
<tr>
<td>24</td>
<td>Yan, Y.</td>
<td>6</td>
<td>24</td>
<td>Yan, Y.</td>
<td>0.0090</td>
<td>24</td>
<td>Zha, X.</td>
<td>0.0000</td>
</tr>
<tr>
<td>25</td>
<td>Arlitsch, K.</td>
<td>5</td>
<td>25</td>
<td>Arlitsch, K.</td>
<td>0.0087</td>
<td>25</td>
<td>Adu, K.</td>
<td>0.0000</td>
</tr>
<tr>
<td>26</td>
<td>Canet, C.</td>
<td>5</td>
<td>26</td>
<td>Canet, C.</td>
<td>0.0087</td>
<td>26</td>
<td>Partridge, H.</td>
<td>0.0000</td>
</tr>
<tr>
<td>27</td>
<td>Zhang, J.</td>
<td>5</td>
<td>27</td>
<td>Zhang, J.</td>
<td>0.0087</td>
<td>27</td>
<td>Rowley, J.</td>
<td>0.0000</td>
</tr>
<tr>
<td>28</td>
<td>Cobos, J.</td>
<td>5</td>
<td>28</td>
<td>Cobos, J.</td>
<td>0.0087</td>
<td>28</td>
<td>Liu, Y.</td>
<td>0.0000</td>
</tr>
<tr>
<td>29</td>
<td>Gimenezchornet, V.</td>
<td>5</td>
<td>29</td>
<td>Gimenezchornet, V.</td>
<td>0.0087</td>
<td>29</td>
<td>Han, Z.</td>
<td>0.0000</td>
</tr>
<tr>
<td>30</td>
<td>Carot, A.</td>
<td>5</td>
<td>30</td>
<td>Carot, A.</td>
<td>0.0087</td>
<td>30</td>
<td>Lo, P.</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Notes: Authors appearing in the three metrics are marked in bold; authors appearing in two metrics are marked in italics.
Furthermore, Zhuang, X. was ranked 11th in the closeness centrality ranking, while his ranking increased to 4th in the degree centrality ranking. This finding indicated that Professor Zhuang, X. has high intra-community influence and is a popular actor in the Library Hi Tech community.

Rosenberg, V. and Jenkins, A. show similar pattern in the centrality indicators. They were ranked fifth and sixth in the degree centrality and fourth and fifth in the closeness centrality. This result showed that Rosenberg and Jenkins play important role and have some influence in the co-authorship network, while the collaboration with scholars in other community is limited.

3.4 Cohesive subgroup analysis
Cohesive subgroup analysis is an analysis method that demonstrates the social network structure by grouping actors that are intimately connected with one another. In social groups, the structural cohesion model can be used to demonstrate the collaboration and cohesive groups among authors in one specific research field. Pajek was performed to identify the cohesive subgroup in this study. There are several indicators that can be used to describe the cohesive subgroup analysis: components and k-core, which were illustrated in the following sections.

3.4.1 Component. The first step for a cohesive subgroup analysis is to identify the "component" which indicated closely connected group of actors within the network. Component is a maximal connected subgraph, indicating members' identification with a collectivity. Thus, researchers in one component are highly connected to each other while disconnected to scholars in another component. Cohesiveness in social network demonstrates the collaboration among scholars who were tightly closed with frequent interactions.

In this study, the components were set to at least five. The result generated 27 clusters and the details of each cluster are shown in Table VII. As indicated, clusters 3, 4 and 11 are the largest cluster that each included ten authors. The representative authors are Fourie (Table VII), Si and Jennings, respectively. Each of the clusters has their own co-authorship network, demonstrating the collaboration of authors in each filed.

Figure 2 illustrated the co-authorship network of cluster 11. Among researchers in the cluster, Professor Jennings, B. is the representative actor and he collaborates with seven researchers: Markey, K.; Swanson, F.; Jenkins, A.; St Jean, B.; Rosenberg, V.; Yao, X.X.; and Frost, R.L. They published one article "Designing and testing a web-based board game for teaching information literacy skills and concepts" in 2008 in Library Hi Tech. Furthermore, result in Figure 3 suggested that Professor Markey, K. collaborated with the most number of co-authors in cluster 11.

3.4.2 k-core analysis. The above degree-based analyses of co-authorship network may fail to capture certain vital connectivity among authors. Thus, the cores decomposition of a graph introduced by Seidman (1983) is used in this study to capture the connectedness among authors. The k-core is a maximal subgraph, in which each point is at least adjacent to k of other nodes; all the nodes within the k-core have a greater degree than or equal to k (Scott, 2000).

The results from current study of k-core analysis are depicted in Figure 3, which indicated that the k-core value mainly ranged from 5-core to 7-core. Six different tightly connected subgroups in Library Hi Tech community were identified. Comparing the result in Figure 3 and the degree ranking in Table VI, two influential cohesive groups were revealed. The first one is sub-group (d) in Figure 3, which includes eight scholars (Markey, K., Forest, R., Jenkins, A., Rosenberg, V., St Jean, B., Jennings, B., Swanson, F. and Yao, X.) who were ranked within top 13 in Table VI and all appearing in at least two centrality indicators. The second one is sub-group (a) in Figure 4 and is formed by seven researchers...
4. Discussion

The results from this study revealed some interesting findings and the research questions were answered as follows.

First, the co-authorship network in Figure 3 suggested that a large group of authors was connected with each other by Markey, who is Professor in the School of Information at the University of Michigan. Her research focuses on subject access, visual resources and gaming for information literacy. Thus, two groups of researchers who were interested in teaching literacy through a game were connected. Group 1 include eight researchers (Markey, Swanson, Jenkins, Jennings, Jean, Roseenberg, Yao and Frost), while group 2 include three authors (Markey, Leeder and Rieh). This finding provided important knowledge network of gaming for information literacy and thus researchers who are studying similar topics in this area will be able to find appropriate social network to join, or find an important key researcher in this area – Professor Markey. In addition, Professor Markey is also ranked as top 1 researcher in *Library Hi Tech* in terms of degree centrality and closeness centrality, and ranked number 2 by betweenness centrality. This indicated that while Professor Markey is not the most productive author (according to Table I), she is playing an important role in co-authorship network of *Library Hi Tech* connecting researchers who share similar research interests.
Second, the results from Figure 4 suggested six large co-authorship social networks within which researchers in similar research interested were connected. Each co-authorship network focused on different research topics included scientific data sharing platforms (group a), information visualization (group b), library service access (group c), literacy gaming (group d), access to electronic research data (group e) and digital libraries (group f). Thus, researchers who need to address the above research topics will be able to find possible partners from major social networks. Furthermore, the results from Table III suggested that the average number of authors per article increased from 1.67 in 2006 to 2.31 in 2017, while Figures 3 and 4 revealed large social networks that some research topics need more researchers from different areas involved. In the future, cross-field cooperation is expected to increase as librarian-related studies need to incorporate researchers in different fields such as computer technology, culture difference, data analysis, art, education, etc.

Third, the authorship analysis of this study suggested that the number of multi-authored articles published in *Library Hi Tech* are increasing, with the majority of the articles are collaborated by two authors, followed by three authors. This result is consistent with similar study conducted by Koseoglu (2016) in *Strategic Management Journal* which showed that “multi-authored articles dominated solo work, and this domination increased over the past periods.” Most of the multi-authored articles are limited to papers with two or three authors.

Furthermore, Amjad *et al.* (2018) conducted an interesting research that examined the relationship between topic drift and the scientific impact of an author. The results suggested that authors who stick to one topic showed higher impact than authors work on multiple topics. The study provided a future direction for co-authorship network analysis research that the combination of studies on research topics and co-authorship may provide a more insightful trend for the knowledge generation and collaboration network in scientific research.
5. Conclusion

The objective of this study is to explore the co-authorship network in *Library Hi Tech* community. The results from descriptive analysis suggested that around 42 percent of the articles were written by single author, and a limited number of articles were written by more than six co-authors. The result also showed an increasing trend of multi-authored articles, suggesting the collaboration among researchers in librarian research field becomes popular.

The results from three social capital indicators (degree centrality, closeness centrality and betweenness centrality) revealed that Professor Markey, K. is the most important scholar in this community in terms of the degree centrality (indicating popularity in the network), closeness (indicating the author influence in the network) and betweenness (indicating the interdisciplinarity of authors in the network). Second, Professors Fourie, I. and Li, X. were also ranked high in betweenness centrality, suggesting that they are important connectors among different research fields that link researchers from different subgroups. Furthermore, the cohesive subgroup analysis indicators (component and k-core) revealed several subgroups in the community in which the authors are closely connected with each other within the subgroups. Two groups of authors showed significant importance in the *Library Hi Tech* community that are frequently collaborated, and play important role to link different subgroups.

As researchers’ collaboration can be described from different perspectives to show a more comprehensive co-authorship network, the contribution of this study is to illustrate the application of different metrics in the analysis of co-authorship in *Library Hi Tech* in past 12 years. The results identified key actors, important subgroups and the research trend.
References


Yu-Wen Huang is currently doctoral student at the Department of Information Management, National Chung Cheng University in Taiwan. She received her MBA Degree in Information Management from Nanhua University. She is interested in e-commerce and MIS-related research topics.

Dr Chin-Shan Wu is currently Professor at the Department of Information Management, National Chung Cheng University in Taiwan. He received his PhD Degree in Information Management from National Sun Yat-sen University in 2005. Dr Wu has published in academic conferences and referred journals including *Information & Management, Decision Support Systems, Electronic Commerce and Research Applications, Computers in Human Behavior, Cyberpsychology & Behavior, Behavior & Information Technology* and others. His research focuses on e-commerce, management information systems, consumer behavior, decision making and internet marketing. Dr Chin-Shan Wu is the corresponding author and can be contacted at: cswu.mis@gmail.com
Tolerance analysis in scale-free social networks with varying degree exponents

Kwok Tai Chui
Department of Electronic Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong, and
Chien-wen Shen
National Central University, Taoyuan City, Taiwan

Abstract

Purpose – There are many complex networks like World-Wide Web, internet and social networks have been reported to be scale-free. The major property of scale-free networks is their degree distributions are in power law form. Generally, the degree exponents of scale-free networks fall into the range of (2, 3). The purpose of this paper is to investigate other situations where the degree exponents may lie outside the range.

Design/methodology/approach – In this paper, analysis has been carried out by varying the degree exponents in the range of (0.5, 4.5). In total, 243 scenarios have been generated with varying network size of 1,000, 2,000 and 4,000, and degree exponents in the range of (0.5, 4.5) using interval of 0.05.

Findings – The following five indicators have been investigated: average density, average clustering coefficient, average path length, average diameter and average node degree. These indicators vary with the network size and degree exponent. If certain indicators do not satisfy with the user requirement using degree exponents of (2, 3), one can further increase or decrease the value with tradeoff. Results recommend that for degree exponents in (0.5, 2), 26 possible scale-free networks can be selected whereas for (3, 4.5), 41 possible scale-free networks can be selected, assuming a 100 percent deviation on the network parameters.

Originality/value – A tolerance analysis is given for the tradeoff and guideline is drawn to help better design of scale-free network for degree exponents in range of (0.5, 2) and (3, 4.5) using network size 1,000, 2,000 and 4,000. The methodology is applicable to any network size.

Keywords Library networks, Social network, Tolerance analysis, Complex network, Degree exponent, Scale-free network

Paper type Research paper

1. Introduction

Scale-free networks have gained more and more interest in the light of a research paper by Barabasi and Albert (1999) (Aziz et al., 2016). Many networks were found to be scale-free, for instance, World-Wide Web (Barnett and Jiang, 2016; Wang et al., 2014), Internet of Things (Qiu et al., 2017; Guo et al., 2013), e-mail (Portela et al., 2016; Yang and Yang, 2014), wireless sensor networks (Huang et al., 2017; Peng et al., 2016), crisis informetrics (Hossain et al., 2015), semantic network (Steyvers and Tenenbaum, 2005), patent litigation (Lee et al., 2017) and food web (Layman et al., 2015; Navia et al., 2016). They are characterized by the degree distribution which follows power law in terms of P(k) ~ k^{-\gamma} for large k where \gamma denotes the degree exponent of the network. The error and attack on the efficiency of scale-free network can be found in Crucitti et al., 2003. In this paper, simulation and analysis have been carried out with degree exponents 0.5 \leq \gamma \leq 4.5 which are 1.5 beyond the typical range of consideration. Scale-free networks will be generated based on variation of two parameters, n and \gamma. For network size, the values are 1,000, 2,000 and 4,000 whereas the possible \gamma is set using 0.05 interval. Thus, 243 cases can be formed. Five common parameters, average density, average clustering coefficient, average path length, average diameter and average node degree, are being selected for performance evaluation of the networks. Also, the goal is to perform tolerance analysis in 0.5 \leq \gamma \leq 2 and 3 < \gamma \leq 4.5 in order to explain the tradeoff when \gamma exceeds the boundary, 2 \leq \gamma \leq 3.
The contribution can be summarized as:

1. Analysis of scale-free networks with degree exponents $0.5 \leq \gamma \leq 4.5$ using five common network parameters, average density, average clustering coefficient, average path length, average diameter and average node degree.

2. Tolerance analysis includes a tolerance index that has been first proposed to give a guideline for users who want to design a network beyond typical degree exponents $2 \leq \gamma \leq 3$. Results recommend that for $0.5 \leq \gamma < 2$, 26 possible scale-free networks can be selected. For $3 < \gamma \leq 4.5$, 41 possible scale-free networks can be selected.

The research and practical implications of this paper is that it is applicable to the study of network of library services. Social network, computer network and semantic network are involved for library services, depending on how the analysis is formulated. The properties of growing and preferential attachment of scale-free networks are essential for library services. Growing is related to adding or removing nodes from the network. The preferential attachment defines the probability that the new nodes connect to the existing nodes.

This report is organized as follows. Literature review is first summarized in Section 2. The methodology of generating scale-free networks and tolerance analysis will be given in Section 3. The performance evaluation of the generated scale-free networks will be discussed in Section 4. In Section 5, tolerance analysis will be made. Finally, a conclusion is drawn in Section 6.

2. Literature review

Typically, many research papers have reported that the degree exponents in real-world networks follow $2 \leq \gamma \leq 3$ (Aziz et al., 2016; Navia et al., 2016; Peng et al., 2016). However, it can be seen from literature that some networks can have degree exponents outside the boundaries. Examples were disease outbreaks (Levin and Finley, 2017), Facebook (Timár et al., 2016), epidemic process (Castellano and Pastor-Satorras, 2016), protein domain (Park et al., 2001), food web (Montoya and Solé, 2002), co-authorship in HEP (Newman, 2001), human sexual-contacts (Liljeros et al., 2001). Table I summarizes the network size ($n$) and degree exponent ($\gamma$) of each scale-free network.

Those networks summarized in Table I play important roles in today’s era. Thanks to the growth of computation power, big data and Internet-of-Things technologies, more data can be captured and analyzed. Therefore, it is believed that the general value of the degree exponents for real-world scale-free network is no longer follow $2 \leq \gamma \leq 3$ and the boundary should be extended. Yet, based on authors understanding, there is no analysis on the network properties for scale-free network with degree exponents in range of (0.5, 2) and (3, 4.5).

3. Methodology

The methodology comprises of two parts. First, scale-free networks are generated based on different settings of $n$ and $\gamma$. Second, tolerance analysis is performed on the ranges $0.5 \leq \gamma \leq 2$ and $3 < \gamma \leq 4.5$.

<table>
<thead>
<tr>
<th>Network</th>
<th>Network size, $n$</th>
<th>Degree exponent, $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease outbreaks (Levin et al., 2017)</td>
<td>n/a</td>
<td>1.8</td>
</tr>
<tr>
<td>Facebook (Timár et al., 2016)</td>
<td>10,000</td>
<td>1</td>
</tr>
<tr>
<td>Epidemic process (Castellano and Pastor-Satorras, 2016)</td>
<td>10,000</td>
<td>3.5</td>
</tr>
<tr>
<td>Co-authorship in HEP (Newman, 2001)</td>
<td>56,627</td>
<td>1.2</td>
</tr>
<tr>
<td>Human sexual-contacts (Liljeros et al., 2001)</td>
<td>2,810</td>
<td>3.4</td>
</tr>
<tr>
<td>Protein domain (Park et al., 2001)</td>
<td>876</td>
<td>1.6</td>
</tr>
<tr>
<td>Food web (Montoya and Solé, 2002)</td>
<td>154</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Table I. The network size and degree exponent of selected scale-free networks
3.1 Generating scale-free networks

Various scale-free networks are generated using different settings of $n = (1,000, 2,000, 4,000)$ and $\gamma = (0.5, 0.55, \ldots, 0.445, 0.45)$. A total of 243 cases are formed. It is noted that the maximum $n$ is set as 4,000 for simplicity and faster simulation and the methodology and tolerance analysis can be applied to any value of $n$. In order to obtain a fair analysis, each network will be simulated ten times and the averaging of five typical network parameters, average density ($D_{\text{avg}}$), average clustering coefficient ($C_{\text{avg}}$), average path length ($L_{\text{avg}}$), average diameter ($d_{\text{avg}}$) and average node degree ($k_{\text{avg}}$) are computed via following equations. Table II summarizes the simulation details.

The formulations of $D_{\text{avg}}$, $C_{\text{avg}}$, $L_{\text{avg}}$, $d_{\text{avg}}$, $k_{\text{avg}}$ are illustrated as follows:

1. Network density defines as the portion of the connection possibilities in a network that are actual connections. The density of a $p$th network $D_p$ is defined as a ratio of the number of edges $E_p$ to the number of possible edges. Thus, we have:

$$D_p = \frac{2E_p}{n(n-1)}, \quad p = 1, \ldots, 10. \quad (1)$$

Since every network will be repeatedly simulated ten times, the average density can be further obtained:

$$D_{\text{avg}} = \frac{1}{10} \sum_{p=1}^{10} D_p. \quad (2)$$

2. Clustering is an important property of scale-free networks. Here is an obvious example, people tend to make friends who are also friends with each other, resulting in sets of people among which many edges exist, while a set made from randomly chosen people would have a much smaller number of edges between them. To measure the clustering in a network, a common measure is the clustering coefficient. The clustering coefficient of the $i$th node of $j$th network is given by:

$$C_{i,p} = \frac{2E(i,p)}{k(i,p)(k(i,p)-1)}, \quad p = 1, \ldots, n, \quad (3)$$

where $n$ is the network size, $k(i)$ is the number of neighbors in $i$th node and $E(i)$ is the number of edges that $k(i)$ has. The average clustering coefficient in ten networks can then be computed as:

$$C_{\text{avg}} = \frac{1}{10} \sum_{p=1}^{10} \left( \frac{1}{n} \sum_{i=1}^{n} C_{i,p} \right). \quad (4)$$

3. Path length is a measure of the efficiency of information transport on a network in which a short average path length supports quick transfer of information and thus

<table>
<thead>
<tr>
<th>Network size, $n$</th>
<th>Degree exponent, $\gamma$</th>
<th>Number of simulations</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000, 2,000, 4,000</td>
<td>(0.5, 4.5) with interval 0.05</td>
<td>10</td>
<td>$D_{\text{avg}}, C_{\text{avg}}, L_{\text{avg}}, d_{\text{avg}}, k_{\text{avg}}$</td>
</tr>
</tbody>
</table>

Table II. Simulation details of the scale-free network
reduces costs. Let $d_{i,j,p}$ be the distance between nodes $i$ and $j$ of $p$th network. The average path length of $p$th network is:

$$L_p = \frac{1}{0.5(n)(n-1)} \sum_{i < j} d_{i,j,p}. \quad (5)$$

Similarly, the average path length of ten networks is thus:

$$L_{avg} = \frac{1}{10} \sum_{p=1}^{10} L_p. \quad (6)$$

(4) The diameter of a network defines as the length of the longest of all the computed shortest paths between all pair of nodes within the network. It describes the linearity of the network. The diameter of the $p$th network can be calculated by:

$$d_p = \max_{i < j} \{d_{i,j,p}\}. \quad (7)$$

The average diameter of ten networks is further computed as:

$$d_{avg} = \frac{1}{10} \sum_{p=1}^{10} d_p. \quad (8)$$

(5) The degree $k$ of a node is the number of edges connected to this node. Different nodes have different degrees, this variability is characterized by the degree distribution $P(k)$, which provides the probability that a node has exactly $k$ edges. The average node degree of $p$th network is measured by:

$$k_p = \frac{1}{n} \sum_{i=1}^{n} k(i). \quad (9)$$

Thus, the average node degree of ten networks can be obtained:

$$k_{avg} = \frac{1}{10} \sum_{p=1}^{10} k_p. \quad (10)$$

3.2 Tolerance analysis

In this subsection, it is assumed that users prefer the degree exponent follows $2 \leq \gamma \leq 3$ but still allow some tolerances $\Delta_{\gamma=2}$ and $\Delta_{\gamma=3}$ in the variation of $D_{avg}$, $C_{avg}$, $L_{avg}$, $d_{avg}$ and $k_{avg}$. $\Delta_{\gamma=2}$ denotes the tolerance index for percentage difference between the network parameters in $0.5 \leq \gamma < 2$ and $\gamma = 2$. Likewise, $\Delta_{\gamma=3}$ denotes the tolerance index for percentage difference between the network parameters in $3 < \gamma \leq 4.5$ and $\gamma = 3$. The results for $\Delta_{\gamma=2}$ and $\Delta_{\gamma=3}$ are useful to guide users to select acceptable scale-free networks outside $2 \leq \gamma \leq 3$. For instance, users set the maximum allowable $\Delta_{\gamma=2}$, then they can refer to the tolerance matrix tables (see Tables AI–AVI and to be discussed in Section 4). They can pick any network with every network parameter obeys the maximum allowable $\Delta_{\gamma=2}$.
The analysis is divided into two parts, $0.5 \leq \gamma < 2$ and $3 < \gamma \leq 4.5$:

1. $0.5 \leq \gamma < 2$

The $\Delta_{y^2}$ for the five network parameters can be computed as:

$$\Delta_{y^2} = \frac{100(D_{avg,\gamma} - D_{avg,\gamma=1})}{D_{avg,\gamma}}$$

2. $3 < \gamma \leq 4.5$

Similarly, $\Delta_{y^3}$ can be calculated using:

$$\Delta_{y^3} = \frac{100(D_{avg,\gamma} - D_{avg,\gamma=3})}{D_{avg,\gamma}}$$

Then, for both $\Delta_{y^2} = 2$ and $\Delta_{y^3} = 3$, the values will be further rounded up to the nearest value in $(5, 10, 95, 100)$ for any computed value in Equation (11) or Equation (12) lies within $(0, 100)$. For computed value $> 100$, $\Delta_{y^2} = 2$ or $\Delta_{y^3} = 3$ will be assigned as “n/a.”

