Scientometrics

An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

AKADÉMIAI KIADÓ

Editors

Thelwall*; Bart Thijs; Benno Torgler; Peter van den Besselaar; Thed N. van Leeuwen; Anthony F.J. Van Raan*; Péter Vinkler*; Ludo Waltman*; Howard D. White*; Erjia Yan; Ping Zhou; Michel Zitt*

**Price Medal Laureate*

Journal authors

Publishing ethics

Open Access & Springer

Discover content

SpringerLink

Books A-Z

Journals A-Z

Video

Other services

Instructors

Librarians (Springer Nature)

Societies and Publishing Partners

Advertisers

Shop on Springer.com

About Springer

About us

Help & Support

Contact us

Press releases

Impressum

Legal

General term & conditions

California Privacy Statement

Rights & permissions

Privacy

How we use cookies

Manage cookies/Do not sell my data

Source details

Scientometrics

Scopus coverage years: from 1978 to Present

Publisher: Springer Nature

ISSN: 0138-9130 E-ISSN: 1588-2861

Subject area: Social Sciences: General Social Sciences | Social Sciences: Library and Information Sciences | Computer Science: Computer Science Applications

Source type: Journal

CiteScore 2020
5.2

SJR 2020
0.999

SNIP 2020
1.565

View all documents > | Set document alert | Save to source list | Source Homepage

CiteScore | CiteScore rank & trend | Scopus content coverage

CiteScore 2020

$$5.2 = \frac{7{,}079 \text{ Citations } 2017 - 2020}{1{,}349 \text{ Documents } 2017 - 2020}$$

Calculated on 05 May, 2021

CiteScoreTracker 2021

$$5.5 = \frac{7{,}660 \text{ Citations to date}}{1{,}394 \text{ Documents to date}}$$

Last updated on 06 March, 2022 • Updated monthly

CiteScore rank 2020

Category	Rank	Percentile
Social Sciences — Library and Information Sciences	#21/235	91st
Computer Science — Computer Science Applications	#166/693	76th

View CiteScore methodology > | CiteScore FAQ > | Add CiteScore to your site

2.13. What are quartiles?

Quartiles are bands of serial titles that have been grouped together because they occupy a similar position within their subject categories. The quartiles are:

- Quartile 1: serial titles in 99-75th percentiles
- Quartile 2: serial titles in 74-50th percentiles
- Quartile 3: serial titles in 49-25th percentiles
- Quartile 4: serial titles in 24-0th percentiles

A title might have a different quartile within each different subject area it is included in. For example, Serial Title A might be categorized in “Oncology”, with a CiteScore percentile of 84%, and “Cancer Research”, with a CiteScore percentile of 73%; These percentiles equate to Quartile 1 and Quartile 2 respectively.

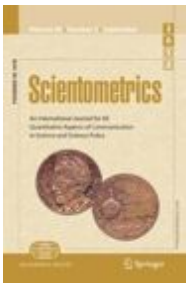
The Top 10% of serial titles can also be viewed.

Skip to main content

Scientometrics
All Volumes & Issues
ISSN: 0138-9130 (Print) 1588-2861 (Online)

In this issue (26 articles)

Skip to main content



Scientometrics

An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

AKADÉMIAI KIADÓ

Editors

EDITOR-IN-CHIEF
Wolfgang Glänzel, Wolfgang.Glanzel@kuleuven.be
Price Medal Laureate

HONORARY EDITOR-IN-CHIEF AND FOUNDER:
Tibor Braun, Hungary
Email: braun@mail.iif.hu
Price Medal Laureate

For authors

Submission guidelines Ethics & disclosures Contact the journal Submit manuscript

Explore

Online first articles Volumes and issues

Scientometrics (2020) 122:1407–1443
https://doi.org/10.1007/s11192-019-03342-2

Identifying collaboration dynamics of bipartite author-topic networks with the influences of interest changes

Diana Purwitasari[1,2] · Chastine Fatichah[2] · Surya Sumpeno[1,3] · Christian Steglich[4] · Mauridhi Hery Purnomo[1,3]