4. Performance evaluation of generated scale-free networks

The simulation platform for the scale-free networks is MATLAB R2011a. It is worth mentioning that the ComplexNetworksPackage Toolbox has utilized during the simulation (Muchnik, 2013). Section 3 comprises five parts, which analyzes and discusses the performance of scale-free networks in each of the network parameters, $D_{avg}$, $C_{avg}$, $L_{avg}$, $d_{avg}$ and $k_{avg}$.

4.1 Average density $D_{avg}$

The average density $D_{avg}$ over different network size, $n = 1,000, 2,000$ and $4,000$ for degree exponent $0.5 \leq \gamma \leq 4.5$ is shown in Figure 1. It can be seen that for all network sizes, $D_{avg}$ of scale-free network decreases when $\gamma$ increases. This is trivial because fewer edges forms when $\gamma$ increases.

The slope of the curves is deepest in $0.5 \leq \gamma < 2$, follows by $2 \leq \gamma < 3$ and $3 < \gamma \leq 4.5$. It is noted that there is a sudden drop in average density in $0.7 \leq \gamma \leq 0.8$. Also, the slope of the curves in $2 < \gamma \leq 4.5$ is approximately steady with only a small variation when $\gamma$ increases.

4.2 Average clustering coefficient $C_{avg}$

The average clustering coefficient $C_{avg}$ over different network size, $n = 1,000, 2,000$ and $4,000$ for degree exponent $0.5 \leq \gamma \leq 4.5$ is shown in Figure 2. The shape of three curves are bell-shaped centered at $\gamma = 1.95, 1.95$ and 2 for $n = 1,000, 2,000$ and 4,000, respectively.

Tolerance analysis in scale-free social networks
For $0.5 \leq \gamma \leq 1.5$, $C_{\text{avg}}$ is fluctuating in which the curves increase and decrease so that there are local maximum and minimum points. The curves are symmetric between $1.5 < \gamma \leq 2.5$. For $2.5 < \gamma \leq 4.5$, the slope of the curves is approximately steady with only a small variation when $\gamma$ increases.

4.3 Average path length $L_{\text{avg}}$

The average path length $L_{\text{avg}}$ over different network size, $n = 1,000$, 2,000 and 4,000 for degree exponent $0.5 \leq \gamma \leq 4.5$ is shown in Figure 3. As $\gamma$ increases from 0.5 to 2, average path length does not vary because the number of edges in the network can be considered as "massive" and more and more nodes are directly interconnected. The threshold is at $\gamma = 2$. As $\gamma$ increases from 2 to 4.5, fewer and fewer nodes are directly interconnected, the average path length increases.
4.4 Average diameter $d_{avg}$

The average diameter $d_{avg}$ over different network size, $n = 1,000, 2,000$ and $4,000$ for degree exponent $0.5 \leq \gamma \leq 4.5$ is shown in Figure 4. It can be seen that the shape and pattern of Figures 3 and 4 are similar. Since $d_{avg}$ considers the maximum path length between every possible nodes, the value in Figure 3 is higher than that in Figure 4. It is worth mentioning that for $0.5 \leq \gamma < 2$, $d_{avg}$ is approximately constant and equals to 4, it can be deduced that the scale-free network will never become a fully-connected network for any $\gamma$.

4.5 Average node degree $k_{avg}$

The average node degree $d_{avg}$ over different network size, $n = 1,000, 2,000$ and $4,000$ for degree exponent $0.5 \leq \gamma \leq 4.5$ is shown in Figure 5. It can be seen that the shape and pattern
of Figures 1 and 5 are similar. Since higher density implies more number of edges, the node degree increases. The same finding can be drawn as in average density, for average node degree that in $2 < \gamma \leq 4.5$, the average node degree is approximately steady.

5. Tolerance analysis of generated scale-free networks
The tolerance analysis of generated scale-free networks is carried out in two parts. First, the cases for degree exponent $0.5 \leq \gamma < 2$ at $n = 1,000, 2,000$ and $4,000$ are considered. Second, the cases for degree exponent $3 < \gamma \leq 4.5$ are considered. Aforementioned, it is assumed that users prefer the degree exponent follows $2 < \gamma \leq 3$ but still allow some tolerances in the variation of network parameters. It is noted that based on authors' understanding, tolerance analysis of scale-free network with degree exponents beyond $(2,3)$ is first addressed. In addition, in this paper, we are not proposing new algorithm in which it is not related to any improvement in accuracy, computational power, complexity, etc. As a result, performance comparison is not available.

5.1 Tolerance analysis with $0.5 \leq \gamma < 2$
It is recommended users to select scale-free networks that do not contain any network parameter with $\Delta_{\gamma \geq 2} > 100$, labels as “n/a.” Table III summarizes the acceptable (recommended) scale-free networks using $n = 1,000, 2,000$ and $4,000$. For the details of tolerance matrix with $0.5 \leq \gamma < 2$, using $n = 1,000, 2,000$ and $4,000$, it is suggested to refer to Tables AI–AIII.

It can be seen that for $n = 1,000$, scale-free networks with $1.5 \leq \gamma < 1.95$ (10 possible networks) can be selected. For both $n = 2,000$ and $4,000$, scale-free networks with $1.6 \leq \gamma < 1.95$ (8 possible networks) can be selected. In total, 26 possible networks can be selected in $0.5 \leq \gamma < 2$.

5.2 Tolerance analysis with $3 < \gamma \leq 4.5$
It is recommended users to select scale-free networks that do not contain any network parameter with $\Delta_{\gamma = 3} > 100$, labels as “n/a”. Table IV summarizes the acceptable (recommended) scale-free networks using $n = 1,000, 2,000$ and $4,000$. For the details of
Table III. Recommended scale-free networks with $0.5 \leq \gamma < 2$ and $n = 1,000$, 2,000, and 4,000

<table>
<thead>
<tr>
<th>$n$</th>
<th>$n = 1,000$</th>
<th>$n = 2,000$</th>
<th>$n = 4,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>1.5</td>
<td>1.55</td>
<td>1.6</td>
</tr>
<tr>
<td>$D_{avg}$</td>
<td>90</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>85</td>
<td>85</td>
<td>80</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>60</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>90</td>
<td>75</td>
<td>70</td>
</tr>
</tbody>
</table>
Table IV. Recommended scale-free networks with $3 \leq \gamma \leq 4.5$ and $n = 1,000, 2,000$ and 4,000.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\gamma$</th>
<th>3.1</th>
<th>3.15</th>
<th>3.2</th>
<th>3.25</th>
<th>3.3</th>
<th>3.35</th>
<th>3.4</th>
<th>3.45</th>
<th>3.5</th>
<th>3.55</th>
<th>3.6</th>
<th>3.65</th>
<th>3.7</th>
<th>3.05</th>
<th>3.1</th>
<th>3.15</th>
<th>3.2</th>
<th>3.25</th>
<th>3.3</th>
<th>3.35</th>
<th>3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>70</td>
<td>65</td>
<td>96</td>
<td>60</td>
<td>80</td>
<td>70</td>
<td>95</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>55</td>
<td>70</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>5</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>35</td>
<td>35</td>
<td>45</td>
<td>45</td>
<td>50</td>
<td>60</td>
<td>65</td>
<td>75</td>
<td>5</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>30</td>
<td>45</td>
<td>50</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>30</td>
<td>30</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

| $n$     | $\gamma$ | 3.45 | 3.5 | 3.55 | 3.6 | 3.65 | 3.7 | 3.65 | 3.7 | 3.05 | 3.1 | 3.15 | 3.2 | 3.25 | 3.3 | 3.35 | 3.4 | 3.45 | 3.5 | 3.55 | 3.6 | 3.65 | 3.7 |
|---------|----------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|
| $C_{avg}$ | 40 | 65  | 85  | 70  | 80  | 70  | 25  | 50  | 50  | 60  | 60  | 40  | 55  | 65  | 75  | 70  | 70  | 55  | 85  | 75  | 75  |
| $L_{avg}$ | 45 | 50  | 65  | 70  | 70  | 75  | 5   | 10  | 15  | 20  | 25  | 35  | 35  | 45  | 50  | 55  | 55  | 70  | 75  | 80  |
| $d_{avg}$ | 60 | 65  | 80  | 90  | 100 | 90  | 5   | 15  | 20  | 25  | 30  | 40  | 45  | 50  | 70  | 60  | 70  | 85  | 90  | 100 |
tolerance matrix with $3 < \gamma \leq 4.5$, using $n = 1,000, 2,000$ and $4,000$, it is suggested to refer to Table AIV–AVI in Appendix.

It can be seen that for $n = 1,000$, scale-free networks with $3.1 < \gamma \leq 3.7$ (13 possible networks) can be selected. For both $n = 2,000$ and $4,000$, scale-free networks with $3.05 < \gamma \leq 3.7$ (14 possible networks) can be selected. In total, 41 possible networks can be selected in $3 < \gamma \leq 4.5$ which is much more than that for scale-free networks using $0.5 \leq \gamma < 2$.

6. Conclusions

In this paper, network analysis has been carried out for scale-free networks with degree exponents $(0.5, 4.5)$ using network sizes of 1,000, 2,000 and 4,000. In total, 243 scale-free networks have been analyzed based on five common network parameters, average density, average clustering coefficient, average path length, average diameter and average node degree. These networks have been evaluated and discussed in aspect of each parameter. In addition, tolerance analysis (with maximum of 100 percent deviation) has been proposed to facilitate users to select scale-free network beyond the typical degree exponent $(2, 3)$. Results recommend that for $(0.5, 2)$, 26 possible scale-free networks can be selected. For $(3, 4.5)$, 41 possible scale-free networks can be selected. It is worth mentioning that the tolerance can be adjusted (depend on the user requirement) and this relates to the number of possible scale-free networks beyond degree exponents in $(2, 3)$. The results imply that researchers can devote more efforts in analyzing scale-free networks with degree exponent outside the boundary $(2, 3)$. Especially for the higher end in which the network parameters are more similar to that in typical networks. In addition, it can be concluded that the typical range of degree exponents $(2, 3)$ of scale-free networks should be extended. There are two major limitations in this project which recommended as future work. First, the interval between each possible degree exponents is 0.05 which may be lowered to allow a more detailed analysis. Second, the network size can be increased to an order of 10,000 although the computation time will be significantly increased. The number of possible scale-free networks may vary with the network size but it is believed that the analysis will be valid when the network size increases.

References


Further reading


### Table AI.
Details of tolerance matrix with $n = 1,000$, $0.5 \leq \gamma < 2$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
<th>1</th>
<th>1.05</th>
<th>1.1</th>
<th>1.15</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

### Table AII.
Details of tolerance matrix with $n = 2,000$, $0.5 \leq \gamma < 2$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
<th>1</th>
<th>1.05</th>
<th>1.1</th>
<th>1.15</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

### Table AIII.
Details of tolerance matrix with $n = 4,000$, $0.5 \leq \gamma < 2$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
<th>1</th>
<th>1.05</th>
<th>1.1</th>
<th>1.15</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>3.05</td>
<td>3.1</td>
<td>3.15</td>
<td>3.2</td>
<td>3.25</td>
<td>3.3</td>
<td>3.35</td>
<td>3.4</td>
<td>3.45</td>
<td>3.5</td>
<td>3.55</td>
<td>3.6</td>
<td>3.65</td>
<td>3.7</td>
<td>3.75</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>$D_{avg}$</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>95</td>
<td>70</td>
<td>65</td>
<td>95</td>
<td>60</td>
<td>80</td>
<td>70</td>
<td>95</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>55</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>35</td>
<td>35</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>50</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>n/a</td>
<td>5</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>35</td>
<td>35</td>
<td>45</td>
<td>55</td>
<td>55</td>
<td>50</td>
<td>65</td>
<td>85</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>$Tolerance index \Delta_{\gamma = 3}$ (%) (n/a indicates $\Delta_{\gamma = 3} &gt; 100$)</td>
<td>3.8</td>
<td>3.85</td>
<td>3.9</td>
<td>3.95</td>
<td>4</td>
<td>4.05</td>
<td>4.1</td>
<td>4.15</td>
<td>4.2</td>
<td>4.25</td>
<td>4.3</td>
<td>4.35</td>
<td>4.4</td>
<td>4.45</td>
<td>4.5</td>
</tr>
</tbody>
</table>

**Table IV.** Details of tolerance matrix with $n = 1,000$, $3 < \gamma \leq 4.5$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>3.05</th>
<th>3.1</th>
<th>3.15</th>
<th>3.2</th>
<th>3.25</th>
<th>3.3</th>
<th>3.35</th>
<th>3.4</th>
<th>3.45</th>
<th>3.5</th>
<th>3.55</th>
<th>3.6</th>
<th>3.65</th>
<th>3.7</th>
<th>3.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>30</td>
<td>30</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>40</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>80</td>
<td>85</td>
<td>95</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>30</td>
<td>30</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>$Tolerance index \Delta_{\gamma = 3}$ (%) (n/a indicates $\Delta_{\gamma = 3} &gt; 100$)</td>
<td>70</td>
<td>95</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

**Table AV.** Details of tolerance matrix with $n = 2,000$, $3 < \gamma \leq 4.5$

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>3.05</th>
<th>3.1</th>
<th>3.15</th>
<th>3.2</th>
<th>3.25</th>
<th>3.3</th>
<th>3.35</th>
<th>3.4</th>
<th>3.45</th>
<th>3.5</th>
<th>3.55</th>
<th>3.6</th>
<th>3.65</th>
<th>3.7</th>
<th>3.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{avg}$</td>
<td>30</td>
<td>30</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>$C_{avg}$</td>
<td>70</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>80</td>
<td>80</td>
<td>100</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$L_{avg}$</td>
<td>90</td>
<td>95</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$d_{avg}$</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>$k_{avg}$</td>
<td>30</td>
<td>30</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>$Tolerance index \Delta_{\gamma = 3}$ (%) (n/a indicates $\Delta_{\gamma = 3} &gt; 100$)</td>
<td>70</td>
<td>95</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

**Table AVI.** Details of tolerance matrix with $n = 4,000$, $3 < \gamma \leq 4.5$
About the authors

Kwok Tai Chui received the BEng Degree in Electronic and Communication Engineering – Business Intelligence Minor, with first-class honor and PhD Degree from City University of Hong Kong in 2013 and 2018, respectively. He was the recipient of 2nd Prize Award (Postgraduate Category) of 2014 IEEE Region 10 Student Paper Contest. Also, he received Best Paper Award in IEEE The International Conference on Consumer Electronics-China, in both 2014 and 2015. He has more than 20 research publications in referred international journals and conferences. His research interests include complex network, wireless communication, pattern recognition, healthcare, machine learning algorithms and optimization. He has served as various Editorial Positions and Guest Editors in referred international journals and conferences. Kwok Tai Chui is the corresponding author and can be contacted at: ktchui3-c@my.cityu.edu.hk

Dr Chien-wen Shen received his PhD Degree from Northwestern University in 2002. He is currently Associate Professor at Department of Business Administration, National Central University. His research interests are innovation technology, supply chain management, decision science, machine learning and social network.
More descriptive norms, fewer diversions

Boosting Chinese researcher performance through social media

Wenzhi Zheng
Business School, Huaqiao University, Quanzhou, China
Yenchun Jim Wu
Graduate Institute of Global Business and Strategy, National Taiwan Normal University, Taipei, Taiwan, and
Yue Lv
Business School, Huaqiao University, Quanzhou, China

Abstract

Purpose – The purpose of this paper is to analyze the relationship between researchers’ social media (SM) behavior and their academic performance.

Design/methodology/approach – A sample of 362 researchers was recruited from the colleges of management of 52 Chinese universities. A factor analysis of eight indices retrieved from the 362 data items was conducted. A total of 24 Chinese researchers were interviewed and given a robust test.

Findings – The results indicate that Chinese general social media (GSM) is insufficient to support academic research and it is difficult for scholars to enhance the visibility of their academic performance using GSM platforms, which can actually induce addiction. University resources, management systems, and working environment affect how scholars apply SM.

Research limitations/implications – The authors examined the researchers’ SM behavior by giving them a questionnaire and interview; however, this approach proved inadequate. The academic performance of researchers is affected by numerous factors, but the authors only considered SM behavior.

Practical implications – It is suggested that universities apply academic social media (ASM) indicators to measure researchers’ contributions so that they self-regulate their SM usage attitudes. Also, universities should also promote ASM platforms.

Originality/value – This study analyzed scholars’ GSM usage and academic performance, and the moderating effect of university level on the relationship between need for competence and relatedness and need for autonomy. This comprehensive analysis contributes to the scholarly SM usage literature.

Keywords China, Social media, Self-determination theory, Academic performance, Bibliographies, Social norm

Paper type Research paper

Introduction

With the rapid development of social media (SM), increasingly more scholars are using academic social media (ASM) to enhance their academic performance. However, Chinese SM platforms differ from those in other countries (Hoffmann et al., 2015; Erdt et al., 2016; Yu et al., 2016; Jeng et al., 2017). First, its general social media (GSM) platforms are independent. China has developed its own GSM platforms, such as renren.com, WeChat, and micro-blogging sites as substitutes for platforms such as Facebook, Twitter, and MySpace (Li and Chen, 2014; Huang et al., 2015; Zheng et al., 2016). Second, many Chinese scholars are unfamiliar with ASMs such as ResearchGate, Academia.edu, and Mendeley because of language barriers and differences in reading habits. For example, China accounts for the lowest proportion of publications on ResearchGate (less than 4 percent) when considering...
Web of Science publications from 2013 (Thelwall and Kousha, 2015). Generally, GSM is intended to promote social and individual life goals, whereas ASM is for academic purposes. Yet although Chinese scholars may obtain relevant information and communicate through the science.com academic blog, xmuchong.com, and dxy.cn, more generally, most scholars use GSM to replace some functions of ASM, such as the WeChat group for academic communication (Hao et al., 2015; Li et al., 2015a; Pun, 2015).

It is questionable whether Chinese scholars’ GSM usage enhances their academic performance. It offers not only the possibility of promotion but also the shock that innovative technological developments cause to human society. SM is a “double-edged sword” for education, because although it can promote learning, it can also lead to addiction and poor time management (Lau, 2017). It is challenging to understand the extent to which SM has permeated universities and their faculty (Kosinski et al., 2015). As individuals, university faculty may find SM addictive and time-wasting. However, they should also consider external social norms from Chinese culture such as power distance, collectivism, risk aversion, and long-term orientation (Zhang and Hummert, 2001; Gelfand and Harrington, 2015). Injunctive or descriptive social norms may influence researchers’ motivations and behavior (Li et al., 2015b).

The present study explored these phenomenon the basis of self-determination theory (SDT), focusing on the following three issues: first, whether addictive behavior or other adverse consequences arise from Chinese scholars’ GSM use, after their core needs have been met; second, how universities should respond to the addictive behavior, given different individual resources; and third, how Chinese GSM functions, and whether researchers can improve their academic performance through GSM. We analyzed the relationship between researcher behavior and academic performance under the influence of external environmental factors.

Literature review and hypothesis

Literature review and research model

Previous studies have indicated that SM platforms have become a type of learning environment that can help individuals grow (Lin and Wang, 2013; Wang et al., 2013; Legaree, 2015). In the learning literature, new terms such as “social media learning” and “digital scholarship” have been introduced (Claros and Cobos, 2013; Pretto and Curró, 2017). According to SDT, people pursue the satisfaction of fundamental psychological needs for competence, autonomy, and relatedness while engaging in various activities (Ryan and Deci, 2000). On SM platforms, when students or educators enjoy independent communication, browsing, quick access to information, sharing resources, and collaboration, they have higher autonomous motivation and meta-cognition, perceived usefulness, playfulness, and ease of use, and they experience higher satisfaction of their psychological needs; this ultimately leads to improved education performance (Søreb et al., 2009; Koh et al., 2010; Legaree, 2015). Researchers have shown that students and educators with high perceived competence, autonomy, and relatedness achieve better learning outcomes. Therefore, schools should create working environments that enable SM platforms to be utilized effectively and enhance the performance of both educators and students.

Previous research has also shown that schools have different strategies and attitudes toward SM, and that teachers from different disciplines differ in their views on the function and relevant norms of SM (Jeng et al., 2015; Jeng et al., 2017). Thus, external social norms influence teachers’ SM attitude as well as their working efficiency and performance (Meishar-Tal and Pieterse, 2017; Wu et al., 2017). However, other studies have reported that SM has adverse effects on learning outcomes, because while SM can improve learning efficiency, it also promotes nonacademic behaviors (i.e. it acts as a diversion). If SM is used to fulfill social and nonacademic needs only, academic performance suffers (Ravizza et al., 2014).
According to SDT, two ways that SM influences academic performance. One is the influence of SM on learning efficiency, including the interesting and satisfaction from SM technology usage, as well as the external environment of schools and the social attitudes and norms around the new technology. These comprehensively influence the learning and work efficiency of the teachers. The other is that, as a new technology platform, SM can also promote individual initiative and self-discipline. Differing attitudes regarding the purpose of using SM can lead to different behaviors and, consequently, different effects on academic performance. Therefore, we proposed the research model depicted in Figure 1 as the basic framework of this study.

Research hypotheses

New technology has brought great convenience and satisfaction to humans. The internet and other information technologies offer motivational benefits such as freedom, ownership, communication, enjoyment, and accessibility. ASM offers university faculty a new option for balancing their work and personal lives (Jeng et al., 2015), by meeting their basic needs and conferring psychological well-being (Ryan and Deci, 2000). Satisfying the needs for autonomy, competence, and relatedness is necessary for promoting intrinsic motivation (Megwalu, 2015), which is positively related to beneficial outcomes.

Claros and Cobos (2013) introduced SM learning, a collaborative environment for composing multimedia-interactive learning objects. With the aid of this type of environment, which can be designed to provide tools such as SM platforms and methodological guidelines, researchers may be able to perceive themselves as being competent in using SM. This could support the development of their cognitive and meta-cognitive skills, specifically for researchers to be able to understand their learning process and construct new knowledge (Blaschke, 2014). SM could improve people’s attitudes toward science, as well as their ecological engagement (Wilson and Boldeman, 2012), environmental awareness, and creative thinking (Karahan and Roehrig, 2015). Moreover, novel technologies such as SM can overcome temporal spatial barriers, accelerating SM learning and thereby facilitating the realization of knowledge sharing and collaborative research. Thus, SM makes it feasible and convenient to conduct joint research (Fernández-Díaz et al., 2017). Furthermore, ASM platforms, such as Mendeley and ResearchGate, provide researchers with support such as social research networking and document and citation management, as well as research collaboration opportunities (Van Noorden, 2014; Jeng et al., 2015). SM makes learning ubiquitous, providing more opportunities, access, and learning tools (Veletsianos and Kimmons, 2012).

In China, the GSM platforms available to teachers are developing rapidly, although ASM platforms are not widely used (Vaughan and Gao, 2016). SM includes sites that enable personal knowledge management (e.g. 360doc.com as well as blog, video, and image management sites), interpersonal communication (e.g. Tencent QQ messaging, online groups, and real-time chats), and learning goal management and learning assessments such as reflections, other people’s comments, and critiques (e.g. the many WeChat groups for academics) (Hao et al., 2015; Li et al., 2015a). SM makes academic exchange, interpersonal relationship management, and information searching convenient for Chinese scholars, thereby meeting their autonomy, competence, and relatedness needs. This promotes their
intrinsic motivation, which is beneficial for improving their academic performance. Therefore, we proposed the following hypothesis:

\[ H1. \text{SM that meets researchers' basic requirements is positively related to their academic performance.} \]

Although SM platforms improve researchers’ learning efficiency by meeting their needs, the process is affected by organizational climate (Thelwall and Kousha, 2015). Organizational climate and social norms can affect individual behaviors, especially for personal descriptive norms, which refer to the perceived popularity of a behavior among an individual’s significant others, and is more easily internalized into an individual’s intrinsic motivation (Kim et al., 2015). Research work is relatively free and unrestrained, and the injunctive norms are limited to scholars (Fogelberg, 2016). Many Chinese universities have endeavored to create personal learning environments and assist their faculty in becoming more self-directed learners and researchers in the field of SM (Greenhow and Lewin, 2015; Hong et al., 2016). Good universities, such as the 36 universities of Project 985, are attempting to build a robust academic SM environment, including the construction of user-friendly ASM platforms to stimulate researcher creativity through academic discussion, seminars, and information sharing (Xu et al., 2015). This also helps scholars to create a strong communication environment or research circle. Second, a leading university can guide its faculty to use SM correctly by implementing a set of performance assessment indicators system. Moreover, these universities often gather high-profile scholars who can set a learning example for their colleagues.

To gain recognition and attention in their circle, researchers might improve their mobile learning and SM-learning abilities (Søreb et al., 2009; Kreijns et al., 2014). SM enables scholars to seek academic information from new friends in various research fields (Dermentzi et al., 2016), thus enhancing their research abilities as they fulfill such social norms as peer evaluation and social expectation. We believe that because reputable universities (i.e. those listed in Projects 985 and 211) offer their faculty access to high-quality academic resources, their faculty enjoy a higher reputation in academic circles, which could promote their academic standards and self-requirements. By contrast, researchers from universities not listed in either project (non-Project 211 universities) have insufficient resources and therefore exhibit relatively poorer academic performance. Hence, we proposed the following hypothesis:

\[ H2. \text{University level (UL) moderates the influence of researchers’ basic needs on their academic performance; Chinese researchers at Project 211 universities have more research resources and examples and exhibit higher academic performance than do those at non-Project 211 universities.} \]

In addition to environmental factors, individual behavior is a central factor in academic SM performance. Cognitive evaluation theory posits that people engaging in work behavior feel interested when external factors meet their basic needs (Lee et al., 2015). The satisfaction of basic needs can increase researchers’ intrinsic motivation, which in turn promotes spontaneous SM behavior (Kreijns et al., 2014).

Two types of SM behavior exist: academic SM behavior (ASMB) and nonacademic SM behavior (NASMB) (Ravizza et al., 2014). ASMB pertains to digital learning platforms that can improve academic performance (Kuo et al., 2017). Many teachers use digital learning environments as a convenient communication tool (Pacurar and Abbas, 2015) and an effective means of obtaining information (Jeng et al., 2015; Dermentzi et al., 2016; Botting et al., 2017; Wang et al., 2017). However, NASMB impedes academic performance through two mechanisms. The first is addiction. In the process of SM usage, researchers can perceive fun, ease of use, and satisfaction because meeting basic needs leads to intrinsic motivation, which can promote addiction. Although it seems reasonable to assume that researchers are rational,
the theory of rational addiction suggests that people consider the future influence of consuming items they are addicted to. Many scholars have realized that SM represents a future trend and have therefore endeavored to regulate themselves as they adapt to it. Moreover, under the influence of social norms, researchers practice self-discipline by exerting self-control and judgment and by controlling their intentions, emotions, and behavior (Larose et al., 2003). However, this type of control can be arduous; because of abstinence reactions, addicts experience strong negative emotions and even physiological discomfort when attempting to terminate or reduce the consumption of items they are addicted to.

The other hindrance of NASMB is multitasking (Van Der Schuur et al., 2015; Wu, 2017), which may take three forms: dual-tasking, rapid attention switching, and continuous partial attention (Lau, 2017). SM multitasking has a negative effect on academic performance, including academic outcomes, study-related behaviors and attitudes, and perceived academic learning (Van Der Schuur et al., 2015). There are two explanations for these negative effects, which are the time displacement hypothesis and the limited information processing capacity hypothesis. For researchers, multitasking leads to reduced efficiency and increased anxiety. Low work efficiency may impair their competence and relatedness. They may experience emotional distortion and a dependence on SM, which aggravates their anxiety and ultimately affects their academic performance.

Chinese researchers also have access to network communication, user-generated content (UGC), and other SM intention and behavior. Academic vs nonacademic UGC behavior differ in their influence on academic performance. That of academic UGC is positive, because it can meet researchers’ competence and relatedness needs (Aladwani and Almarzouq, 2016). However, these innate needs are difficult to satisfy when they conflict with external social norms. In the face of students and the public, many under achieving researchers with weaker academic backgrounds adopt other approaches to realize their self-worth in order to achieve a sense of competence and relatedness (Costa, 2015).

In addition, being raised in a collectivist culture, Chinese people adhere to the norm of moderation and value their reputations. Consequently, Chinese researchers might use UGC to gain a stronger sense of competence. Because their work is widely applicable, many university researchers divert some of their energy to social services, teaching, and class management and generate research-unrelated content for these social circles and communities (Toubia and Stephen, 2013; Zhang et al., 2017). Consequently, researchers might create UGC to gain a stronger sense of competence. Therefore, in addition to using SM to accomplish limited teaching goals (e.g. classroom discussions), university researchers show case and communicate their ideas by creating research-unrelated content, forum posts, communication, and seeking autonomy or a sense of existence. However, the time spent on this multitude of tasks might jeopardize their academic performance (Adebiyi et al., 2015). Therefore, we proposed the following hypotheses:

**H3.** Different SM behaviors mediate the relationship between researchers’ basic requirements and their academic performance.

**H3a.** The link between researchers’ basic requirements and their academic performance is mediated by ASMB.

**H3b.** The link between researchers’ basic requirements and their academic performance is mediated by NASMB.

**Method**

*Participants*

In this study, data were collected from researchers teaching at colleges of management in Chinese universities. Relative to the practice of business, management as a field of study is
most closely connected with society; researchers in this field therefore have more options and liberty in their SM behavior. To examine the academic performance of researchers, 52 colleges of management were randomly selected from three categories of Chinese universities and then screened on the basis of whether they had blogs and Weibo accounts. Ultimately, 362 effective samples were collected. Specifically, 163 of the researchers were teaching at Project 985 research-focused universities (45 percent), 123 were from Project 211 but non-Project 985 teaching- and research-focused universities (34 percent), and 76 were from universities not listed in either project (21 percent). Considering the influence of gender on academic research (Abramo et al., 2015), gender was also included as a variable: 294 of the participants were male (81.2 percent) and 68 were female (18.8 percent). Similar to the gender distribution in other parts of the world (Cimenler et al., 2014), Chinese researchers are predominantly male. These statistics match the actual situation in China. Therefore, we can consider them representative for researching the SM behavior of Chinese management scholars.

**Measurement and data sources**

To study the SM behavior of researchers, we examined blogs, Weibo, and WeChat, which are influential SM platforms among Chinese scholars. Mobile app usage by Chinese netizens in 2016 indicated that most of them (95.1 percent) are users of mobile instant messaging (primarily WeChat and QQ), totaling 695 million people. WeChat circle (85.8 percent), QQ space (67.8 percent), and Weibo (37.1 percent) are the top three SM apps in China (CNNIC, 2016). We employed content analysis to code the SM behavior of these researchers.

To measure individuals’ basic needs or SM behavioral intentions, the number of followers was used as a proxy for need for competence and relatedness (NCR) (Zhu and Chen, 2015; Lin et al., 2016), and the update frequency for need for autonomy (NA). The variables used to measure ASMB included the professionalism of SM content and the professionalism of UGC. NASMB was measured by behaviors such as forum posting and generating blocked content. Regarding institutional environments, Projects 985 and 211 are two widely recognized classifications in China. Specifically, Project 985 comprises the top research-focused universities, the best among the Project 211 universities. Concurrently, Project 211 universities that are not in Project 985 are research and teaching-focused universities. Non-Project 211 universities are generally teaching-focused universities offering undergraduate degrees.

When the content posted on Weibo was insufficient, we used other SM platforms such as WeChat to collect the ten latest images posted by a researcher and analyzed the content of the images. Specifically, content professionalism was measured by examining the content of updates, images posted, and other UGC. In other words, we retrieved and interpreted the ten most recent self-generated or reposted items for each researcher in the sample.

Additionally, we used the following criteria for coding the content: more than half of the contents are related to the field of management = 5, 3 to 5 contents are related to management = 4, 1 to 3 are related to management = 3, 1 is related to management = 2, and 0 is related to management = 1. Similar criteria were adopted to code the other variables. During the coding process, two researchers conducted the coding separately and then their results were compared. When the results were identical, they were adopted; otherwise, the coding was further discussed and any results for which a consensus was not reached were omitted. Of the 384 items in the data set, we reached a consensus on 362 of them (matching success rate, 94.27 percent).

To avoid homologous variance, we adopted the h-index, a method commonly employed for measuring the academic contributions of scholars (Hirsch, 2005). The h-index is a traditional measure of author-level bibliometric that attempts to measure both the productivity and citation impact of a scholar’s publications. This bibliometric measure is
based on a scholar’s most cited papers and the number of citations (Hassan et al., 2017). Using the China National Knowledge Infrastructure and Google Scholar for searches, we calculated and entered the h-index values for the participants.

**Data analysis and hypothesis testing**

After standardizing the raw data, we conducted a factor analysis of eight indices retrieved from the 362 data items. In the results, the professionalism of SM content and UGC were clustered as ASMB, and forum posting and generating blocked content were clustered as SMB. This produced a six-factor model, and a confirmatory factor analysis result of the model showed that the six-factor model, \( \chi^2/df = 4.697 \), goodness-of-fit index (GFI) = 0.879, comparative fit index (CFI) = 0.888, Tucker-Lewis index (TLI) = 0.886, incremental fit index (IFI) = 0.887, and root mean square error of approximation (RMSEA) = 0.074, was a closer fit to the data than the eight-factor model, \( \chi^2/df = 6.604 \), GFI = 0.779, CFI = 0.828, TLI = 0.816, IFI = 0.827, and RMSEA = 0.104. Therefore, the six-factor model was adopted.

We subsequently conducted a correlation analysis of the six factors (Table I). Significant correlations were observed between the h-index and other predictive variables, except for GSM. Therefore, we explored the relationships between researchers’ SM behavior and their academic performance by analyzing their basic needs and motivation.

**Mediating effect**

After controlling for gender and UL, we referred to the analytical procedures of the 2-1-1 mediation model proposed by Zhang et al. (2009) to conduct a mediating effects analysis for the measured variables (see Table II). The following aspects were examined: first, the direct effect of arguments on the dependent variables (confirmed by M5 and M6 in Table II); second, the direct effect of arguments on the intervening variables (M7 and M8 in Table II show that the h-index correlated significantly and negatively with NASMB, \( r = -0.108, p < 0.01 \), but not with ASMB); third, the direct effect of the intervening variables on the dependent variables (M1-M4 in Table II) show that ASM correlated significantly and positively with NCR, \( r = 0.227, p < 0.001 \) and NA (\( r = 0.120, p < 0.01 \)). ASM correlated significantly and positively with NA (\( r = 0.122, p < 0.01 \)) but not with NCR. Fourth, the interaction effects of arguments and the intervening variables on the independent variables were tested. M4 and M7 were not significant; there was thus no mediating effect of ASMB, meaning that ASMB does not mediate the relationship between researcher needs and the h-index; thus, \( H3a \) was rejected.

However, M10 indicated that the interaction of NA and NASMB correlated significantly with the h-index. The explanatory power of NA on the h-index increased, suggesting that NASMB (negative) exerted partial mediating effects on NA and the h-index. The explanatory power of NA on the h-index increased from M6 (\( r = 0.117, \)

<table>
<thead>
<tr>
<th>Pearson</th>
<th>Gender</th>
<th>UL</th>
<th>NCR</th>
<th>NA</th>
<th>ASMB</th>
<th>NASMB</th>
<th>h-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1</td>
<td></td>
<td>-0.122*</td>
<td>-0.139**</td>
<td>-0.101</td>
<td>-0.191**</td>
<td>0.051</td>
</tr>
<tr>
<td>University level (UL)</td>
<td>1</td>
<td></td>
<td>0.549***</td>
<td>0.017</td>
<td>0.015</td>
<td>-0.275**</td>
<td>0.301**</td>
</tr>
<tr>
<td>Need for competence and relatedness (NCR)</td>
<td>1</td>
<td></td>
<td>0.283**</td>
<td>0.249**</td>
<td>-0.005</td>
<td>0.213**</td>
<td></td>
</tr>
<tr>
<td>Need for autonomy (NA)</td>
<td>1</td>
<td></td>
<td>0.138**</td>
<td>0.106*</td>
<td>0.133*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASMB</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>NASMB</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.116*</td>
<td></td>
</tr>
<tr>
<td>h-index</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table I.** Correlation coefficients for the variables

**Notes:** *p < 0.05; **p < 0.01
<table>
<thead>
<tr>
<th>Variables</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
<th>M10</th>
<th>M11</th>
<th>M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCR</td>
<td>0.227***</td>
<td>0.002</td>
<td>0.193***</td>
<td>0.117**</td>
<td>0.194***</td>
<td>0.131**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>0.120**</td>
<td>0.122**</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASMB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.108**</td>
<td>-0.108**</td>
<td>-0.121**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASMB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCR×UL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.258***</td>
</tr>
<tr>
<td>NA×UL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.068</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.187</td>
<td>0.150</td>
<td>0.115</td>
<td>–</td>
<td>0.164</td>
<td>0.142</td>
<td>–</td>
<td>0.139</td>
<td>0.176</td>
<td>0.156</td>
<td>0.179</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is h-index. **p < 0.01; ***p < 0.001
The size of the mediating effects was $-0.158$ ($-0.121 \times 0.131$) and the proportion of the mediating effects was 54.67 percent ($0.158/(0.158 + 0.131)$), meaning that the researchers’ autonomous motivation can positively predict their academic performance, and in absence of NASMB, the effect is improved; therefore, $H3b$ was supported.

**Moderating effect**

We further analyzed the moderating effect of UL by setting it as a moderator to examine its effect on the relationships of NCR and NA with the h-index. After controlling for gender, we found that UL exhibited moderating effects, described as follows (see Table II).

In M5 and M11 of Table II, although NCR had a direct positive predictive effect on the h-index, UL moderated the relationship between these two variables. As shown in Figure 2, when UL was high (High Univ), NCR (F-few, F-more) had a larger effect on the h-index. By contrast, in M12, UL did not moderate the influence of NA on the h-index. In conclusion, UL significantly moderated the influence of the number of followers on the h-index but did not moderate the influence of NA on the h-index. Therefore, $H2$ was partially supported.

**Robust test**

To verify the research reliability and assess how to improve academic performance in terms of SM use, we conducted an in-depth interview of 24 Chinese teachers (male = 15, female = 9; $E_{age} = 38.2$) from three universities and coded the results. The Nvivo 8.0 qualitative analysis program was employed to code the data. Two researchers coded the data separately and compared the results. If the results were identical, we proceeded with the study. When the results differed, a discussion was held. If a consensus was not reached after this discussion, these results were deleted.

The coding process was mainly based on three components of the SDT, academic performance (including academic achievement and academic reputation), the scholars’ SMBs, and other related factors. First, a total of 405 tertiary nodes were created to code the data using the numbers N101-N124. Second, in addition to open coding, axial coding was used to aggregate the open-coding results, thereby yielding accurate explanations and associating the data with established concepts, as described in Table III.

According to Table III, SM mainly meets scholars’ needs for autonomy and relatedness. Relatedness requirements include seeking mutual help and a favorable academic atmosphere. In terms of communication content, the interviewees perceived openness, pertinence, and recent research. Therefore, they reported favorable atmosphere and belonging when commenting on academic communication. Autonomy requirements include SM entertainment, which can provide relaxation as well as self-regulation and control for scholars. Second, regarding SMB, the interviewees stated that China lacks ASM and that the functions of

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Moderating effect of UL on the relationship between NCR and the h-index.
SM-based academic evaluation and prestige signaling are weak. Third, in terms of specific behaviors, ASMBs mainly result from academic communication between scholars. NASMB information was difficult to capture during the interview process. However, the relevant information can be acquired from the frequencies of SM use and nonacademic communication, such as chatting online by using WeChat or QQ. Finally, the role of academic atmosphere, including academic assistance, communication topic, and the prohibition of SM use in the classroom, is strongly related to the semantics of SM behavior and academic performance. In summary, because the dimensions and items of the four variables were correlated, the in-depth interview supported the aforementioned analysis results.

**Discussion and conclusions**

*Results and discussion*

A hypothetical model was tested to examine whether researchers’ satisfying their basic needs through SM affects their academic performance. First, new technologies, such as SM, that supported self-determination enhanced researchers’ performance, because the technologies induced satisfaction and interest, enabled autonomous choices and decisions, highlighted and demonstrated ability, and led to acceptance and approval from their peers. These results indicated that the main function of SM is to satisfy researchers’ basic needs, such as competence (by enabling researchers to search for the latest information) and relatedness (by facilitating the building of interpersonal connections) (Gerow et al., 2013; James et al., 2015; Shen et al., 2017). To establish relations with peers and students, Chinese researchers must constantly update their knowledge and provide timely updates on cutting-edge academic information in their fields. During the process of coding, we discovered that the followers of many high-achieving scholars were insiders of academic circles. Moreover, satisfying the NA has a positive influence on creative performance (Liu et al., 2011). This finding agrees with the results of previous research (Kanfer et al., 2017).

Second, ASMB did not mediate the relationship between researchers’ basic needs and academic performance. Because Chinese researchers have no experience with (or access to) ASMs such as Mendeley, it is difficult for them to improve their academic performance through SM. Few Chinese scholars use ResearchGate, Academia.edu, and Mendeley, making it difficult for them to obtain academic information or share knowledge by using SM. The GSM (WeChat and Weibo) data in the present study indicate that Chinese scholars have been unable to effectively collaborate over SM. The interview data were also used to retest the conclusion that GSM cannot meet researchers’ competence needs (only five nodes; Table III). This is consistent with the conclusions of previous studies of academic exchange groups (Hao et al., 2015; Li et al., 2015a). GSM can partially meet scholars’ basic needs, but GSM cannot fully replace the functionality of ASM or improve academic performance. Moreover, the relationship between the altmetric and bibliometric at the author level was poor (Ortega, 2015; Yu et al., 2016; Hassan et al., 2017; Lali et al., 2017).