Received: 18 June 2019 / Published online: 14 January 2020

Abstract
Knowing driving factors and understanding researcher behaviors from the dynamics of collaborations over time offer some insights, i.e. help funding agencies in designing research grant policies. We present longitudinal network analysis on the observed collaborations through co-authorship over 15 years. Since co-authors possibly influence researchers to have interest changes, by focusing on researchers who could become the influencer, we propose a stochastic actor-oriented model of bipartite (two-mode) author-topic networks from article metadata. Information of scientific fields or topics of article contents, which could represent the interests of researchers, are often unavailable in the metadata. Topic absence issue differentiates this work with other studies on collaboration dynamics from article metadata of title-abstract and author properties. Therefore, our works also include procedures to extract and map clustered keywords as topic substitution of research interests. Then, the next step is to generate panel-waves of co-author networks and bipartite author-topic networks for the longitudinal analysis. The proposed model is used to find the driving factors of co-authoring collaboration with the focus on researcher behaviors in interest changes. This paper investigates the dynamics in an academic social network setting using selected metadata of publicly-available crawled articles in interrelated domains of "natural language processing" and "information extraction". Based on the evidence of network evolution, researchers have a conformed tendency to co-author behaviors in publishing articles and exploring topics. Our results indicate the processes of selection and influence in forming co-author ties contribute some levels of social pressure to researchers. Our findings also discussed on how the co-author pressure accelerates the changes of interests and behaviors of the researchers.

Keywords Longitudinal network analysis · Scientific collaboration dynamics · Research interest changes · One mode co-author network · Bipartite (two-mode) author-topic network · Stochastic actor-oriented model

Mathematics Subject Classification 68T30 · 68U15 · 90B15 · 91B16 · 91C20 · 91D30

✉ Diana Purwitasari
diana@if.its.ac.id

Extended author information available on the last page of the article

JEL Classification C31 · C38 · C44 · D80 · D85

Introduction

Collaboration studies on bibliographic data of scientific publications (Jung 2015; Xia et al. 2017) provide services like impact evaluation on articles and authors (Hirsch 2005) or recommendations for searching experts (Lin et al. 2017) along with scientific articles (Beel et al. 2016). Collaborations through co-authoring with experienced or international partners can become one of the factors that help in career advancement (Iglič et al. 2017). It is not surprising to have collaborations across disciplines (Iefremova et al. 2018; Purwitasari et al. 2017) which on longer period can cause topic drift or changes of researchers' interest (Amjad et al. 2018). Therefore, researchers often have options to give more attention to their current topics, create variations, or entirely change their interests through the collaborations (Siciliano et al. 2018).

The performance of collaborations is often symbolized by stronger tie formation between researchers such as more co-authoring number in relations of supervisors and doctoral students (Shibayama 2019). Joint publications as the results of co-authorships are often used to capture social interactions of researchers. For that purpose, the structure of co-author networks generated from the publications is preferred instead of citation networks (Fu et al. 2014). As co-authorship evolves over time, tie formations and changes in researchers' interest show the dynamics of co-author networks in longitudinal analysis. The interest changes over time is also accounted for measuring publication productivity (Amjad et al. 2018) or finding researchers with temporal rank differentiated by publication venues (Daud et al. 2010) since some venues have higher ratings (Meho 2019). In preceding studies on collaboration dynamics overtime, network properties like centrality or clustering coefficient were used to explain tie formations of researcher cliques (Abbasi et al. 2011; Hou et al. 2013). Moreover, the centrality of bipartite or two mode author-topic networks became indicators to determine core authors (Abrahams et al. 2019). They applied the bipartite networks to link researchers and research areas of published articles in a cross-sectional study. The research areas or topics were manually identified. Although the studies showed that core authors have roles in expanding research interest, the expansion extent of research interest particularly to the collaboration dynamics has not been discussed.

Recent works on finding the collaboration contributing factors acknowledged preferences to researchers' cliques, associated organizational units or same interest of scientific fields (Ferligoj et al. 2015) along with researchers' experience (Iglič et al. 2017). As researchers adapt to social environments forming by collaborations through co-authoring, their curiosities are possibly adjusted to the topic interest of the co-authors. The assumptions have yet to be empirically tested in the academic social networks. In a real-life academic data, some co-authored articles are not complemented with the information of scientific fields. Therefore, definite interests of the researchers are not clearly stated. Studies seeking mechanisms to explain network evolution with attention to the expansion of research interest and the absence of scientific fields remain scarce. Thus, to understand the contributing factors for those mechanisms, evolution analysis on the dynamics of co-author networks is essential and crucial.