Concurrently, NASMB negatively mediated the relationship between NA and academic performance. The satisfaction of basic requirements could have had a greater effect on researchers’ performance, but only with fewer NASMBs. Generally, researchers’ performance can be enhanced by academic behavior on ASM (Botting et al., 2017). On GSM platforms, the

<table>
<thead>
<tr>
<th>SDT Dimensions</th>
<th>Behaviors Dimensions</th>
<th>Atmosphere Dimensions</th>
<th>Performance Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatedness</td>
<td>Frequency</td>
<td>Academic assistance</td>
<td>Academic prestige</td>
</tr>
<tr>
<td>Competency</td>
<td>A-communication</td>
<td>Communication topic</td>
<td>Academic performance</td>
</tr>
<tr>
<td>Autonomy</td>
<td>NA-communication</td>
<td>Prohibition</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Items</th>
<th></th>
<th>Items</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatedness</td>
<td>37</td>
<td>Frequency</td>
<td>75</td>
<td>Academic assistance</td>
</tr>
<tr>
<td>Competency</td>
<td>5</td>
<td>A-communication</td>
<td>19</td>
<td>Communication topic</td>
</tr>
<tr>
<td>Autonomy</td>
<td>55</td>
<td>NA-communication</td>
<td>15</td>
<td>Prohibition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Items</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>Academic prestige</td>
</tr>
<tr>
<td>24</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Academic performance</td>
</tr>
<tr>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

**Table III.** Axial-coding items
goal of researchers is not professional but personal; behaviors motivated by personal life and entertainment-seeking are not conducive to academic performance (Nicholas et al., 2014). Because NASMB can lead to addiction and multitasking problems (Van Der Schuur et al., 2015; Lau, 2017), new technologies may weaken scholars’ creativity and academic though Chinese researchers mainly use SM as a virtual space to play, and they waste a considerable amount of time doing so. Gaining peer approval for academic performance remains challenging. This conclusion is consistent with a study of NEHU Scholar SM usage in India (Chakraborty, 2012), and the investigation in Nature (Van Noorden, 2014).

Finally, UL moderated the relationship between NCR and academic performance, but it did not moderate the relationship between NA and academic performance. Studying SM technology adoption behavior through social influence theory is innovative. Previous studies have shown that subjective norms influenced by external forces could affect behavior and performance by changing individual attitudes (Venkatesh et al., 2003; Sharma et al., 2016). Facing the impact of SM information and norms, scholars may adopt strategies such as compliance, internalization, and identity (Cheung and Lee, 2010). Leading universities confer an information advantage or supportive external environment for their faculty, which can meet their basic needs and promote their academic activity by enhancing their self-regulation of SM usage. Moreover, the results of this study show that scholars with high competence and relatedness needs were prone to conform to descriptive norms (Venkatesh et al., 2003; Cheng et al., 2011; Beullens and Vandenbosch, 2016). UL and other aspects of institutional atmosphere can enhance academic performance when SM can meet competence and relatedness needs, but UL does not moderate the influence of NA on academic performance. In other words, UL can improve access to academic information and other facilities by promoting SM usage and encouraging scholars to use SM in an appropriate manner, establishing productive norms that can enhance their academic performance. On the contrary, if SM is used only to meet NA in the absence of regulatory or social norms, it cannot effectively improve academic performance.

Implications for research and practice
The purpose and value of individuals’ SM usage as well as information collection methods have become a popular topic in recent years (Li et al., 2015b). In this context, the present study analyzed scholars’ GSM usage and academic performance in addition to the moderating effect of UL. This comprehensive analysis contributes to the scholarly SM usage literature. Past researchers have focused on the influence of cognitive and other intrinsic factors on individual behavior (Kanfer et al., 2017). Drawing on SDT, this study examined the combined influences of external social norms on individual SM motivation and behavior. Researchers’ SM behavior can be understood as a means of pleasing their own or others’ requirements. The motivation underlying this type of behavior maybe self-improvement, maintaining existing academic circles, raising status among peers, and image management (Dermentzi et al., 2016). Therefore, this study considered researchers’ individual variables such as SM use motivation and behavior, and it further revealed the relationship between the altmetric (such as researchers’ views and their SM followings) and bibliometric (h-index) at the author level.

Practically, The SDT and social influence perspectives underlying this study suggest that new SM technologies have the potential to meet certain needs of scholars and thus improve their work efficiency and academic performance (Ryan and Deci, 2000; Gruzd et al., 2012; Kim et al., 2015). However, our results indicate that Chinese GSM is insufficient to support academic research; it is difficult for scholars to improve their academic performance using GSM platforms, which can actually induce addiction. University resources, management systems, and working environment affect how scholars apply SM. Those universities with abundant resources and a supportive atmosphere could meet scholars’ competence and relatedness needs, or discourage them from indulging in diversions, thereby enhancing their
performance. By contrast, in the absence of standards or norms to guide behavior, the freedom to partake in SM diversions is not conducive to academic growth.

This study has some limitations. First, we tested SM researchers’ motivation through interviews and questionnaire survey; however, this approach was insufficient to test individual internet motivations. Second, the academic performance of researchers is affected by numerous factors, but we only considered SM behavior. Moreover, our conclusions are based on a small sample of researchers from Chinese universities; further verification of our findings is warranted. However, we found that the UL and SM environments are factors that considerably influence academic performance. Therefore, future studies should further examine the influence of university environments, SM academic exchange spaces, SM integration, and types of SM usage.

Conclusions
Chinese society expects universities to lead in science and technology and social development through academic research (Chen et al., 2013; Hao et al., 2015). SM behavior is an aggregate of multiple activities promoted by SM usage intention, which is induced by researchers’ comprehensive assessment of social influences, their individual needs, and the SM technologies they use (Gruzd et al., 2012). If a university wants to promote its faculty’s academic performance, they should encourage researchers to use GSM and other new technologies for academic purposes while restricting their autonomy in SM usage (Al-Aufi and Fulton, 2015). To create an atmosphere where SM supports academic research, we suggest that universities apply ASM indicators to measure researchers’ contributions, so that they self-regulate their SM usage attitudes. Universities should promote ASM platforms, such as official WeChat accounts (Zhu, 2016), to assist researchers in screening the information, which can then meet their competence needs. Therefore, many business opportunities are available for the third-party application suppliers such as Baidu and CNKI to cooperate with academic libraries. Such suppliers can help researchers to build individual academic accounts, collect academic information, share their latest research, track peer progress, establish and maintain connections, and join academic circles.

References


Corresponding author
Yenchun Jim Wu can be contacted at: wuyenchun@gmail.com
Bibliometric networks and analytics on gerontology research

Chien-wen Shen, Duong Tuan Nguyen and Po-Yu Hsu
Department of Business Administration, National Central University, Taoyuan City, Taiwan

Abstract

Purpose – The purpose of this paper is to bibliometrically analyze the gerontology-related research articles for a comprehensive understanding of the gerontology literature.

Design/methodology/approach – This study employed the approach of visual analytics on 32 journals with a total of 99,204 articles published after 2000 to identify the main subfields, keywords, and growth trend. The investigated journals are either open access online or listed in the Social Sciences Citation Index. In addition, the 200 most frequently cited papers were analyzed through bibliographic coupling, co-word, and co-citation analysis.

Findings – The selected most cited papers were mostly published before 2007, and psychiatry and psychology were the top research subfields. Dementia, older adult, and Alzheimer’s disease were the three most frequently occurring keywords, both in Author Keywords and KeyWords Plus. While coupling analysis yielded 12 research groups, co-word analysis classified the most frequently used 20 Author Keywords into two categories. Four research clusters were identified by the co-citation analysis.

Originality/value – This research provides a comprehensive view of the gerontology research as well as an understanding of the subfields and their interrelations. It also provides government departments with directions for formulating and executing policies affecting older people not only in setting academic and professional priorities but also in understanding the key topics related to older people.

Keywords Elderly, Network analysis, Co-citation analysis, Bibliometric analysis, Co-word analysis, Gerontology

1. Introduction

The aging population problem has become increasingly prominent because it has direct impacts on the size and quality of the workforce as well as the financial integrity of health care and pension systems (Bloom et al., 2011). Especially when the older people use more health services than the other segments of the population, the need for gerontology research has become more evident in relation to the economic aspects of population aging (Zubair and Norris, 2015). Gerontology is the scientific discipline that investigates the process of aging from middle age to old age, including the study of the physical, mental, and social changes among older people, the social impact of population aging, and the application of this knowledge to policies and programs (Hooyman and Kiyak, 2008). This field has attracted many scholars to investigate its different aspects. In the early 1990s, Vasil and Wass (1993) investigated the image of the older people in the media. They reported that the description of older people in the media was generally inadequate and did not accurately reflect older people in the USA. Levers et al. (2006) reviewed the theoretical and empirical literature identifying the causes of vulnerability among older people. Topo (2009) reviewed the literature on strategies for using science and technology to support people with dementia and their caregivers. She determined that most studies have focused on caregivers’ needs but few have explored the benefits of using technology to support people with dementia. Bloch et al. (2011) conducted a meta-analysis on psychotropic drugs to elucidate the relationship between falling among older people and psychotropic drugs.

Although several studies have applied systematic analysis to the gerontology literature, most of these metadata studies on aging among older people have investigated a specific area: dementia and Alzheimer disease (Ansari et al., 2006; Asghar et al., 2017; Baldwin et al., 2003;...
Guido et al., 2015), cardiovascular disease (Ugolini et al., 2013), long-term care (Phay, 2011), psychology and physical activity (Müller et al., 2016), and successful aging (Martinson and Berridge, 2014). Ang and Kwan (2017) sorted studies on geriatrics and gerontology published on Journal Citation Reports between 2007 and 2014 by impact factor to create a dataset. Their findings contributed to the understanding and provided research directions for researchers in this field. However, the data source was limited, and the search was limited to a relatively short period. The relationships between the studies and the authors have been not evidently indicated.

Clearly, the gerontology has been drawing a demanding research field that a lot of scholars entered and invested on exploring this field with a wide range of topics. As well as specific investigations, reviews of the gerontology literature may provide researchers with greater insight into this field and clarify potential research directions. Nonetheless, few literature reviews on gerontology have been conducted from a holistic perspective. Therefore, the objective of this study is to undertake the approach of bibliometric networks and analytics on the research articles in the field of gerontology in order to provide a comprehensive understanding toward this field. This study employed the approach of visual analytics to analyze the status of geriatric research as well as the approach of network analysis to explore the relationship between the authors and the distribution of research areas. The investigated journals are either open access online or listed in the Social Sciences Citation Index. Under the network analysis, the 200 most frequently cited papers were examined through bibliographic coupling, co-word, and co-citation analysis. From the findings of this study, the researchers could identify the recent situation of the gerontology research and focus on the critical and challenging problems. The institutions could also refer the results of this study to formulate and implement the policies and practices on the research activities in the gerontology field.

2. Literature review
The literature in the gerontology field generally addresses on specific areas such as the psychology of aging, diseases that affect older people, and health care services and assistive technology for older people. Research on the psychology of aging has increased appreciably since the 1990s. Sáiz et al. (1998) used the Psycit database to conduct bibliometric analysis on the psychology of aging by using data from research conducted from 1991 to 1995. The results revealed the critical role of the psychology of aging in contemporary psychology; the USA was the country with the most publications on this topic, followed by Germany, Australia, the UK, and Sweden. The most commonly used keywords were aged, very old, aging, elder, adulthood, major depression, age differences, drug therapy, and Alzheimer’s disease. Additionally, Baldwin et al. (2003) focused on the literature on ethics and dementia during the period 1980-2000. Their findings indicated that four topics dominated the literature during this period: professional care, end-of-life issues, decision making, and treatment. Additionally, they found that two methods were excessively applied to investigate these topics, namely, surveys and studies soliciting responses to predefined issues; and relatively few in-depth, open-ended qualitative studies were conducted. Besides, the lack of research on the ethical challenges facing family members and professionals was discussed.

Ansari et al. (2006) and Sorensen (2009) have conducted bibliometric analyses of the literature on Alzheimer disease. The results indicated that the USA led among 50 countries for publications on Alzheimer disease with 38.05 percent of the publications; the UK and the Netherlands took the second and third places with 19.95 and 6.06 percent of publications, respectively. Furthermore, author network analysis was applied to identify the relationships between authors (Sorensen, 2009). Similarly, Asplund et al. (2012) compared the developments in stroke research between countries over time. The results showed that
the USA dominated with 28.7 percent of the published articles and 36.2 percent of citations. In terms of population size and gross domestic product, several small European countries, Israel and Taiwan ranked highest. A negative association between the burden of stroke (disability-adjusted life-years lost) and the number of articles per population was found. The study highlighted the importance of multinational collaboration in this research field, particularly in countries with many people with stroke.

Ugolini et al. (2013) focused on cerebrovascular and cardiovascular disease (CCD) rehabilitation; the data used in their study (covering 1967-2008) were downloaded from the PubMed database. The findings indicated that the number of publications on CCD rehabilitation increased 8.6-fold over the 40 years of data, whereas those on all-disease rehabilitation increased 7.8-fold over the same period and those on cerebrovascular diseases increased 5-fold over the last 15 years. However, article quality (i.e. impact factor) in this field decreased over this period. 44.4 percent came from the European Union and 30.3 percent from the USA. The commonest keyword was “stroke” (Ugolini et al., 2013). Guido et al. (2015) found that more knowledge on frontotemporal dementia (FTD) had become available since the mid-1990s. The most mentioned topics included advanced neuroimaging techniques, genotype-phenotype heterogeneity, and animal models. These findings indicated the attraction of FTD, and more generally gerontology, to researchers.

Phay (2011) reported that the five most prominent topics in the field of long-term health care services are gerontology, psychiatry, healthcare science, nursing, and epidemiology. Furthermore, the role of physical activity and aging as key research directions over the past several decades was affirmed (Müller et al., 2016). Asghar et al. (2017) investigated assistive technology that supports patients with dementia. The aforementioned studies have reported that the USA is the major contributor to this research field.

The summary of literature review is shown in Table I, which highlights the information of authors, data, and method. As we can see that all of studies are based on a specific field of gerontology such as psychology of aging, dementia, Alzheimer’s disease, and long-term health care. Besides, the approach of bibliometric networks has not been addressed yet in the gerontology research. Hence, this research aims to bridge some of the existing gaps.

3. Methods
3.1 Data set description
The data used in this study were obtained from papers published in the relevant gerontology journals. Most of these journals are listed in the Social Sciences Citation Index. The data set contains 32 journals with a total of 99,204 articles published in the period 2000-2016. Additionally, nine are open access journals (Journals of Gerontology Series A: Biological Sciences and Medical Sciences, Journals of Gerontology Series B: Psychological Sciences and Social Sciences, Geriatrics and Gerontology International, European Journal of Ageing, Zeitschrift für Gerontologie und Geriatrie, Canadian Journal on Aging – Revue canadienne du vieillissement, Generations: Journal of the American Society on Aging, Turkish Journal of Geriatrics – Türk Geriatri Dergisi, and Topics in Geriatric Rehabilitation). To analyze the author, country, reference, and other factors, the data must be verified for accuracy and identifiable information. A total of 2,520 (2.54 percent) original works written by anonymous authors were removed; 96,674 (97.46 percent) data remained in multiple forms, including articles, bibliographies, biographical items, book reviews, corrections, and editorial materials. Subsequently, these data were sorted and selected on the basis of two criteria: article format and English-language usage. Ultimately, 31,163 pieces of data accounting for 31.41 percent were selected for further analysis.
3.2 Research methodology

This study explored the use of Author Keywords and KeyWords Plus in the retrieved articles. Author Keywords are the list of terms selected by the authors to best represent the content of their paper (Li et al., 2009). Meanwhile, KeyWords Plus is the keywords maintained by the Web of Science. It is extracted from the titles of the cited references by Thomson Reuters by applying an automatic computer algorithm. Keywords from KeyWords Plus are terms or phrases that appear frequently in the titles of an article’s references. They sometimes do not appear in the title of the article as the Author Keywords (Zhang et al., 2016). Analyzing the Author Keywords, the number of keywords, and the relationship between these allowed this study to determine the research trends in the gerontology field. Additionally, KeyWords Plus was used to investigate the knowledge structure of each scientific field. Therefore, Author Keywords were used for analysis at the document level, whereas KeyWords Plus was used for analysis at the scientific field level. Scientific keywords may change over time, but the KeyWords Plus can help researchers explore a greater number of articles using different terms and increase article visibility, which may not have been possible using previous search methods (Garfield and Sher, 1993; Zhang et al., 2016).
Visual and trend analyses were employed in this study. Based on the sorted data, the number of studies, country, subfields, Author Keywords, and KeyWords Plus were analyzed year by year. Additionally, author network analysis was employed. The three common methods of exploring the relationship between contexts and mining their knowledge structure are bibliographic coupling, co-word, and co-citation. To explore the relationship between research contexts and knowledge structure, the three common methods are bibliographic coupling network, co-word network, and co-citation network analysis, of which bibliographic coupling employs citation analysis between studies. In this study, the bibliographic coupling relationship was explored between related studies (the top 200 cited citations) that shared at least one bibliographic coupling. Where two studies had similar citations, a stronger bibliographic coupling strength between the two studies was indicated. It also implied that the studies’ topics and knowledge are similar. The network of bibliographic couplings generated by both studies was static. Their references did not change after publication (Kessler, 1963). A co-word network is when studies used the same Author Keywords. Greater replication of similar keywords indicated that studies were, to a great extent, similar. Furthermore, the keywords were grouped into the same sub-domain to explore the research priorities (Cambrosio et al., 1993). Co-citation indicates that two studies used the same references, and they appeared in an initial publication as well as later publications. This indicated that the two studies shared a research topic. Moreover, where more citations were used in both studies, it indicated stronger similarities and co-citation relationships between the two studies (Small, 1973). According to the above discussion, our research procedure can be shown in Figure 1, which includes the steps of data collection, visual analytics, and author network analysis.

4. Findings

4.1 Visual analytics

Let us start with the discussion of visual analytics on research subfields. Figure 2 shows details on the research in specific subfields for each year from 2000 to 2016. The research during this period can be divided into several subfields. Most studies were published on psychiatry (5,664 articles; 18.18 percent), psychology (3,309; 10.62 percent), psychiatry and psychology (1,499; 4.81 percent), nursing (1,469; 4.71 percent), educational research (1,004; 3.22 percent), and healthcare science and services (837; 2.68 percent). Other subfields accounted for less than 2 percent of the total, including women studies (609; 1.95 percent), sport sciences (599; 1.92 percent), and rehabilitation (519; 1.67 percent). However, 15,654 (50.23 percent) studies could not be classified into specific sub-domains. A slight decrease in the total number of publications in 2003 and 2004 is clear. Overall, the annual distribution of each sub-domain differed. The largest proportion comprised unclassified sub-domain literature. Meanwhile, the total number of articles published annually has increased gradually during this period, doubling within 17 years from 1,146 papers in 2000 to 2,691 papers in 2016. However, a slight decrease in the number of published papers occurred in 2003 from 1,296 articles to 1,229 articles. Subsequently, the trend of increases resumed in subsequent years. Although other
subfields gradually increased year on year with less than 100 studies each year, psychiatry and psychology both exhibited more sudden changes. Specifically, psychiatry increased from approximately 200 studies in 2006 to reach a peak of more than 400 studies in 2014 after a gradual increase from 2000 to 2014. However, the number of the studies of this topic decreased slightly in 2015 and 2016. Similarly, psychology peaked in 2014 with approximately 200 studies, but decreased over the following two years.

We also conducted visual analytics on the most frequently used Author Keywords. As shown in Figure 3, the top 10 Author Keywords was identified: aging (2,844; 9.13 percent), dementia (2,787; 8.94 percent), depression (1,737; 5.57 percent), elderly (1,621; 5.20 percent), older adults (1,209; 3.88 percent), Alzheimer disease (1,171; 3.76 percent), aged (830; 2.66 percent), cognition (686; 2.20 percent), mortality (619; 1.99 percent), and long-term care (609; 1.96 percent). These can be divided into three categories: diseases (dementia, Alzheimer disease, mortality, and long-term care), spirit (depression and cognition), and elders (aging, elderly, older adults, and aged). Figure 3 shows the increasing use of the top 10 keywords over the study period. The increase in use for each keyword differed but all occurred relatively quickly. In particular, the number of studies using the keyword older adults increased 158-fold between 2000 and 2017, followed by cognition (143-fold), aging (68-fold), and dementia (46-fold).

This study also examined the top 10 KeyWords Plus to explore the most frequently used words in references from 2000. The commonest keywords were health (3,779; 12.12 percent), dementia (3,385; 10.86 percent), older adults (3,253; 10.44 percent), Alzheimer disease (3,062; 9.83 percent), age (2,626; 8.43 percent), adults (2,406; 7.72 percent), people (2,305; 7.40 percent), prevalence (2,298; 7.37 percent), care (2,239; 7.18 percent), and population (2,134; 6.85 percent). Besides, the growth of top 10 ranking KeyWords Plus over this period is presented in detail in Figure 4. In particular, the keyword older adults had the fastest and most significant growth in KeyWords Plus growth factor from 0 in 2000 to 3,253 in 2016. Studies were increasingly related to the title of the manuscript. In addition, the top 10 keywords can be classified into two groups on the basis of lexical relevance: population related (older adults, age, adults, people, and population) and health related (health, dementia, Alzheimer disease, prevalence, and care). The analysis indicated that dementia,
Figure 3.
The top 10 ranking Author Keywords between 2000 and 2016.

Figure 4.
The growth of top 10 ranking KeyWords Plus between 2000 and 2016.
older adult, and Alzheimer disease are the three most frequently occurring keywords, both in Author Keywords and KeyWord Plus. This finding shows that scholars are increasingly studying these areas.

4.2 Author network analysis
Subsequently, the top 200 studies among more than 30,000 publications were selected, analyzed, and summarized to identify the key topics of research in the field of seniors in 2000. This section discusses the relationship between studies through co-word, bibliographic coupling, and co-citation analysis and explores the contemporary situation. The bibliographic coupling of the top 200 studies was analyzed, and the top 200 studies were categorized into 12 groups depending on their similarity in terms of Author Keywords and KeyWords Plus (Figure 5). Group 1 included words related to the risk of falling and prevention; Group 2 included words related to mini nutritional and elderly; Group 3 included words related to medications and older adults; Group 4 included words related to lifespan; Group 5 included words related to dementia and memory; Group 6 included words related to leg strength, power training, and disablement; Group 7 included words related to the risk of falling and accidental falls; Group 8 included words related to oral health and daily living; Group 9 included words related to Alzheimer disease and mild cognitive impairment; Group 10 included words related to dementia, pathology, and lesions; Group 11 included words related to older drivers

Figure 5.
Bibliographic coupling network map
and motor vehicle collisions; and Group 12 included words related to falls, nursing facility, and hospitalization.

The network between top 20 Author Keywords which were cited most among the first 200 cited authors is shown in Figure 6. They can be divided into two categories according to their common use. The first group is mainly related to the basic description (age, women, disability, risk, health, risk factors, performance, men, strength, adults, skeletal muscle, older adults, body composition), and the second group contains health-related words (community, mortality, dementia, Alzheimer's disease, decline, population, older adults). The network between the top 20 Author Keywords that were the most cited of the 200 most cited studies is shown in Figure 6. They can be divided into two categories depending on their use. The first group contains keywords related to the basic description (age, women, disability, risk, health, risk factors, performance, men, strength, adults, skeletal muscle, older adults, and body composition), and the second group contains health-related words (community, mortality, dementia, Alzheimer disease, decline, population, and older adults).

The co-citation of the top 20 most cited keywords of the 200 most cited studies was analyzed to determine the relationship among the studies. This study used Medical Subject Headings (MeSH) terms to explore study titles. MeSH terms comprise the controlled vocabulary thesaurus developed by the National Library of Medicine. After each study was read, the MeSH term was marked with an asterisk. Words with an asterisk were generally related terms, and this study was based on screening criteria. Words were divided into four clusters. The four groups identified using the MeSH term were mainly used to explore the scope of the topic: Specifically, the green cluster was used to explore frail elderly and life function; the yellow cluster, mental disorder; the purple cluster, accidental wrestling; and the red cluster, psychology. The red cluster is associated with the yellow and green clusters and is unrelated to the purple cluster. This implies that psychology has a strong connection with

Figure 6.
The co-author network map
mental disorder, and frail elderly and life function. Similarly, the yellow cluster is associated with the red and green clusters and is unrelated to the purple cluster, indicating that mental disorders has a strong relationship with frail elderly and life functions and psychology. The purple cluster is associated only with the green cluster, and a strong correlation exists between the two clusters, representing a strong relationship between frail elderly and life and accidental wrestling. In addition, Folstein et al. (1975) (shown as Folstein (1975) in Figure 7), Guralnik et al. (1994) (shown as Guralnik (1994) in Figure 7), Guralnik et al. (1995) (shown as Guralnik (1995) in Figure 7), and Ferrucci et al. (2000) (shown as Ferrucci (2000) in Figure 7) had the strongest co-citation relationships. That is, the green cluster was the principal cluster of the literature network, and a close correlation existed between the green cluster and the other three.

5. Conclusion
In this study, the literature on gerontology from 2000 was bibliometrically analyzed. The USA was the most productive country for publishing research in this domain during the period 2000-2016. Furthermore, most publications in this field came from the top 10 publishing nations, which accounted for more than 80 percent of the total published studies. Psychiatry was the most frequently discussed sub-field, followed by psychology. The three most frequently used keywords were aging, dementia, and depression. The most frequently used keywords were categorized into three main groups: diseases, spirit, and elders. Additionally, the results for KeyWords Plus were similar: the most frequently used keywords were health and dementia. Older adults was the keyword with the fastest growth in this field. The top 10 KeyWords Plus were classified into two categories: population related and health related. The keyword whose use increased the fastest in both categories was older adults. The commonest keywords for the two categories were dementia, older adults, and Alzheimer disease, which shows that dementia among older adults has become a key research topic worldwide. This result provides research directions for fellow researchers conducting studies on gerontology.
Additionally, the 200 most frequently cited papers were analyzed through bibliographic coupling, co-word, and co-citation analysis. This study explored the topics that scholars have focused on in gerontology since 2000. The selected most cited papers were mostly published before 2007, and the USA was the dominant country for publication. The 200 major citations were input for bibliographic coupling analysis; this analysis yielded 12 groups that were similar in terms of Author Keywords and KeyWords Plus. Co-word analysis of the top 20 most cited keywords, the four clusters and their inter-relationships were identified. In particular, the frail elderly and life functions cluster was the most cited cluster and had interrelations with the other three clusters. Moreover, the psychology cluster had a strong connection with the mental disorder and frail elderly and life function cluster but no connection with the accidental wrestling cluster. This paper highlighted the most crucial research topics and subfields in the gerontology field.

Researchers can use the findings of this study to determine directions for subsequent research as this study provides a comprehensive view of the research field as well as an understanding of the subfields and their interrelations. This paper empowers scholars to use their resources to investigate this field. Additionally, this paper provides government departments with directions for formulating and executing policies affecting older people not only in setting academic and professional priorities but also in understanding the key topics related to older people.

In short, this study conducted the approach of bibliometric networks and analytics on modern research on gerontology to provide an exhaustive understanding of this field from an academic perspective. However, the study had some limitations due to constraints in time and resources. First, this study collected only papers written in English since 2000. Therefore, research in other languages was omitted in data collection. Future research should consider using different languages from other databases or Google Scholar as screening criteria. Second, this study sorted and selected only articles to form its data set. Future research should use additional document types to construct a data set. Last, future research should increase the number of most cited references for author network analysis to more profoundly understand this topic.

References


Corresponding author
Chien-wen Shen can be contacted at: cwshen@ncu.edu.tw
Indexing Arabic texts using association rule data mining

Ramzi A. Haraty
Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon, and
Rouba Nasrallah
Lebanese American University, Beirut, Lebanon

Abstract
Purpose – The purpose of this paper is to propose a new model to enhance auto-indexing Arabic texts. The model denotes extracting new relevant words by relating those chosen by previous classical methods to new words using data mining rules.

Design/methodology/approach – The proposed model uses an association rule algorithm for extracting frequent sets containing related items – to extract relationships between words in the texts to be indexed with words from texts that belong to the same category. The associations of words extracted are illustrated as sets of words that appear frequently together.

Findings – The proposed methodology shows significant enhancement in terms of accuracy, efficiency and reliability when compared to previous works.

Research limitations/implications – The stemming algorithm can be further enhanced. In the Arabic language, we have many grammatical rules. The more we integrate rules to the stemming algorithm, the better the stemming will be. Other enhancements can be done to the stop-list. This is by adding more words to it that should not be taken into consideration in the indexing mechanism. Also, numbers should be added to the list as well as using the thesaurus system because it links different phrases or words with the same meaning to each other, which improves the indexing mechanism. The authors also invite researchers to add more pre-requisite texts to have better results.

Originality/value – In this paper, the authors present a full text-based auto-indexing method for Arabic text documents. The auto-indexing method extracts new relevant words by using data mining rules, which has not been investigated before. The method uses an association rule mining algorithm for extracting frequent sets containing related items to extract relationships between words in the texts to be indexed with words from texts that belong to the same category. The benefits of the method are demonstrated using empirical work involving several Arabic texts.

Keywords Precision, Recall, Arabic text, Auto-indexing, Frequent sets, Rule-based data mining

Paper type Research paper

1. Introduction
Indexing text documents consists of analyzing the content of the text in order to retrieve its subject. It is a very important task related to information retrieval: a field that is considered imperative in the computer science domain due to the need of exploring different topics in our daily lives. For instance, most of us use “Google,” a search engine used to search and retrieve information about different topics.

Manual indexing is very difficult and needs an immense human effort from an expert, since one needs to read the entire text and analyze it holistically. The longer the text, the longer the time needed.

Indexing texts is needed in many domains and for different types of texts such as: articles in newspapers and magazines, online articles, archiving, documenting, helping in e-mail spam detection, web page content filtering, and automatic message routing. But most importantly it is used for information retrieval.

There are two types of indexing techniques: thesaurus-based indexing and full text-based indexing (Khoja, 2001). In the thesaurus-based indexing technique, the words chosen to represent the document do not need to exist in the text. However, their synonyms exist. The synonyms might be chosen by the documenter if the search for the text is usually done by using the
synonym and not the exact word. Note that the synonym is not only the correspondent of the term in the dictionary. For example, if we are talking about a football player, we can consider his/her name as an index based on the thesaurus system. The problem of thesaurus-based indexing is the difficulty in implementation because of the need of human intervention. The implementation should depend on a file containing the synonyms of the terms. This file should continuously be manually updated. In full-text based indexing, the indexing relies on choosing terms or phrases that already exist in the text. This type is much easier in concept.

The significance behind this work is not having the documenter read the document, choosing the key words and building the subject of the document or the subject headings, but a computer program that picks the key terms of the text automatically and returns them as output to the documenter. This will facilitate the work and spare the documenter time and effort. Therefore, the goal is to automate one part of the work. In other words, the proposed solution is to extend the set of relevant words extracted from the text based on relations of words extracted from the prerequisite texts of the same category of the text to be indexed.

In this paper, we present a full-text based auto-indexing method for Arabic text documents. This is a very important task to be achieved since we have many Arabic publications such as books, articles in newspapers as well as online blogs that require analysis and indexing. Our auto-indexing method extracts new relevant words by using data mining rules. The method uses an association rule mining algorithm for extracting frequent sets containing related items to extract relationships between words in the texts to be indexed with words from texts that belong to the same category. The benefits of our method are demonstrated using empirical work involving several Arabic texts.

The rest of the paper is organized as follows: Section 2 provides a literature review and related works. Section 3 discusses the processing steps of the algorithm. Section 4 outlines stem word extractions. Section 5 delineates weight calculation and index extraction. Section 6 presents the experimental results and Section 7 concludes the paper.

2. Literature review

Over the years, a number of methods have been proposed related to the classification of algorithms for text categorization. These methods are based on classification algorithms such as Naïve Bayes proposed by McCallum and Niagam (1998), decision trees by Sahami et al. (1998), neural networks by Joachims (1998) and Harrag and El Qawasmeh (2009), and Support Vector Machine or SVM by Dumais (1998) and Turney and Pantel (2010). In Yang and Liu (1999), some of these methods were compared. The results show that SVM is the best among others regarding the document classification.

As for the Arabic language, numerous works have been done to process texts. Haraty and Khatib (2005) introduced a procedure that extracts temporal elements from a document. Haraty and Hamid (2002) presented a technique to segment hand-written Arabic text, and Haraty and Ghaddar (2004) put forward two neural networks were built to classify already segmented characters of handwritten Arabic text. Al-Harbi et al. (2008) discussed automated document classification considering it an important text mining task. Text classification aims to automatically assign the text to a predefined category. The authors proposed a solution based on linguistic features. They generated the feature frequency of the lexical features that they have extracted and then they calculated the importance of each feature locally (for each class) based on $\chi^2$. Seven data sets were used including essays, pieces of literature, poems, web, forums and others, and each data set contained different classes. The assembled corpus comprised 17,658 texts with more than 11,500,000 words. A tool was implemented to extract features using SVM and C5.0 (classification algorithms) to decide which class the tested texts belong. The results showed that, in general, the C5.0 classifier is more accurate.

In Al-Anzi and AbuZeina (2017), the authors used singular-valued decomposition as a feature reduction technique as well as for producing semantic rich features and to truncate the
term-document matrix (produced by latent semantic indexing). They also proposed to remove small words in the preprocessing phase based on certain length threshold. In addition, they presented a comparison between the different classification methods and it was shown that cosine similarity measure is a good option to be considered for Arabic language text classification.

In Alazzam and Alsmady (2017), the authors proposed an approach that consists of a learning phase in which sample documents are selected and classified manually. The distribution approach presented in their proposed solution consists of passing the manually classified texts into more than one machine (one class per machine). The text classification is done based on LSA, and cosine similarity, and can be applied to big data. The proposed approach reduced the time required for classification.

Al-Molijy et al. (2012) presented a method to automatically create and index books written in the Arabic language. The method depends largely on text summarization and abstraction processes to collect main topics and statements in a book.

Khader et al. (2014) presented a tool to assist the indexing of offline handwritten historical documents. The tool has a semi-automatic interface that can deal with possible errors in text segmentation.

Jilani and Mohammed (2016) offered a technique for the concatenation of the individual stand-alone Arabic characters which are extracted and recognized from image frames. The authors used Unicode format of Arabic characters for the concatenation of extracted characters.

Khoja (2001) proposed a tagger that uses rule-based techniques and statistics to solve the problem of auto-indexing. The Arabic parts-of-speech (POS) studied are divided into: verbs, nouns, and particles. All affixes are removed using a stemmer to produce the stem word of each term in the text.

Gawrysiak et al. (2002) presented the shortcomings of the unigram and n-gram document representation techniques used in text mining, and proposed a solution that takes into consideration the position of a word in a text document. In Larkey and Connell and Billhardt et al. (2000) approaches to handle co-occurrences were introduced.

Mansour et al. (2008) proposed a method based on morphological analysis (using grammatical rules for extracting stem words) and on a technique (which computes the spread of the word in a document) for assigning weights to words.

Sharef et al. (2014) used a new classifier called FRAM for Arabic text categorization. Their experiments were divided into two stages: the training stage and the test stage. During the training stage, they pre-processed the documents that initially belong to a specific category and then extracted the important keywords. After that, they formed a database of the feature terms and used it in the testing stage. In the testing stage, they considered a set of categorized documents as uncategorized, and classified them using a classifier. They measured the categorization performance by using several standard techniques of performance evaluation. The purpose of using the classifier FRAM was to calculate the frequency ratio for each feature of the new document based on the candidate features of the training phase. They concluded that FRAM was better than three Bayesian learning classifiers: MBNB, MNB and NB.

Hattab and Hussein (2012) introduced a three-model system. The first deals with the pre-processing phase. Then the result of the first model is cleaned while going through the second model consisting of error detection and correction of the corpus used. After that, comes the classification model where it is decided to which domain the text belongs. In the classification model, each document of the training corpus will be classified. Then, the N-gram profile will be generated by splitting the text into tokens consisting only of letters while removing all digits. The frequency of occurrence of each N-gram is calculated and sorted from the highest frequency to the lowest. The similarity of the N-grams profile of each document is compared with the profiles of all the documents in the training classes. For each N-gram in the document profile, search is performed for the N-gram in the class profile, and then the difference between their positions is calculated.
For N-grams that are not found in the class profile, a maximum value is assigned. The second measure, after all N-grams in the document profile have been exhausted, is computing the sum of the distance measures. The detection and correction algorithm outperformed the Bayes algorithm by about 10 percent. Without checking misspelling errors, accuracy is 68.85 percent, while the average accuracy for the classification system with misspellings detection and correction is 71.77 percent.

El-Shishtawy and El-Ghannam (2012) considered an accurate Arabic root-based lemmatizer for information retrieval purposes. The aim was to build a lemmatizer with minimum sufficient resources. Lemmatizer transforms inflected word form to its dictionary lemma look-up form. For example, in the case of nouns and adjectives, the lemma form is the singular indefinite (masculine if possible) form. In the case of verbs, it is the perfective third person masculine singular form. Through it, they can collect more information about the word to be stemmed and its context to generate more accurate word features. The system generates accurate lemma form and its relevant morpho syntactic features that support information retrieval purposes. Morpho syntactic features are also required to capture the important semantic senses of the language. Collected information about words also include word pattern. During the implementation of the approach, the authors took into consideration the following features: the power and generality of rule-based stemmers and the accuracy of dictionary based approaches, the Arabic context morpho syntactic rules used to expect the correct word category and then verified. For example, the algorithm uses the word pattern and the category of its previous word for identifying the current word, and the simple rules used to re-categorize nouns as adjectives. The presented algorithm proves that accurate results for POS tagging (part of speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context), can be achieved when using inherent features and rules of Semitic languages like Arabic. Ambiguity can be resolved using metadata about patterns, roots, and infixes’ indications. Analysis is aided with auxiliary dictionaries and syntax rules to produce a lemmatizer.

In the work of Daher (2002) and Haraty et al. (2003) a model that consists of four layers was proposed, each independently implemented. The layers are connected and exchange information. The model consists of phases, where the first one is reading the document; the second is extracting stem words (but this phase can easily be plugged off the system and replaced by another stemming technique), the third and fourth phases consist of assigning the weight and selecting the index words. In the first phase, the model uses a stop-list containing terms (words/phrases) that should not be taken into consideration while choosing the index words since they do not contribute to the meaning of the text. After removing the stop-list terms, the second phase begins. In this phase, the remaining words are considered candidate words that need to be stemmed. The stem words will form an array that is used as input of the third phase in which the weight calculation is done. In the fourth phase, the index words are selected. All the phases of the model are important, nonetheless the phase in which the weight is calculated is considered the core of the system. Three factors influence the weight of the stem word: the frequency of appearance of the word in the text is one of the factors. The second is the count of the stem words for that word. The authors proposed adding another factor which is the spread of the word in the text. This is to prevent choosing a word as an index if it appears in a specific part of the document and not in all of it because in this case it is less likely that the word represents an index word of the text.

In Mahmood et al. (2014), the authors proposed a method that analyzes medical blogs where patients describe their conditions by presenting the symptoms, problem identification and treatment. The analysis of these blogs will result into extracting positive and negative association rules from text data sets – negative association rules from the frequent item sets, and the opposite which is the extraction of positive association rules from infrequent item sets. They used inverse document frequency (IDF) weighting techniques that assigns high
weight to terms that occur rarely in a document since it helps understanding the terms that carry special significance in a certain type of document corpus. In other words, IDF gives the higher weight to the words that have strong connection to the topic of the text. Then they applied association rule data mining techniques using the Apriori FISinFIS algorithm on the index words with the highest weight by IDF. This generated frequent and infrequent item sets. Then they proposed finding among the items in frequent item sets ones that are infrequent together and among the items in infrequent item sets ones that are frequent.

Bhujade and Janwe (2011) highlighted the importance of using association rules in text mining. The phases of the proposed solution stressed on pre-possessing and stemming. The authors applied the GARW association rule algorithm on indexes selected from a text based on the IDF algorithm.

In Qureshi et al. (2015), the authors proposed an associative classifier by combining the association rule data mining algorithm – Apriori – with pruning methods to classify text documents. The results showed improvement in the accuracy of classification compared to other classical methods. They extracted the association rules related to several text categories using Apriori and only kept the rules with higher confidence and support for each text category. This allowed them to extract the different association rule sets related to different classes of text documents. After this, they found the category of texts that has the maximum number of rules that matches the text to be classified. If more than one category is considered relevant based on a certain threshold, they choose the category with maximum confidence average.

3. Preprocessing steps
Prior to any indexing or classification work done on texts of any language, pre-processing should be done. This is especially true with indexing, which consists of extracting the most relevant words out of a text. Pre-processing was presented in previous work of Daher (2002) and Mansour et al. (2008) that relies on the frequency and spread of words in a text to be considered as indexes. Our proposed solution uses the indexes extracted using the work done in Daher (2002) and Mansour et al. (2008) and associate to it new relevant words. Before we elaborate more on our proposed solution, we will review the pre-processing work done in Daher (2002) and Mansour et al. (2008) as it is considered as a preliminary step in text indexing (see Figure 1).

First, the pre-processing algorithm starts by rhyming the words of the text. This helps identifying verbs and nouns, which are the two types of words in the Arabic language. Also, this step is used to identify pronouns that are attached to words. Second comes the step of removing stop-words out of the text to be indexed. These words should not be taken into consideration while extracting the index terms.

Third, this pre-processing relies on step one to differentiate between the two types of terms: verbs and nouns, to be able to perform the right stemming technique that changes from one type to another.

Fourth, stem words should be extracted. Stemming consists of extracting a word’s root to avoid extracting duplicate indexes with the same meaning. This step allows the auto-indexing algorithm to treat different forms of the same term as one word and not consider it independent. Note that different forms of the word mean for example, singular or plural form, or adjective or subject. This improves the indexing mechanism by reducing redundancy; for example, a word and its plural have the same meaning and contribute equally to the subject of the entire text. Therefore, they should not be treated differently. We should not forget that the count of stem terms of a certain word in the text is one of the factors that contribute to the weight calculation; therefore, stemming is a preliminary phase of the auto-indexing work.

3.1 The rhyming step
Rhyming is the first step in the stemming mechanism. In this step, we need to take into consideration that in the Arabic language we have two types of words: verbs and nouns.
Each type has a different rhyming that should be relied on to differentiate between verbs and nouns. In addition, one can identify through rhyming the tense of the verb (past, present or future) and whether a noun is singular or plural. It is also used to identify attached pronouns so they can be removed. Note that in the Arabic language all verbs rhyme with different standard validations of the word “فعل” (i.e. did).

3.2 Removing stop-list terms

The second step consists of removing stop-list terms that are words that do not contribute to the meaning of the text. A predefined stop-list is used to remove all the stop-words from the text to be indexed. As an example, in the English language we can consider the words “it,” “the,” “never,” “where,” “that,” “numbers,” etc. as stop-list terms. However, we should mention that these can also be used to identify whether the word that follows is a verb or a noun.

On a more thorough level, we can also talk about stop-list phrases. These are sentences with no contribution to the meaning of the text – “To all my dearest readers” – for example, is a sentence that does not add any value to the text meaning. The auto-indexer should not take into consideration the stop-list terms and phrases.