We propose a stochastic agent-based model from researcher perspective on bipartite (two-mode) author-topic networks to disclose the premises of collaboration preferences on the influences of interest changes. The model includes longitudinal data analysis in

an elaborate empirical setting to get a better understanding of the relations of researchers, their interest, and time influences. The model is used to investigate to what degree researcher relations through co-authorship accelerates their behaviors, besides whether their topic interest affects them. Our experiments use a carefully chosen standard dataset of a real-life academic social network without predefined scientific fields. Therefore, our contributions in designing models to analyze the collaboration dynamics without information of scientific fields are summarized as follows:

- To get relations of researchers and topics from article metadata, the first phase includes preprocessing steps such as clustering article keywords to obtain the topics and then mapping them to articles for inferring research interest of the researchers. The preprocessing steps are parts of data preparation.
- The second phase is variable extraction from bipartite author-topic networks to get sociomatrices of the researchers, their individual covariates, as well as their behaviors on interest changes over time.
- Finally, we specify models using various potential changes of network structural for comprehensively investigating the evolution of bipartite networks. Then, in-depth analysis of covariates and behaviors of co-authors to support our hypotheses conclude the empirical experiments. The hypotheses are about collaboration contributing factors and the influence of co-authors in expanding research interest.

Related works on collaboration dynamics in academic networks

The proposed model using co-author and author-topic networks aims to obtain insight of driving factors in collaborations with a concern on co-author influence in research interest expansion. The following reviews address some potential factors, i.e. node or edge, in the formation of collaboration ties from the network perspective of co-author and author-topic.

Influencing factors for tie formation in collaboration networks

For understanding collaboration dynamics in co-author networks, some studies started the observations since the initiation of the first node in co-author networks (Abbasi et al. 2011). The networks were formed from selected publications having a particular keyword exists in either titles, abstracts or keywords. The studies found that the affiliation of countries and institutions could influence the collaboration growth. Extending to several associated keywords but with more controlled affiliation, other studies demonstrated comparable findings in Chinese communities (Hou et al. 2013) and South African institutions (Abrahams et al. 2019). Without assessing the author property of affiliation, collaboration ties in those studies were controlled by authors with higher centrality node degrees from the network analysis. The tied authors shared keywords as research topic substitution, and made the topic as node property became the influencing factor instead of affiliation.

Different studies observed manually collected data of node properties such as gender, academic position, and research discipline rather than using publication metadata (Siciliano et al. 2018). Investigation on those author properties fixated on the interest change or network churn through researcher behavior for either exploration or exploitation on topics. Collecting data manually for co-author networks was applied as well from questionnaires to generate a network of knowledge and social ties between researchers, defined as

self-reported collaborations (Bozeman and Corley 2004). The studies observed different aspects of collaboration dynamics from the policymaker setting to consider strategies in designing research grants. Questionnaires usage avoided deceptive conclusions from the practice of making colleagues honorary co-authors. Nevertheless, institutions such as the governments observe research outputs through publications to create further initiatives for advancing national innovation. Thus, a publication-based measure using co-author networks is still in favor, even though co-authorship reveals a partial indicator of collaboration. Questionnaires were also used to investigate the benefit of author relationships in mentorship in the context of tie formation (Shibayama 2019).

Above-mentioned works confirmed that either the node factor of author properties (i.e. affiliation, research topics) or the edge factor of author relation (i.e. mentorship) have roles in the network growth. Those works examined the collaboration dynamics with pre-defined keywords (Abbasi et al. 2011; Hou et al. 2013) (Abrahams et al. 2019) or manual validation requirement (Siciliano et al. 2018). Current works delimit the scientific domains instead of keywords from existing publication metadata. Even though node factor is still anticipated to influence the tie formation, but the properties of authors originate from the metadata to avoid the manually collected process. Aside of those differences, our studies emphasize to identify influencing factors of tie formation without the properties of authors, such as situations of researcher expertise scarceness (Zhu et al. 2017; de Siqueira et al. 2018) while considering interest change perspective in shaping the network dynamics.

Bipartite exploration on author-based networks

Since prominent keywords, which could be treated as topics in a broader context, present significant concepts of academic articles, same keywords often co-occur within texts on similar themes or areas. Due to the need for tracking trends in academic articles, one-mode network of articles or keywords derived from two-mode network of article-keyword was constructed to understand key points of articles (Li et al. 2016). The article-keyword network was also used in particular network type of author-article-keyword to quantify author relationships for citation analysis (Wang et al. 2018). By incorporating keyword information, the analysis revealed more insights of knowledge domain mapping for authors.

An information retrieval application for searching scientific articles also took advantage on author-keyword networks to semantically represent knowledge domain (Wen and Junping 2014). The networks were constructed by processing records retrieved from a library catalogue of certain university on one selected domain. With manageable number of keywords since the catalogue records had been well organized, the network size constraint did not occur. Another author-keyword network study was about community detection based on a measure of group cohesion with the legendarily InfoVis 2004 dataset of conference papers (Renoust et al. 2014). InfoVis dataset has been cautiously prepared and the keywords have been carefully prepared by the authors, which leads to an anticipated number of keywords.