3.3 Identifying nouns and verbs

Identifying nouns and verbs is carried out in order to extract the stem of a word based on its type. This consists of two steps:

1. Some of the stop-list terms are always followed by verbs, while others precede nouns. So if a word is preceded by any of the stop-list terms that are always followed by a verb, then it is a verb. The same is applied to identifying nouns. For example, the word “لم” that means “never,” is a stop-list term that is always followed by a verb.
Distinguishing between verbs and nouns can rely on rhythm of the word itself. Some of the used rhymes are specific for verbs; for example, “فعل” (i.e. does) or “فعل” (i.e. do). Others are for nouns, for instance “فاعل” (i.e. doer) or “فعل” (i.e. done).

In the case of words that cannot be identified using these two methods, we examine the pronouns attached to the word. Some of the pronouns are only attached to nouns, while others are attached to verbs.

4. Stem word extraction

Stemming allows the auto-indexing algorithm to treat different forms of the same term as one word and not consider them independent. The count of stem terms of a certain word in the text is one of the factors of the weight calculation; therefore, stemming is a preliminary phase of the auto-indexing work.

4.1 Extracting stem words from verbs

Different stemming procedures need to be taken into consideration in the “Stemming Algorithm” because we have different types of verbs in the Arabic language. The output of any algorithm is the original form of the verb.

4.1.1 Checking attached prefix/suffix pronouns. The first step in the stemming algorithm is to remove the pronouns that are attached to a word as suggested by Gillman (1990). One of the characteristics of the verbs in the Arabic language is having two forms of pronouns: attached and discrete. The discrete pronouns are stop-list terms and are eliminated in the phase that handles removing them. A word may consist of attached pronouns that are very important to identify in order to separate them from the word. We have a defined finite list that contains all the pronouns.

4.1.2 Checking verbs against the “five verbs”. Another characteristic of the Arabic language is having five standard verbs known as the “five verbs” and can be found in Deeb (1971) as well as Kindery et al. (1996). One of the important things about these verbs is that they are always in the present tense and they usually end with “n” or “ن”. Also, non-essential letters are attached to these verbs but not considered as pronouns. Therefore, these attached letters are not detected in the phase in which attached pronouns are removed. As an example, the verb “فعالون” (i.e. they do), where “ي” and “ون” are non-essential. However, they are not removed with the attached pronouns. Rhyming is used instead of pattern matching to detect whether a verb belongs to the “five verbs” set or not.

4.1.3 Checking verbs against the “ten-verb-additions”. The third characteristic of the Arabic verbs is the ten different derivation formats that come from the original three letter verb. These derivations, similar to the “five verbs,” have essential letters and non-essential letters. Three of the ten-derivations are obtained by adding one letter to the original stem verb, five are obtained by adding two letters, and the other two derivations are obtained by adding three letters. Rhyming is also used in this case to detect the verbs that are considered a derivation.

4.2 Extracting a stem word from a noun

Different factors affect the difficulty of stemming a noun. The first is the different forms in which a noun might occur in the text, whether singular, double or the plural form. Needless to say that these forms also differ when the noun addresses a male or a female. The steps of extracting a stem word from a noun are: check attached prefix and suffix pronouns; check the noun against the “five nouns”; restore a noun to its singular form; and compare the noun to the “common derivations”: M-derivations, T-derivations, and miscellaneous derivations.
5. Weight assignment and index selection

The previous work presented in Daher (2002) and Mansour et al. (2008) consisted of selecting index words out of the text based on their count as well as their spread along the text. However, many words that do not appear frequently or are not spread along the text might be relevant and should be considered as candidate indexes. We propose using association rule data mining methods to relate indexes selected based on Daher (2002) and Mansour et al. (2008) to new words that are relevant. Pre-requisite texts of the same category of the text to be indexed are used to extract words related to each other using the association rule data mining algorithm. In Section 5.1 we re-introduce the weight calculation method presented in Daher (2002) and Mansour et al. (2008) since our work depends on it.

5.1 Weight calculation

The weight calculation phase is considered an important step, on which our work relies to extend the list of index words; and therefore, will be described in the following sub-sections.

5.1.1 Factors affecting the weight. The weight of a word is calculated depending on three factors: the count of the term in a text document, the count of the stem words for that word, and the spread of that word. The last factor was introduced based on the fact that if a word appears frequently in a specific part of the text, it does not necessarily reflect the subject of the text. The more spread the word is along the text, the more it should be considered an index word. These three factors are used in formulas to calculate the weight of the words.

5.1.2 Formulas and terms used. In this section, we present the formulas used in Daher (2002) and Mansour et al. (2008) to calculate the count of the word in a text and its spread along it. Based on these formulas the indexes of the text to be indexed are selected and then they are related to new words using our approach. Interested readers are referred to Daher (2002) and Mansour et al. (2008) for more details. Calculating the count of a word and its stem is not of any difficulty since it can be done by counting the occurrence of each term in the text. The calculation of the spread factor takes more effort, and needs to be calculated using a complex formula. The weight calculation is performed using the following formula:

\[ w = m \times sm \times f \]

In this formula, \( m \) is the count of a certain word in the text, \( sm \) is the count of stem words for a certain word in the text. The last factor which is the spread of the word within the document is represented by \( f \). The original formula was developed in (Daher (2002) and Mansour et al. (2008). This formula takes into consideration the following factors:

- \( d \) is the word’s position in the text, i.e. its distance. It is found by counting the words preceding it in the text.
- \( ad \) is the average of all distances for each stem word of each term in the text, defined as:

\[ ad = \left[ \frac{\sum_{i=1}^{sm} d_i}{sm} \right] \]

- \( id \) is the ideal distance between every two occurrences for each stem word:

\[ id = \left[ \frac{N}{sm+1} \right] \]

Notice that \( N \) is equal to the count of all terms in document.
• $\text{aid}$ is the average of all ideal distances for a stem word and is given as:

$$\text{aid} = \frac{\sum_{i=1}^{sm} i \times id}{sm}$$

$$\sum_{i=1}^{sm} i = \frac{sm \times (sm + 1)}{2} \text{ and } id = \left\lfloor \frac{N}{sm+1} \right\rfloor, \text{ aid becomes } \left\lfloor \frac{N}{2} \right\rfloor$$

• $g$ is the difference between $\text{aid}$ and $\text{ad}$ called gap, i.e. $g = \text{aid} - \text{ad}$.

Note that $g$ may have a positive or a negative value. The smaller $g$ is, the more the word is spread in the text.

5.2 Proposed solution for index selection

In Daher (2002) and Mansour et al. (2008), the authors proposed a method to index Arabic texts based on the frequency of appearance of words and their spread along the text. Of course pre-processing techniques and stemming were presented by the authors as well. To enhance the work done in this field, we propose taking advantage of the relations between words in a text. This will allow generating new index words out of the text that might not appear frequently in the text; however, they are somehow related to the topic and should be considered index words. Introducing association rule based data mining techniques to index Arabic text is an innovative proposed solution. In addition to its novelty, the proposed solution will make the auto-indexer that is based on relations of words more accurate since it will expand the number of relevant indexes. For example, if we have a text from the category of "استخراج النفط" (i.e. oil extraction), the word "بئر" (i.e. barrel) might not be considered an index word of the text, having taken into consideration the count of the word and of its stem, as well as the spread of the word in the text. So, why not take advantage of the fact that this word is related to the term "النفط" (i.e. petroleum) and they both appear together frequently in texts of the same category! In this case, the use of relations between words and their integration in an auto-indexing system will allow the extraction of new relevant words related to the subject of the text. In our example, a new relevant index word will be "بئر" (i.e. barrel), not appearing frequently in the text; however, contributing to its meaning.

To generalize, we take TEXT1 as the text to be indexed. TERM1 is not an index of TEXT1. However, TERM2 is an index. Many texts from the same category of the TEXT1 are used. If TERM1 and TERM2 appear frequently together in these texts, they are related together. We assume that since TERM1 is related to TERM2 that is an index of TEXT1, then TERM1 should be considered a relevant word of the text. This means that TERM1 contributes to the meaning of the text and should be considered an index word.

The proposed solution is based on four phases: collecting the pre-requisite texts, generating item-sets of frequent patterns, extracting relevant frequent sets, and selecting the candidate index words.

5.2.1 Collecting the pre-requisite texts. First, the work relies on choosing pre-requisite texts from the same category of the document to be indexed in order to build relations between words. These texts are considered the “fuel” of the engine that represents the proposed model.

The average number of words that index the text is considered to be around five as per an Arabic language expert who helped us in the manual indexing of the sample texts used in the experimental results. The pre-requisite texts can be indexed using, for example, the model proposed by (Daher, 2002). Out of the index words of each pre-requisite text, the top five terms with the highest weights are considered to represent the text the most; and therefore, are chosen.
We end up having a number of sets each corresponding to a text and containing five terms. Words in these sets appear more likely in other texts of the same category of the text to be indexed. This is the main reason behind using relations of words in texts of the same category. For example, all texts belonging to the category of “استخراج النفط” (i.e. oil extraction) have words in common such as “النفط” (i.e. petroleum), “الدollar” (i.e. dollar), and “البرميل” (i.e. barrel). These words appear frequently together in a text from this category and are considered to be related to each other. The more pre-requisite texts we have, the more relations of words we have; and therefore, the results become more accurate.

The pre-requisite sets of words will be used in the second phase of our model in which the relation between words is extracted and the words that appear frequently together will be stated. This is done using the “Apriori” algorithm.

5.2.2 Generating item-sets of frequent pattern. Next, we need to build sets of items appearing frequently with each other; and therefore, related. The Apriori algorithm is used to achieve this task. Apriori is an algorithm that generates frequent item sets and association rules over transactional databases. After identifying the frequent individual items in a database, Apriori extends them to larger item sets that appear frequently often in the database. According to Bhalodiya et al. (2013), Apriori generates the candidate item-sets of frequent patterns that appear together. It is used to extract association rules of items in different sets and generalize the trends of the database.

In our case we will use this algorithm to extract associations between words of the different sets, extracted from the pre-requisite texts, which will highlight the relations between words in texts of the same category. Based on this, we can extend the candidate indexes of a text. The relations between words are used in the indexing mechanism that we are proposing in order to extract more relevant words out of the text. The algorithm takes as an input the sets of words extracted out of the pre-requisite texts. Out of these sets a list of items is formed. This list contains all the words that appear in the pre-requisite sets but with a condition that the word will not appear twice in the list. Next, lists of transactions are formed. Each transaction corresponds to one of the pre-requisite sets. Every transaction contains the indexes of the words of the set it represents in the list containing all the pre-requisite words. To make this clearer, the following example is presented:

(1) $S_1$, $S_2$ and $S_3$ three pre-requisite sets as follows:
- $S_1 = \{\text{term}_1, \text{term}_2, \text{term}_3, \text{term}_4\}$.
- $S_2 = \{\text{term}_2, \text{term}_4, \text{term}_5, \text{term}_1\}$.
- $S_3 = \{\text{term}_5, \text{term}_3, \text{term}_7, \text{term}_6\}$.

(2) $L$ is the list containing all the pre-requisite words in a distinct way:
- $L = \{\text{term}_1, \text{term}_2, \text{term}_3, \text{term}_4, \text{term}_5, \text{term}_6, \text{term}_7\}$, notice that the first term in the list $\text{term}_1$ has the index 0, etc.

(3) $T_1$, $T_2$ and $T_3$ are the three transactions corresponding respectively to $S_1$, $S_2$ and $S_3$:
- $T_1 = \{0, 1, 2, 3\}$ since $\text{term}_1$ in $S_1$ is the first in $L$, $\text{term}_2$ is the second, etc.
- $T_2 = \{1, 3, 4, 0\}$.
- $T_3 = \{4, 2, 6, 5\}$.

These transactions are then used as input of the algorithm responsible for extracting the frequent sets of words related to each other. The output will be sets of words appearing frequently together in these transactions.
5.2.3 Extracting relevant frequent sets. Third, not all the frequent sets extracted in the previous phase by Apriori should be considered relevant. Some of them might not contain any of the indexes of the document we are trying to index using traditional methods. In this case the words in these sets are not related to any of the index words extracted from the text to be indexed. That is why this phase is introduced to extract only the frequent sets that are relevant.

After dealing with the pre-requisite texts and extracting the frequent sets of words that are related together, we should deal with the text to be indexed.

We built on the auto-indexing model proposed in Daher (2002) to extract the index words out of the document that we wish to index. We considered that the top five words with the highest weights represent the text the most.

Only the frequent sets that contain at least one of the index words of the text to be indexed are considered relevant because it presents new terms that are related to one of the indexes chosen using traditional auto-indexing methods. This is helpful to the auto-indexing mechanism since it expands the number of candidate relevant index words. Some of the words that were not considered as index words of the text to be indexed are now being taken into consideration.

5.2.4 The output of the proposed solution. Our proposed solution relates the outputs of previous solutions to new relevant words. The relevant frequent sets returned in the third phase of the proposed solution are the output of our work. These are sets of related words that appear frequently together in texts of the same category of the text to be indexed.

5.2.5 General algorithm. In this section, we present a high level algorithm of the model with the proposed solution discussed above:

```
// for text to be indexed
1. Apply indexing techniques to auto-index the document to be indexed

2. Transaction T = Select top 5 index words of docToBeIndexed
   // (indexed words sorted from highest weights to lowest)

// for each pre-requisite text
3. Transaction t_x = Select top 5 index words of pre-requisite doc // (indexed words sorted from highest weights to lowest)

//using Apriori
4. Array[] ar = getfrequentsets(t_1,..., t_n)

// Select all the combinations of the index words
5. Array[] ar1 = PowerSetPart(T)

// Select all the relevant frequent sets that contain index words
6. Array[] result
   int i = 0
   Foreach set in ar1 // for each combination of indexes
   {
      Foreach set1 in ar // for each frequent set
      {
         if set1.contains(set)
         {
            result[i] = set1
            i++
         }
      }
   }
```
First, the top five weighted index words are selected out of the text to be indexed. These indexes form transaction T. Then, the top five weighted index words are selected out of each of the pre-requisite texts chosen. These indexes form \( t_1, \ldots, t_x \) each corresponding to a text of the pre-requisites. Apriori is applied taking as input \( t_1, \ldots, t_x \). The output will be the sets of words that appear frequently together. The frequent sets that contain any of the index words of the text to be indexed represented by T are chosen. This will allow us to expand the set of words that are more likely to be indexes of the indexed text, which will produce more relevant words extracted out of this text. The implementation done in Daher (2002) and Mansour et al. (2008), on which we relied to get the top five index words based on the frequency of their appearance and spread along the text, was done using Visual Basic 6. We implemented our approach using C# language and Microsoft Visual Studio.

6. Experimental results

6.1 Case studies

Our case study consists of 37 texts selected arbitrary from the same category of "استخراج النفط" (oil extraction). We chose this category due to the variety of subjects that it has. Also, the topics of the texts in this category are related to "Petroleum" which is considered a very important topic frequently discussed in the Arab countries. Therefore, we consider the subject of "Oil Extraction" very important and validating our approach using texts having this topic would be very interesting. The category should be specified since we need to have access to the pre-requisite text from the same category of the text to be indexed to help achieve the work based on relations of words. The texts have different lengths so we can have various case studies on which we test the validity of our approach. The variety of scenarios makes the results more reliable and trustworthy.

The texts of the chosen category were used as pre-requisite texts needed in our approach in order to extract the frequent item sets of words related to each other to help extend the sets of relevant index words. Notice that these index words are selected based on the frequency of words, their stems, as well as the spread of the words along the text to be indexed. The experiments were conducted on 25 texts selected arbitrarily. To validate our results we used the help of an Arabic language expert to manually index the texts chosen in our case study. Also, the expert helped us choosing relevant and irrelevant indexes out of the words returned as indexes by the previous method’s (Daher, 2002) auto-indexer and those selected using our proposed solution.

6.2 Calculations

The experiments are based on calculating two factors used to validate the usefulness of the approach proposed in this paper. These factors are: the precision and the recall. The precision measures the percentage of relevant index words retrieved programatically out of all the words retrieved as candidate index words whether relevant or irrelevant. The recall measures the percentage of relevant words retrieved programatically out of the index words retrieved manually using the help of an expert in the Arabic Language. This means it measures the completeness of retrieval of relevant index words. First, we will start by defining some terms used in the calculations:

- \( N \): total number of words in a text;
- \( I \): the most useful (relevant) index words found by manual indexing;
- \( RR \): retrieved Relevant index words using the software we implemented;
- \( RI \): retrieved Irrelevant index words using the software we implemented;
- \( NRR \): none retrieved relevant words. \( NRR = I - RR \);
- \text{precision} = \frac{RR}{(RR + RI)}; \text{ and}
- \text{recall} = \frac{RR}{(RR + NRR)} = \frac{RR}{I}.

Any word that is returned as an output by the automatic auto-indexer, and the manual indexing, is considered relevant. Others that are returned programatically but not manually are irrelevant. As we explained earlier, our approach retrieves relations between words presented as item sets. Each word appearing in an item set with a relevant word is checked to be relevant as well. If it is selected by the manual indexing, then it is relevant.

The calculations of precision and recall will be affected by the new relevant words returned as an output by our method.

### 6.3 Results analysis

Our method resulted in finding new relevant index words that are not extracted when using traditional methods based on the frequency of appearance of the word in the text. This was done by the association rule based data mining approach that we have integrated into the approach presented by Daher (2002) and Mansour et al. (2008) as a new contribution to enhance the process of indexing Arabic texts.

The results of Table I show that our approach retrieves 69 percent of relevant index words out of the whole number of words selected; this is the Precision. Also, it retrieves 76 percent of the words manually determined as index words, represented by the Recall.

We compared our results to the results calculated if the as auto-indexer implemented in Daher (2002) is used alone, without taking into consideration the relations between words. We only compared our work with that of Daher (2002) since it targets directly indexing Arabic texts similarly to ours. The recall was equal to 70 percent in the previous work; the

<table>
<thead>
<tr>
<th>Text No.</th>
<th>(N)</th>
<th>(I)</th>
<th>RR</th>
<th>RI</th>
<th>NRR</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>413</td>
<td>71</td>
<td>57</td>
<td>17</td>
<td>14</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>443</td>
<td>90</td>
<td>64</td>
<td>19</td>
<td>26</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>459</td>
<td>84</td>
<td>62</td>
<td>19</td>
<td>22</td>
<td>0.74</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>466</td>
<td>92</td>
<td>58</td>
<td>21</td>
<td>34</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>475</td>
<td>95</td>
<td>73</td>
<td>18</td>
<td>24</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>477</td>
<td>89</td>
<td>69</td>
<td>24</td>
<td>20</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>7</td>
<td>484</td>
<td>103</td>
<td>83</td>
<td>20</td>
<td>20</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>8</td>
<td>488</td>
<td>71</td>
<td>57</td>
<td>20</td>
<td>14</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>9</td>
<td>502</td>
<td>82</td>
<td>66</td>
<td>33</td>
<td>16</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>508</td>
<td>87</td>
<td>68</td>
<td>21</td>
<td>20</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>11</td>
<td>509</td>
<td>75</td>
<td>53</td>
<td>27</td>
<td>22</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>12</td>
<td>519</td>
<td>88</td>
<td>64</td>
<td>27</td>
<td>24</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>13</td>
<td>531</td>
<td>92</td>
<td>76</td>
<td>31</td>
<td>16</td>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>14</td>
<td>584</td>
<td>80</td>
<td>67</td>
<td>34</td>
<td>13</td>
<td>0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>15</td>
<td>592</td>
<td>91</td>
<td>61</td>
<td>36</td>
<td>28</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>16</td>
<td>613</td>
<td>86</td>
<td>71</td>
<td>36</td>
<td>15</td>
<td>0.83</td>
<td>0.66</td>
</tr>
<tr>
<td>17</td>
<td>643</td>
<td>97</td>
<td>63</td>
<td>44</td>
<td>34</td>
<td>0.65</td>
<td>0.59</td>
</tr>
<tr>
<td>18</td>
<td>680</td>
<td>116</td>
<td>87</td>
<td>29</td>
<td>31</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>19</td>
<td>799</td>
<td>95</td>
<td>80</td>
<td>39</td>
<td>15</td>
<td>0.84</td>
<td>0.69</td>
</tr>
<tr>
<td>20</td>
<td>856</td>
<td>118</td>
<td>90</td>
<td>59</td>
<td>28</td>
<td>0.76</td>
<td>0.64</td>
</tr>
<tr>
<td>21</td>
<td>888</td>
<td>130</td>
<td>98</td>
<td>59</td>
<td>32</td>
<td>0.75</td>
<td>0.62</td>
</tr>
<tr>
<td>22</td>
<td>890</td>
<td>120</td>
<td>91</td>
<td>64</td>
<td>29</td>
<td>0.76</td>
<td>0.59</td>
</tr>
<tr>
<td>23</td>
<td>981</td>
<td>155</td>
<td>113</td>
<td>74</td>
<td>42</td>
<td>0.73</td>
<td>0.60</td>
</tr>
<tr>
<td>24</td>
<td>1,157</td>
<td>202</td>
<td>126</td>
<td>66</td>
<td>76</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>25</td>
<td>1,349</td>
<td>183</td>
<td>155</td>
<td>121</td>
<td>28</td>
<td>0.85</td>
<td>0.56</td>
</tr>
</tbody>
</table>

**Table I.** Experimental results

Average 25.72 0.76 0.69

Arabic texts
precision was equal to 67 percent. This means that the integration of the relation of words feature enhanced the recall by 6 percent and the precision by 2 percent.

Also, the NRR value was reduced by approximately 17.35 percent. This means that the number of non-retrieved relevant words was reduced by our method which in itself is a remarkable enhancement. In other words, reducing the NRR value means that our auto-indexer selects a bigger percentage of relevant index words selected manually by a human intervention.