In short, the feasibility to construct author-keyword networks depends on the number of keywords which might cause the network size constraint. Although article themes are narrowed down to involve certain domains, this study still includes a broad range of keywords. To conduct the approaches of author-keyword networks in our case for investigating collaboration dynamics would be quite problematic because of the network size constraint. Nevertheless, we agreed that set of keywords or topics provide a concise overview of a body of articles. We expect different approaches by clustering article words as substitution (Kong

et al. 2017) and then topic mapping could represent keywords, so constructing author-topic bipartite networks would have no size constraint.

Research hypotheses

Searching academic experts is one kind of recommendation services (Lin et al. 2017). A recommendation process to an expert is suggested through analyzing bibliographic data of her or his scientific publications. The experts are researchers who have in-depth skills in specified fields. Collaborating with co-authors from different fields makes the researchers more likely to have knowledge sharing through interaction and discussion. Then, mutual confidence and shared behavior may be accounted for repeating co-authorships. Later, working repetitively on multi-disciplines might lead to situations with the likelihood of interest changes or widening topic variation of researchers.

We propose a framework for expert recommendation system with regard to the influences of interest changes, as illustrated in Fig. 1. Centered by scholar profiles with interest evolution and fairness factor, the framework contains three modules to process publication metadata of scholars into recommendations, i.e. for finding mentors or references. Basically, Synchronize Module preprocesses researcher metadata such as title, abstract, or author information of published scientific articles, and then Expertise Module generates the preprocessed results into researcher profiles. Finally, Expert Finder Module utilizes the profiles into recommendations.

Single-layered network or also known as one-mode networks such as co-author networks are widely-used in abstracting bibliographic data (Abbasi et al. 2011), although recent approaches tend to utilize multi-layered network (Abrahams et al. 2019). Multi-layered in the proposed framework is defined as multi perspectives of bibliographic data which is one-mode networks of co-author relations and two-mode (bipartite) networks of author-topic relations. The proposed framework considers topic information as part of up to date expertise in researcher profiles. Consequently, the premises of collaboration preferences between authors regarding to co-author influences on interest changes need to be disclosed. Therefore, this paper focuses to analyze important aspects

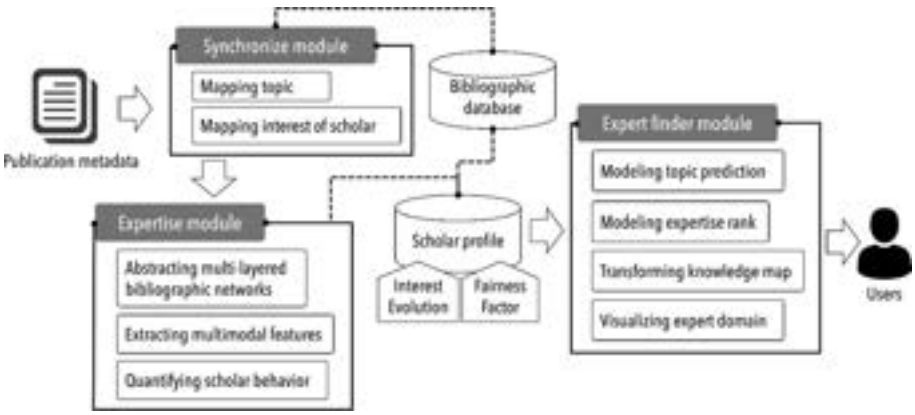


Fig. 1 Proposed framework for expert recommendation system regarding to the influences of interest changes

in "Abstracting multi-layered bibliographic networks" process to construct profiles of researchers or called as Scholar Profile by Expertise Module. Scholar Profile considers the changes of research interest as one of evolution factors that becomes our issues. Scholar Profile also considers fairness factor to control overly self-citation practices and manage unbiased measures for in-depth researcher skills (Purwitasari et al. 2018b). Finally, researcher features extracted in Expertise Module with the perspectives of interest evolution and fairness factor are applied for recommending researchers with high academic level who are productive and have up to date expertise.