Some of the relevant words that do not appear frequently in the texts or that might not be spread along the text, but are somehow related to the subject of the text are now associated with other relevant words in the texts; and therefore, these words are considered by our approach as relevant.

This was not the case in any of the previous works related to the auto-indexing topic. The number of relevant candidate terms selected is increased; therefore, the user will have a wider variety of words related in a more efficient way to the subject of the text being looked at, which allows for the extraction of the subject of the text more accurately.

7. Conclusion

Auto-indexing mechanisms are very important and can be used in many fields such as information retrieval. We should not forget that the language has a huge impact on the difficulty of developing any auto-indexing software. In our paper, we proposed a solution for the “auto-indexing Arabic texts” problem based on association rule data mining, or in other words, relating terms that frequently appear together in texts of the same category of the text to be indexed using the Apriori algorithm. This will increase the number of relevant words extracted out of the text. Some words that were not considered relevant when using the previous methods (that only consider the frequency of the words and their stems as well as the spread of the word in the text) will be considered relevant by our work. The association and relation of words are considered a new idea introduced to enhance the auto-indexing work. In fact, we believe that our proposed solution can be considered a major step in innovative contribution to any auto-indexing system that deals with Arabic texts.

From this context, we propose to associate new words using association rule data mining methods to the ones extracted based on the frequency of words in a text as well as their spread along the text. However, some enhancements will be done in future work on many levels.

First, we noticed that the stemming algorithm used in the approach that selects indexes of a text based on the frequency and spread of the words used a stop-list that does include numbers. It is useful to study the benefits of considering numbers as indexes of a text.

Second, the Arabic language has a very wide vocabulary and complex grammatical rules. This might lead to missing some words in the stop-list such as “توجه” (i.e. Toward), “يحيى” (i.e. Across), “عبر” (i.e. in which); and therefore, this list should be always revised and updated when used in an automatic indexer to avoid irrelevant words being extracted as indexes.

Third, in the Arabic language we have many grammatical rules that can be integrated into the stemming algorithm to avoid considering words with the same stem word and treat them as different indexes; for example, the words “زادة” (i.e. increase), “ترنيد” (i.e. increase) and “متراددة” (i.e. increasing) or the words “خطط” (i.e. planning) and “خطط” (i.e. plans).

Fourth, our proposed solution concentrated on selecting the words that frequently appear in the text to be indexed and associate these words to new ones appearing in the text using data mining association rules. We also plan on integrating the Thesaurus System into the data mining association rules in the future. Ruminate, for instance, the case presented in Section 5.2 and amend it by integrating the Thesaurus System: considering TEXT1 as the text to be indexed – TERM1 is not an index of TEXT1 based on its count in the text and its spread. However, TERM2 is an index. If TERM1 and TERM2
appear frequently together in texts of the same category of the text to be indexed, then they are related together. We assume that since TERM1 is related to TERM2 that is an index of TEXT1, then TERM1 should be considered a relevant word of the text. However, in our solution, related terms TERM1 and TERM2 should both appear in TEXT1. What we propose in future work is relating TERM1 or any of its synonyms to TERM2 or any of its synonyms. In this way, we can still benefit from the association between TERM1 and TERM2 even if these words do not exist in the text to be indexed. This will maximize the number of relevant words extracted as indexes of the text. In addition, thesaurus systems can be used to link different phrases or words with the same meaning to each other, which potentially improves the indexing mechanism.

In addition, in the future, we also plan to provide a used-case based implementation.

References


Further reading

About the authors
Ramzi A. Haraty is an Associate Professor of Computer Science in the Department of Computer Science and Mathematics at the Lebanese American University in Beirut, Lebanon. He is also the internship coordinator for MEPI’s Tomorrow’s Leader program. He received his BS and MS Degrees in Computer Science from Minnesota State University, Mankato, Minnesota, and his PhD in Computer Science from North Dakota State University, Fargo, North Dakota. His research interests include database management systems, artificial intelligence, and multilevel secure systems engineering. He has well over 110 books, book chapters, and journal and conference paper publications. He supervised over 110 dissertations, theses and capstone projects. He is a member of the Association of Computing Machinery, Institute of Electronics, Information and Communication Engineers, and the International Society for Computers and Their Applications. Ramzi A. Haraty is the corresponding author and can be contacted at: rharaty@lau.edu.lb

Rouba Nasrallah received her MS Degree in computer science from the Lebanese American University, Beirut, Lebanon. Her research interests include Arabic text processing.
Supply chain management scholar’s research impact: moderated mediation analysis

Yucheng Zhang
Southwestern University of Finance and Economics, Chengdu, China

Yenchun Jim Wu
Graduate Institute of Global Business and Strategy, National Taiwan Normal University, Taipei, Taiwan

Mark Goh
The Logistics Institute-Asia Pacific and NUS Business School, Singapore, and

Xinhong Liu
Research Institute of Economics and Management, Southwestern University of Finance and Economics, Chengdu, China

Abstract
Purpose – The purpose of this paper is to draw on social capital theory to develop a model to explain the determinants of a supply chain management scholar’s academic research impact.
Design/methodology/approach – Drawing from a database of 450 supply chain management scholars in different countries collected from ResearchGate and the World Bank, the bootstrapping method was applied on the moderated mediation analysis.
Findings – Analysis of the mediating role of a scholar’s social capital suggests that social capital theory has a strong explanatory power on the relationship between a scholar’s research skill and academic impact. To account for the boundary effect at the country-level, the authors further examine if this mechanism differs by country in the supply chain management research context.
Research limitations/implications – The findings from this study are from a single research area, which limits the generalizability of the study. Although the data are collected from different sources, including ResearchGate and the World Bank, it is cross-sectional in nature. The variables in this model do not have strong causal relationships.
Practical implications – The results suggest that supply chain management scholars can reap the benefits of their social capital. Specifically, scholars can enhance their academic impact by increasing their social capital.
Originality/value – The results provide a reference for supply chain management scholars keen on enhancing their academic research impact. It also provides a reference to explain why country-level differences can influence these scholars.

Keywords Social media, Social capital, Universities, Research, Analysis, Bibliographies

Paper type Research paper

1. Introduction
Scientific scholars usually obtain peer recognition and funding opportunities by sharing their research outcomes. The most popular and common method of knowledge sharing is to share their publications through academic social websites. The literature suggests that social networking sites are better suited for networking and maintaining a professional image in the academic community (Dermentzi et al., 2016). Practically, scholars have been adopting academic social websites professionally for their research endeavors because of the

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 71602163), Young Scientists Fund from the Ministry of Education of Humanities and Social Sciences Project in China (Grant No. 16YJC630171), and the National Social Science Foundation of China (Project No. 12CGL023), and Ministry of Science & Technology, Taiwan (MOST 105-2918-I-003-002).
convenience of forming new connections with their peers (Yu et al., 2016; Kuo et al., 2017). Today, scholars can collaborate, publish, and promote their work online (De Vocht et al., 2017). Wang and Chen (2012) further suggest that when more members with similar interests join an online community, network externalities can form, and interaction ties would occur. Furthermore, these users are more likely to participate regularly because they believe that their friends or reference groups are concerned with their participation (Yu et al., 2016).

To date, ResearchGate is the largest online social platform used in the scholar community for knowledge sharing (Yu et al., 2016; Thelwall and Kousha, 2015). ResearchGate was founded by Madisch in 2008, as a means to transform the way researchers conduct their research (Thelwall and Kousha, 2015). On ResearchGate, researchers can exchange research ideas and share articles freely to facilitate collaboration without additional cost from other researchers elsewhere. To some degree, ResearchGate is the “Facebook” equivalent for scholars. On ResearchGate, users can announce their findings, communicate with others, and keep abreast of the other scholars’ publications. The uniqueness of ResearchGate is that its website not only allows member researchers to publish and share their publications but also provides a platform for the researchers to interact with others informally. ResearchGate combines both bibliometrics and altmetrics to create a general performance measurement for organizations and researchers alike (Yu et al., 2016). Traditional bibliometric, a performance metric, is used for evaluating the number and impact of research publications. In ResearchGate, the RG score, an impact point, represents a scholar’s academic impact.

In ResearchGate, supply chain management is a large research group which is rapidly growing in number and quality. Given that international trade has increased rapidly in recent decades, more supply chain management scholars are directing their research to explain how commercial supply chains manage their supply chains efficiently to facilitate trade. This increase is accompanied by the growth in human and social capital in the logistics industry (Hartmann and Herb, 2014). Thus, the study on supply chain management should follow apace with economic development (Ni et al., 2016; Shah and Brueckner, 2012; Lee, 2015). In particular, the study of the global supply chain management scholar’s academic impact can lend an international perspective to supply chain management and can help the community to have a better understanding of this scientific area (Rao et al., 2013).

Overall, this study makes three contributions. First, this research extends previous studies that focused on the scholar’s academic impact. While research has found that the scholar’s research ability, such as research skills, can influence their research performance including academic impact (Dakhli and De Clercq, 2004), we lack sufficient knowledge to explain this. Therefore, Hong and Zhao (2016) introduce an information communication mechanism to explain how network composition influences scientific performance. Liu et al. (2017) use meta-analysis to explore the relationship between social network site use and academic performance. However, these research works overlooked the underlying mechanism. As such, our paper explores the underlying mechanism of the relationship between a scholar’s research skill and the corresponding academic impact by testing the mediation effect of the scholar’s social capital. Second, this study examines the boundary effect of this mediation process. Specifically, the investigation of the economic development as the country-level moderator facilitates an understanding of the country-level variation. Third, multiple data sets were included in our analysis, including ResearchGate and the World Bank database, thus providing a rich and robust empirical foundation for investigation. For example, while Yu et al. (2016) selected 300 supply chain management ResearchGate members, we expanded the database and collected data of 450 supply chain management scholars from universities in various countries with country-level data from the World Bank.

In this paper, we first introduce five hypotheses posited from the theoretical relationships amongst research skill, social capital, and a country’s economic development. Next, we test these hypothesized relationships and analyze whether the hypotheses are supported.
We focus this assessment initially on factors that could determine the research scholar’s academic impact. Pre-supposing a moderated mediation effect, we integrate a moderated regression analysis with path analysis to comprehensively analyze this simultaneous moderation and mediation.

2. Theory and hypotheses development
Several constructs and theoretical lenses can apply to this study. In what follows, we will highlight the key constructs and theory applicable to our work.

2.1 Social capital theory
Social capital theory has been receiving attention in the literature and has been studied at multiple levels, including the individual (Raider and Burt, 1996), organizational (Nahapiet and Ghoshal, 1998), and societal (Dasgupta and Serageldin, 2001). The term social capital has multiple definitions, interpretations, and uses (Di Ciommo et al., 2014; Johnson et al., 2013). In this study, we define social capital as "the collective value of all social networks and the inclinations that arise from these networks to do things for each other. Social capital, in this view, emphasizes specific benefits from the trust, reciprocity, information, and cooperation associated with social networks. Researchers have positioned social capital as a key factor to understanding knowledge creation" (Nahapiet and Ghoshal, 1998), which directly influenced the academic impact of researchers. Individuals are able to access and leverage resources embedded in social networks through interactions (Sharmeen et al., 2014). Researchers who are situated closer to the core of their social network may benefit professionally (Tomás-Miquel et al., 2016).

The strength of social relations indicates how well an individual knows his/her exchange partners. However, explicit knowledge is modifiable; it can be easily transferred from one person to another, frequently without interpersonal interaction. Scholars can easily find and follow other scholars they are interested in, and read and download these scholar’s publications through ResearchGate. Moreover, collaboration with colleagues from different research groups can help researchers to increase their citation counts (Li et al., 2013).

Scholars have applied social capital in the theoretical and empirical studies of a scholar’s academic performance (Gonzalez-Brambila, 2014). For example, McFadyen and Cannella (2004) found that social capital, such as the opportunity of interactions among scholars, have positive relationships with knowledge creation and academic performance. The size and composition of a scientist’s social network (Hong and Zhao, 2016), and the value of social capital (Oranye et al., 2017) can have a significant effect on their scholarly performance. Abbasi et al. (2011), and Evans et al. (2016) have reported that scholars who are better connected to and participated actively in such academic communities achieve more scholastically.

2.2 Research skill and social capital
In the modern academic environ, researchers are more likely to cooperate with each other. International cooperation makes their study more convincing and global. Researchers are more likely to cooperate with other scholars who possess different research skills. Collaboration can deliver greater diversity of resources, skills, knowledge, insights, and perspectives. Empirical studies have shown that scholars tend to collaborate with those who can help them in terms of methodology or theory building. Seibert et al. (2017) have shown that co-authoring is positively correlated to the number of publications in the highest-quality journals for a researcher. For instance, Cantor et al. (2010) have examined 3,116 articles published in seven logistics journals from 1987 to 2007, and have concluded that scholars in the logistics domain value co-authored research. Their results show that co-authorship and the frequency of citations are significantly related. This leads us to posit that cross-disciplinary and international collaboration are conduits for providing knowledge, and
they facilitate an ongoing knowledge exchange for developing a better understanding of the studied area. Research skill is a scholar’s human capital, which refers to the stock the scholar possesses in terms of knowledge and abilities that can help to facilitate the flow of change in scholarly pursuit and scientific progress. Such capital is valuable for a specific research area or method (Dakhli and De Clercq, 2004).

Research skills can be developed through formal training and education aimed at updating and renewing one’s capabilities in order to do well in research collaboration (Dakhli and De Clercq, 2004). As such, research skills can improve one’s ability to possess better social capital in research. Dasgupta and Serageldin (2001), in their review of social capital, concurred with Coleman (1990) and emphasized the importance of research skills for developing social capital. On ResearchGate, some evidence implicitly suggests that supply chain management scholars who have more research skills may have more followers. For example, a supply chain management scholar from ResearchGate, scholar A, has 27 research skills. There are 57 scholars following his personal page. In contrast, scholar B who has 5 skills has only 7 followers. This example suggests that a scholar with more skills may have more followers. In other words, a scholar with more skills may possess greater social capital. Thus, we posit the following hypothesis:

**H1.** A scholar’s research skill is positively related to the scholar’s social capital.

### 2.3 Social capital and academic impact

Scholars, with high social capital have higher academic impact than those with few followers. Amjad *et al.* (2016) found that the number of co-authors and the work of their co-authors who have high academic impact have great impact on the ranking of the author. McFadyen and Cannella (2004) found that when individuals seek partners to exchange information and knowledge, these individuals will be influenced by their partners. The number of followers, as a proxy of social capital, explains how a scholar’s skill influences the scholar’s academic impact. On ResearchGate, scholars who have a large number of followers have a higher academic impact than those with lower social capital. This is because supply chain management scholars with high social capital will influence other scholars who view their study by following their personal pages. These followers can know the recent improvements including the new publications and topics from the scholars they follow. This suggests that scholars who have more followers will have greater academic impact. Hence, we posit the following hypothesis:

**H2.** A scholar’s social capital is positively related to the scholar’s academic impact.

### 2.4 Mediating role of social capital in research skill and academic impact

The above hypotheses suggest that a scholar’s research skill has an influence on the scholar’s social capital which may in turn have a positive effect on the scholar’s academic impact. Therefore, the scholar’s academic impact is directly influenced by the scholar’s research skill through social capital. Research skill is the human capital of a researcher. Kilkenny *et al.* (1999) propose a human capital model of performance, which suggests that high performance results from human capital. In addition, they found that all measures of social capital exert an effect on individual performance. Moreover, Boxman *et al.* (1991) report that there is a direct effect of social capital on individual performance. Social capital adds to, rather than replace, human capital in explaining individual performance. Based on the above argument, we propose a mediating model to explain the relationship amongst research skill, social capital, and academic impact. Thus, we propose the next hypothesis:

**H3.** A scholar’s social capital mediates the relationship between the scholar’s research skill and academic impact.
2.5 National economic development as moderator

The efficiency of the research and higher education system in a country is dependent on the economic development level of that country (Varsakelis, 2006). Countries with high economic development usually invest more resources on research and higher education than countries with low economic development. According to Abbott and Doucouliagos (2003), in most developed countries with a high efficiency research system and higher education system, the universities there are funded largely by public sector bodies, such as the Department of Employment, Education and Training in Australia. In a developed country, the public universities are autonomous bodies established under legislation. A high efficiency research system provides scholars with more opportunities to share, follow, and discuss their studies with others. In such a high efficiency system, the results of a study from a scholar can be better diffused. This system provides scholars with a platform to expand their studies and propagate their results faster and wider than a low efficiency research system. Thus, a research system with high efficiency can magnify and expand the scholar’s academic impact. However, it cannot be concluded that there is no scope for improvement in the efficiency of these systems, but rather, the efficiency of a research system in countries with strong economic development level may be higher than countries with weak economic development. In other words, for a supply chain management scholar, the impact of his or her study can vary depending on the home country’s state of economic development. Figure 1 summarizes the theoretical model of this study. To explain the relationships amongst country economic development, supply chain management scholar’s research skill, and the corresponding academic impact, we propose the following hypotheses:

\[ H4. \text{ A country’s economic development moderates the relationship between a scholar’s research skill and the academic impact.} \]

\[ H5. \text{ A country’s economic development moderates the mediation effect of a scholar’s academic impact between a scholar’s research skill and academic impact.} \]

3. Method

3.1 Research context

Scholars tend to turn to the internet nowadays to communicate and disseminate their research rather than to read the journal articles and to attend conferences (Yu et al., 2016; Dermentzi et al., 2016). Individual and organizational websites are created by researchers and academic organizations to provide academic information. Social websites such as Google Scholar, Microsoft Academic Search (MAS), Mendeley, LinkedIn, Academia.edu, and ResearchGate are social platforms that allow researchers to share their publications. Thelwall and Kousha (2015), by using data from the profile pages of Academia.edu, found that the researcher’s academic impact results from both academic and social capital. Their finding suggests that the academic social network sites are hybrid and are genuinely new additions to the academic communication platform (Fu et al., 2014).

![Figure 1. Theoretical model](image-url)
We collect data from ResearchGate, which is built for sharing knowledge among the researchers. Google Scholar had also been considered for the data collection. However, compared to the information on Google Scholar, the data found on ResearchGate are more complete and comprehensive. With 14 million researchers, ResearchGate is large enough to provide massive open access and free of charge information that we need. In our sample, we need not only basic indicators such as impact points, number of publications, number of downloads, and profile views but also indicators about research skills, such as modeling. These data are not available on Google Scholar. However, these indicators can be easily found and coded from the personal profiles of the researchers on ResearchGate. As such, we decided to call on ResearchGate as our data source.

ResearchGate began when two researchers discovered that the collaboration with other researchers from different locations was difficult. The mission of ResearchGate is to connect researchers from all over the world and provide a platform for them to share and exchange academic findings, output and ideas easily. In ResearchGate, housing over 80 million publications, 9 million researchers, and a million answers to research questions, users can find ways to promote their research and expand their academic impact.

According to the ResearchGate website, the platform has five main functions. First, scholars can share publications, have access to millions of publications, and publish their data. Second, scholars can obtain statistics and find out who has been reading and citing their work. Third, organizations can post job information to hire researchers they need on the platform. Fourth, scholars can post research questions, get answers from others, and find solutions to research problems. Finally, scholars can readily connect and cooperate with other colleagues, peers, co-authors, and specialists in their field.

Besides, ResearchGate provides the overall evaluation of a researcher’s academic impact, which account for the quantity of publications, number of questions asked and answered, and the number of followers. Thus, ResearchGate’s score measures academic reputation based on an individual researcher’s contributions and interactions (Thelwall and Kousha, 2015).

Although the literature has provided mixed results on its popularity, activity, authority and accessibility (Nicholas et al., 2016), ResearchGate is changing the way scholars publish their work and gain peer recognition and social capital (Thelwall and Kousha, 2015). Citations indicate recognition and the social capital of the cited author. This recognition is important for academic careers and research evaluations (Moed, 2006). Logically, scholars should apply a wide range of methods to ensure that their publications are accessible and visible to others (Thelwall and Kousha, 2015). Traditionally, scholars seek to publish their papers in high-impact journals or conferences, to advertise them in social media and their classes (Haglund and Olsson, 2008). However, the new norm for scholars is to list their work on online curricula vitae or to list them on online academic social platforms, such as ResearchGate.

Given that the increased visibility benefits of listing publications in multiple sites online outweigh the cost of the time taken to register them, scholars who embrace the new mode of spreading and sharing research online seem to gain greater recognition and social capital. This multiple online availability of research can presumably increase the citation count and thus increase a scholar’s academic impact. Put simply, ResearchGate provides scholars with a new way to gain citations, thus improving their social capital and academic impact.

3.2 Sampling procedure
Supply chain management scholars play an increasingly prominent role in research and academia. This is evident as the supply chain management studies attract greater international academic impact. On ResearchGate, supply chain management scholars are a large academic group, for the number of publications on ResearchGate focusing on supply
chain management study forms a large share and the number of new publications studying supply chain management is increasing rapidly. The total number of followers in 2016 for the supply chain management discipline is 20,123, an increase of 5,673 in just one year.

One of the authors compiled and collected the data from the personal file webpages of the supply chain management scholars on ResearchGate. In our sample selection process, 450 ResearchGate supply chain management members were selected in the supply chain management group. The personal files of the scholars should have the basic information we need such as nationality, followers, and the RG score. Empty profiles and inactive accounts are abandoned by the data selection procedure as such profiles are deemed useless for our study. We submit that the sample is representative of the supply chain management scholars since the scholars chosen are diversified in terms of nationality and institution.

The sample of scholars collected from ResearchGate is from 70 countries and more than 400 academic institutions. While we can study this new online academic platform on a global perspective, having a variety of data can make our results more convincing and objective. Including scholars from different countries and institutions into our sample can support the understanding of the influence of a country’s economic growth on a scholar’s social capital and academic impact, and to identify the relative success of the countries that are increasing or decreasing their scientific success vis-à-vis others. For these reasons, it is instructive to select scholars from different countries.

The scholars in our sample are mainly chosen as follows: 87 scholars from the USA (19.3 percent), 29 scholars from China (6.43 percent), 27 scholars from the UK (5.59 percent), 24 scholars from France (5.32 percent), 23 scholars from Germany (5.1 percent), 21 scholars from Canada (4.65 percent), 18 scholars from Spain (3.99 percent), and 16 scholars from Iran (3.54 percent). The following countries such as Italy, Turkey, India, Greece, Australia, with scholars, form less than 3.5 percent of the sample size. The institutions of the scholars in the sample are discrete. At most four scholars come from the same institution including the Amirkabir University of Technology, Laval University, Universitat Politècnica de València, and the University of Leuven. Each of these institutions accounts for 0.911 percent of the total number of institutions in the sample. Three scholars come from the following seven institutions: Aalto University, École Nationale Supérieure des Mines de Saint-Étienne, Griffith University, McGill University, Shanghai Jiao Tong University, University of Macedonia, and the University of Minnesota Twin Cities. For the rest of the institutions in our sample, 47 institutions have two scholars, and 305 institutions have only one scholar. This indicates our sample include various universities, which is a good mix of supply chain management scholars from around the world. We have included a table in Table AI that contains the key information of this data set for each country.

ResearchGate tries to combine both bibliometrics and altmetrics to provide an inclusive and objective performance measurement for organizations and scholars. Traditional bibliometrics is a performance indicator measuring the amount and impact of publications such as books, articles, and publications. In ResearchGate, impact points directly measure the academic impact of organizations and scholars (Yu et al., 2016). In our sample, we not only collected data of academic impact but also collected indicators such as the number of followers, number of skills, and topics. These indicators can provide us with a comprehensive and objective view to study the social capital and academic impact of the supply chain management scholars on ResearchGate.

3.3 Measures
Our study focuses on the active researcher members in ResearchGate, which refers to those who interact with peer researchers and ask or answer questions through this social medium. We identify researchers in the supply chain management group. There are 450 ResearchGate members with a ResearchGate score greater than 3.0 who were selected from the SCM group.
This score was selected because 90 percent of ResearchGate members have a ResearchGate score of greater than 3.0, which indicates that our sample is representative. The data on research skills, social capital, and the scholar’s academic impact are coded from ResearchGate. Research skills is measured by the number of research skills a scholar has. The more skills a scholar has indicates s/he has various research skills. Social capital is measured by the number of followers a scholar has. A large number of followers indicate that the scholar has high social capital. A scholar’s academic impact is measured by using a scholar’s impact points, which are computed based on the total impact factors of the journal articles that the researcher from a university has authored/co-authored. Country economic development is measured by the Gross Domestic Product (GDP) of the scholar’s country, taken from the World Bank database. Two variables are controlled, namely, the scholar’s number of research topics and country culture. We control for the number of research topics conducted by a scholar as the more topics a scholar focuses on, the more research skills the scholar will have, which may distort the results. Country culture is a control variable because a scholar in a high power distance country may have more followers. People in high power distance countries tend to accept the inequalities in power distribution in organizations (Clugston et al., 2000). When a scholar has higher level research skills, more people may follow his/her website. We measure the power distance culture of a scholar’s country by using the data found from the World Values Survey and hosted on the website maintained by Hofstede (http://geert-hofstede.com/china.html).

3.4 Analysis strategy
For the mediation tests, we followed the recommendations of MacKinnon et al. (2002), and Preacher and Hayes (2004). The bootstrap method was used to avoid the problems associated with the non-Normal distribution of data. For the moderation test, an interaction term was created and its significance tested (Aiken et al., 1991). The moderated mediation or the “conditional indirect effect” (Preacher et al., 2007) indicating the contingent nature of the mediation effect depend on the moderator(s). Following Edwards and Lambert (2007), we test our model by integrating the moderated regression analysis with path analysis to comprehensively analyze the simultaneous moderation and mediation (Hayes, 2012). Figure 2 summarizes the research model. As noted from Figure 2, in the first stage, social capital regressed on a scholar’s research skills, the host country’s economic development,

![Figure 2. Research model](image)

Note: \( e_{sr} \) and \( e_r \) are the error terms of social capital and academic impacts respectively.
and the interactive term of these two variables. The construct, “academic impact,” is regressed on these three factors as well. In addition, social capital and its interactive term with the host country’s economic development are included as predictors for the moderated mediation analysis. Potential multicollinearity was avoided by centering the variables and expressing the interaction terms as a product of the centered scores of the component variables. Control variables were included in all the regression runs unless otherwise specified. A flowchart found in Figure A1 summarizes our analysis.

4. Results
The means, standard deviations, and bivariate correlations among the studied variables are reported in Table I. The results show that a scholar’s research skills are positively correlated with social capital. In addition, social capital is positively correlated with a scholar’s academic impact.

Our statistical results, as shown in Table I, support \( H1 \), namely, a scholar’s research skills is significantly and positively related to the scholar’s social capital \( (B = 0.32, p < 0.001) \). The same applies for \( H2 \) which posits that a scholar’s social capital is positively related to the scholar’s academic impact \( (B = 0.66, p < 0.001) \). For the mediation hypothesis \( (H3) \), the indirect effect is statistically different from zero, as evidenced by a 95 percent bias-corrected bootstrap confidence interval that contains zero \( (B = 0.23; CI [0.10; 0.44]; \alpha = 0.05) \). These results suggest that a mediation effect exists. Therefore, \( H3 \) is supported.

For the moderation analysis \( (H4) \), a host country’s economic development positively moderates the relationship between social capital and a scholar’s academic impact \( (B = 0.26, p < 0.01) \). This interaction is depicted graphically in Figure 3. The effect of social capital on

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scholar’s research skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Social capital</td>
<td></td>
<td></td>
<td>0.44**</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Country economic development</td>
<td>-0.10*</td>
<td>-0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Scholar’s academic impact</td>
<td>0.05</td>
<td>0.53**</td>
<td>0.15**</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Power distance</td>
<td></td>
<td>-0.01</td>
<td>-0.75**</td>
<td>-0.06</td>
</tr>
<tr>
<td>6</td>
<td>Research topics</td>
<td>0.36**</td>
<td>0.34**</td>
<td>-0.02</td>
<td>0.15**</td>
</tr>
</tbody>
</table>

Table I.
Means, standard deviations, and correlations of variables

Note: *, **Significant at 0.01 and 0.05 level (two-tailed), respectively

Figure 3.
Interaction between social capital and country’s economic development on scholar’s academic impact

[Graph showing interaction between social capital and country’s economic development on scholar’s academic impact]
a scholar’s academic impact is high for a country with higher economic development; this effect is low for a country with low economic development. The results for the different groups supporting $H4$ are presented in Table II (Model 2).

$H5$ proposed the moderated mediation effect used to assess the indirect effect of a scholar’s research skill on the scholar’s academic impact (through social capital) at the high and low levels of a country’s economic development. We used linear regression with maximum likelihood estimates and 5,000 data draw for the bootstrapping procedure. The results suggest that the indirect effect of a scholar’s research skill on the scholar’s academic impact (through social capital) between a country with low economic development and a country with high economic development is significant (estimate = 0.10; CI [0.09; 0.24]; $\alpha = 0.05$). Thus, $H5$ is supported. These results are consistent with the results of earlier studies. Figure 4 summarizes the results.

5. Discussion
This paper seeks to fill the gap in explaining the relationship between a supply chain management scholar’s research skill and the corresponding academic impact. In sum, with the data of 450 supply chain management scholars collected from ResearchGate and the World Bank databases, this paper draws the conclusion that factors including research skills, social capital, and a host country’s economic development level can influence the academic impact of supply chain management scholars. Specifically, the results indicated a scholar’s research skills influences the scholar’s social capital ($H1$), which in turn has a positive effect on the scholar’s academic impact. The statistical results are shown in Table II and Figure 4.

### Table II

<table>
<thead>
<tr>
<th>DV</th>
<th>Model 1 B (M(SE))</th>
<th>Model 2 B (M(SE))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.00 (0.04)</td>
<td>0.02 (0.04)</td>
</tr>
<tr>
<td>Power distance</td>
<td>-0.03 (0.06)</td>
<td>0.11 (0.06)</td>
</tr>
<tr>
<td>Research topics</td>
<td>0.20*** (0.05)</td>
<td>-0.02 (0.04)</td>
</tr>
<tr>
<td>Scholar’s research skills</td>
<td>0.32*** (0.05)</td>
<td>-0.16*** (0.05)</td>
</tr>
<tr>
<td>Social capital</td>
<td></td>
<td>0.66*** (0.05)</td>
</tr>
<tr>
<td>Country economic development</td>
<td>-0.03 (0.07)</td>
<td>0.25*** (0.04)</td>
</tr>
<tr>
<td>Social capital × country economic development</td>
<td></td>
<td>0.26*** (0.04)</td>
</tr>
</tbody>
</table>

**Mediation test**

| Indirect effect                                 | 0.23*** (0.09)    |
| Model $R^2$                                      | 0.21***           | 0.45***           |

Note: ***$p < 0.001$
academic impact (H2). In addition, the scholar’s social capital mediates the relationship between a scholar’s research skills and the corresponding academic impact (H3). This mediation effect is stronger for countries with a higher level of economic development than for those with a lower level of economic development (H4 and H5). We hope that this study can shed light on the research about academic impact and can help scholars to extend their academic impact, thus making the study more generalizable in the future.

Based on the data from ResearchGate and the World Bank, we found that social capital theory can explain the relationship between a scholar’s research skills and the corresponding academic impact. In addition, the economic development of a country moderated this relationship and the mediation effect of social capital. With the findings, our study provides some implications for theory and practice, and suggestions for future research.

5.1 Theoretical implications
This research offers two theoretical implications. First, this paper examined how to explain the relationship between a scholar’s research skills and the scholar’s academic impact by testing the mediation effect of a scholar’s social capital. The results indicated that a scholar’s research skills influence the scholar’s social capital, which in turn influences the scholar’s professional performance. A supply chain management scholar with varied research skills has more followers than one with a fewer number of research skills. More followers suggest that the publications will have a higher chance of being read and cited. The ideas and results from the researchers with a higher social capital skill will be more widespread and stand a better chance of being accepted.

Second, the results indicated that the mediation effect is stronger for countries with higher economic development than those with lower economic development. We found that for a supply chain management scholar, s/he has a higher research impact if s/he works in a country with better economic development. This is because the research system in such a country is more efficient. The study outcomes of the scholar can be better diffused, publicized, and recognized. However, in countries with a lower level of economic development, where the research system is not as efficient as the research system in the developed countries, a valuable publication from a scholar may be undervalued and unrewarded. In extreme examples, a war torn country will pay less attention to scholarly pursuits, thus a scholar’s productivity can be easily underestimated. Though the scholar from a country with a lower level of economic development has equal number of research skills and an equal amount of social capital as the peers from countries with a higher level of economic development, his/her academic impact will be lower than his/her peers. Thus, countries that value academic research should focus on developing the economic environment to improve the academic research impact in their countries.

5.2 Practical implications
Our findings may also be of interest to other scholars who want to improve their research impact because our results can inform this accordingly. The results imply that supply chain management scholars can reap the benefits of their social capital. Specifically, the results indicate that scholars can enhance their academic impact by increasing their social capital. Scholars should cooperate with others to increase their academic impact. In this way, scholars can obtain more citations and feedback on their research. The more the other researchers know about these studies, the higher the likelihood that they will apply the findings and translate the results to practice. Such findings may be used in more practical areas and benefit more people.

In addition, the strong correlation between altmetrics and bibliometrics indicates that the researchers who have greater academic impact can usually enjoy greater social impact among researchers who share similar research interests (Yu et al., 2016). Scholars can also
move to countries with better economic development to improve their academic impact. However, this may be difficult. Many factors such as family, religion, culture, and personal preferences deter scholars from moving from their home country to a better host country (Gawlewicz, 2014).

Finally, our study also provides insights for organizations and institutions such as companies and universities to evaluate the academic performance of their researchers for promotion and funding decisions. A researcher’s academic impact can be measured by evaluating their research skills, social capital, and the state of the country’s economy. Based on the evaluations of these indicators which could reflect a researcher’s academic impact, institutions can make appropriate funding decisions to increase the research outcome from an already existing pool of limited funds.

5.3 Limitations and future research
While the insights gained from this study are important for both theory and practice, this study has four limitations which limit the generalizability of the study but at the same time provide potential avenues for future research.

First, the findings from this study are from a single research area. Future research could replicate the theoretical model in the other fields, such as biomedical research and chemical research, and the samples in the other knowledge-intensive industries.

Second, the data collected from ResearchGate and the World Bank are cross-sectional. Future work could collect longitudinal or experimentation data to examine the causal relationships between the scholars’ academic impact and its antecedents. Specifically, future research could adopt a longitudinal or experimental design to test the dynamic process. For example, scholars can use a case-scenario or a lab experiment to examine where a scholar’s research skills could determine the academic impact of the scholar.

Third, this paper mainly studied the effect of social capital at an individual level. Future research can study the impact at a higher level. For example, it is important to examine whether the university’s reputation could influence a scholar’s social capital. Therefore, future studies could collect more university-level data to examine whether social capital is related to a scholar’s academic impact and explore the potential boundary conditions.

Finally, future research can explore other mediating mechanisms and the potential boundary conditions. Besides the processes examined in this paper, we admit that other mechanisms may exist which mediate the relationships among research skill, social capital, and academic impact. These limitations are worth examining in future studies.

6. Conclusions
This paper has applied social capital theory to explain the factors that determine a supply chain management scholar’s academic impact. Data on 450 supply chain management scholars from various universities were collected and analyzed through a path-analysis approach. A moderated mediation analysis was used in studying the effect of social capital theory on a scholar’s research skills and academic level. The results inform that the mediation effect of a scholar’s social capital explains the relationship between a supply chain management scholar’s research skills and academic impact. In addition, the moderation analysis shows that a country-level moderator can influence the academic impact of the scholar. The results thus provide supply chain management scholars with suggestions about how to enhance their academic research impact. While our study only focuses on the supply chain management domain, the scholars from the other domains can also draw inspiration from this study on how to enhance their academic impact.
References


### Appendix 1

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>Social capital</th>
<th>Scholar's academic impact</th>
<th>Research skills</th>
<th>Publication of university per person</th>
<th>University impact per person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>1</td>
<td>102.00</td>
<td>5.84</td>
<td>0.00</td>
<td>2.90</td>
<td>6.53</td>
</tr>
<tr>
<td>Australia</td>
<td>11</td>
<td>48.46</td>
<td>17.58</td>
<td>12.46</td>
<td>2.35</td>
<td>6.52</td>
</tr>
<tr>
<td>Austria</td>
<td>3</td>
<td>62.33</td>
<td>27.07</td>
<td>5.00</td>
<td>1.00</td>
<td>2.25</td>
</tr>
<tr>
<td>Belgium</td>
<td>6</td>
<td>52.67</td>
<td>32.80</td>
<td>7.33</td>
<td>4.38</td>
<td>14.52</td>
</tr>
<tr>
<td>Brazil</td>
<td>9</td>
<td>45.56</td>
<td>9.76</td>
<td>16.11</td>
<td>4.92</td>
<td>15.42</td>
</tr>
<tr>
<td>Canada</td>
<td>21</td>
<td>41.29</td>
<td>30.45</td>
<td>10.62</td>
<td>5.14</td>
<td>15.72</td>
</tr>
<tr>
<td>China</td>
<td>28</td>
<td>22.86</td>
<td>27.90</td>
<td>4.04</td>
<td>3.50</td>
<td>7.00</td>
</tr>
<tr>
<td>Colombia</td>
<td>4</td>
<td>54.25</td>
<td>11.20</td>
<td>10.75</td>
<td>0.32</td>
<td>0.67</td>
</tr>
<tr>
<td>Finland</td>
<td>10</td>
<td>57.60</td>
<td>23.47</td>
<td>10.60</td>
<td>2.16</td>
<td>0.16</td>
</tr>
<tr>
<td>France</td>
<td>24</td>
<td>77.96</td>
<td>43.02</td>
<td>14.29</td>
<td>1.89</td>
<td>3.60</td>
</tr>
<tr>
<td>Germany</td>
<td>23</td>
<td>51.00</td>
<td>19.06</td>
<td>16.26</td>
<td>2.18</td>
<td>5.16</td>
</tr>
<tr>
<td>Greece</td>
<td>11</td>
<td>74.36</td>
<td>29.88</td>
<td>17.91</td>
<td>1.63</td>
<td>2.57</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>5</td>
<td>60.40</td>
<td>55.93</td>
<td>1.60</td>
<td>4.49</td>
<td>10.82</td>
</tr>
<tr>
<td>India</td>
<td>16</td>
<td>43.94</td>
<td>10.92</td>
<td>16.44</td>
<td>1.75</td>
<td>3.36</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1</td>
<td>6.00</td>
<td>4.21</td>
<td>9.00</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Iran</td>
<td>16</td>
<td>111.19</td>
<td>24.61</td>
<td>18.44</td>
<td>0.87</td>
<td>1.25</td>
</tr>
<tr>
<td>Israel</td>
<td>2</td>
<td>30.00</td>
<td>21.16</td>
<td>3.00</td>
<td>0.88</td>
<td>1.58</td>
</tr>
<tr>
<td>Italy</td>
<td>15</td>
<td>79.73</td>
<td>14.71</td>
<td>12.93</td>
<td>4.96</td>
<td>14.70</td>
</tr>
<tr>
<td>Japan</td>
<td>1</td>
<td>20.00</td>
<td>3.20</td>
<td>5.00</td>
<td>11.13</td>
<td>22.52</td>
</tr>
<tr>
<td>Malaysia</td>
<td>6</td>
<td>78.83</td>
<td>16.12</td>
<td>9.67</td>
<td>1.08</td>
<td>1.75</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>9</td>
<td>41.00</td>
<td>15.67</td>
<td>7.56</td>
<td>3.47</td>
<td>11.57</td>
</tr>
<tr>
<td>New Zealand</td>
<td>1</td>
<td>60.00</td>
<td>32.34</td>
<td>0.00</td>
<td>4.27</td>
<td>13.05</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1</td>
<td>5.00</td>
<td>4.78</td>
<td>4.00</td>
<td>1.20</td>
<td>1.78</td>
</tr>
<tr>
<td>Norway</td>
<td>3</td>
<td>19.00</td>
<td>24.81</td>
<td>15.00</td>
<td>1.04</td>
<td>2.57</td>
</tr>
<tr>
<td>Pakistan</td>
<td>1</td>
<td>15.00</td>
<td>5.48</td>
<td>0.00</td>
<td>0.82</td>
<td>0.26</td>
</tr>
<tr>
<td>Peru</td>
<td>1</td>
<td>596.00</td>
<td>24.42</td>
<td>95.00</td>
<td>0.11</td>
<td>0.25</td>
</tr>
<tr>
<td>Poland</td>
<td>2</td>
<td>89.50</td>
<td>12.42</td>
<td>39.50</td>
<td>1.36</td>
<td>2.13</td>
</tr>
<tr>
<td>Portugal</td>
<td>4</td>
<td>37.75</td>
<td>7.51</td>
<td>21.50</td>
<td>2.66</td>
<td>5.14</td>
</tr>
<tr>
<td>Singapore</td>
<td>2</td>
<td>7.50</td>
<td>6.53</td>
<td>5.00</td>
<td>3.46</td>
<td>7.85</td>
</tr>
<tr>
<td>South Africa</td>
<td>1</td>
<td>71.00</td>
<td>4.87</td>
<td>20.00</td>
<td>2.01</td>
<td>4.03</td>
</tr>
<tr>
<td>South Korea</td>
<td>8</td>
<td>27.75</td>
<td>20.64</td>
<td>7.13</td>
<td>4.98</td>
<td>10.75</td>
</tr>
<tr>
<td>Spain</td>
<td>18</td>
<td>50.00</td>
<td>19.48</td>
<td>7.89</td>
<td>2.69</td>
<td>6.54</td>
</tr>
<tr>
<td>Sweden</td>
<td>5</td>
<td>19.00</td>
<td>9.00</td>
<td>18.00</td>
<td>4.78</td>
<td>13.95</td>
</tr>
<tr>
<td>Switzerland</td>
<td>6</td>
<td>73.33</td>
<td>13.25</td>
<td>11.00</td>
<td>1.25</td>
<td>3.86</td>
</tr>
<tr>
<td>Taiwan</td>
<td>15</td>
<td>21.93</td>
<td>24.96</td>
<td>7.13</td>
<td>5.84</td>
<td>10.55</td>
</tr>
<tr>
<td>Turkey</td>
<td>16</td>
<td>58.25</td>
<td>16.81</td>
<td>7.44</td>
<td>1.74</td>
<td>3.24</td>
</tr>
<tr>
<td>UK</td>
<td>27</td>
<td>58.82</td>
<td>29.16</td>
<td>10.41</td>
<td>3.20</td>
<td>11.13</td>
</tr>
<tr>
<td>USA</td>
<td>87</td>
<td>52.79</td>
<td>41.75</td>
<td>7.64</td>
<td>5.63</td>
<td>19.01</td>
</tr>
</tbody>
</table>

*Table Al.* Key information of this data set for each country
Appendix 2

**Figure A1.** Flowchart of procedure of analysis

Appendix 3

**Figure A2.** Breakdown of scholars by country
Appendix 4

Corresponding author
Yenchun Jim Wu can be contacted at: wuyenchun@gmail.com

Figure A3.
Top 20 institutions