Tie formation becomes an important aspect in "Abstracting multi-layered bibliographic networks" process. Studies about ties of co-author networks from Slovenian national scientific system verified small-world structure and displayed preferential attachment evidence with a clustering level in the researchers' cliques (Ferligoj et al. 2015). The small-world structure means that most nodes in co-author networks have several steps to every other node, defined as a network property of transitive closure (Ebadi and Schiffauerova 2015). The Slovenian libraries officially maintained scientific disciplines and their subcategories of scientific fields to label the researchers. Label information was used to consider preferential attachment in the network studies. However, the label information can be unavailable because of the sparsity problem in a cold start recommendation application with actor networks generated from movie metadata (Zhu et al. 2017). The actor networks are comparable to our problem with co-author networks from scientific publications. This paper differs from prior studies of Slovenian co-author networks by the absence of predefined topics assigned to the researchers. To consider the topic issue, in this study we obtain the topic information by automatically labeling each scientific article with subject areas or topics. Then, from the mapped articles with topic information, we indirectly assign topic information to researchers who publish the articles. Topic labels are often extracted with a generative model of word distributions from abstract texts of one particular publication venue that spanned over some time (Griffiths and Steyvers 2004). Articles required in constructing co-author networks of certain disciplines can come from various publication venues, but their contents should be in the same related scientific fields. To address the topic issue, unsupervised approach of clustering was preferable to identify coherence subjects in topic extraction instead of a generative model (Suominen and Toivanen 2016; Kong et al. 2017; Purwitasari et al. 2017).

While Slovenian networks solely observed authors, the proposed model in this paper develops bipartite networks with node types of authors and topics identified by title-abstract clustering. Aside from network type difference, we also consider issues on clustering level and interest changes of the researchers as mentioned in the first hypothesis (Fig. 2).

H1 Bipartite author-topic networks based on topic interests demonstrate transitive closure and researcher preferences in forming cliques.

Transitive closure means that in the context of a researcher, called as an ego, co-authors of his or her co-authors, tend to become co-authors of the ego. In the research collaboration process, the equivalent mechanisms of transitive closure are revealed in co-evolution of bipartite (two-mode) author-topic networks. In this study, research interests and career age of experts become the researcher characteristics. Career age of an expert is defined from year value of his or her first publication.

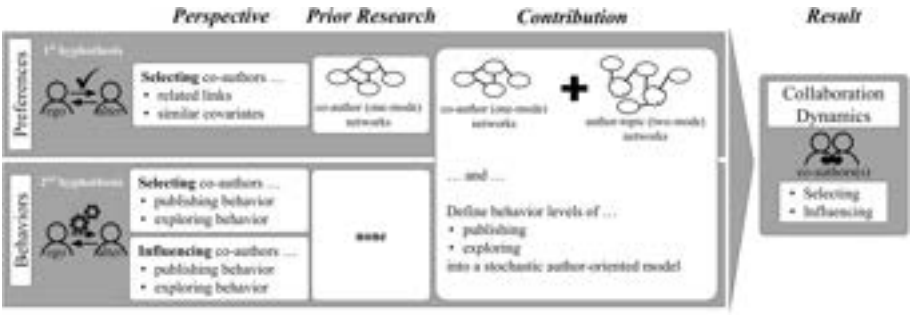


Fig. 2 Proposed model for identifying collaboration dynamics with the influences of interest changes

Researchers can change their interests because of the influences of more experienced co-authors. The consequence of interest changes motivated some studies on expert finder (Amjad et al. 2018; Daud et al. 2010). A model to track individual research profile and quantify outcome during researchers' career was developed to accommodate interest changes (Liang et al. 2016). However, there are scarceness on to what extent the co-author influences are within these studies. Thus, the network structure of researchers and their characteristics are going to be studied as joint dependent variables to avoid bias in the second hypothesis (Fig. 2).

H2 Behavior values from bipartite author-topic networks based on topic interests are associated with experience such that researchers incline to form ties with others in looking for supervision aspect.

The second hypothesis about expertise rank suggests a hierarchy among researchers based on their behavior and experience. Tie formation of co-authors that considers the researcher hierarchy is similar to trade formation among countries based on country income (Manger et al. 2012). Higher expertise of researcher can be translated differently according to the context, such as researcher status as team leader who are often more productive and publish more frequently (Oliveira et al. 2018). Author order in scientific publications could also indicate expertise level (Kosmulski 2012). However, our experiments use open access metadata collected from automatic procedures that may cause missing data. Therefore, the proposed model disregards author orders and replaces with frequent levels in publishing. The experiments investigate the second hypothesis by addressing interdependence between researcher behavior and the network structure of social ties between researchers.

Data preprocessing

This section describes our data and some preliminary works needed on the dataset to provide supplementary information caused by the inexistent of predefined topics assigned to the researchers. The phenomenon of co-authorships between researchers through published articles is presented in Figs. 3 and 4. Then, preprocessing steps from preparing dataset steps to substituting topic steps are displayed in the flowchart of Fig. 5. Preparing dataset includes data sampling, expert selection, and semi validating metadata, whereas substituting topic also covers clustering words of title-abstract as well as topic mapping.

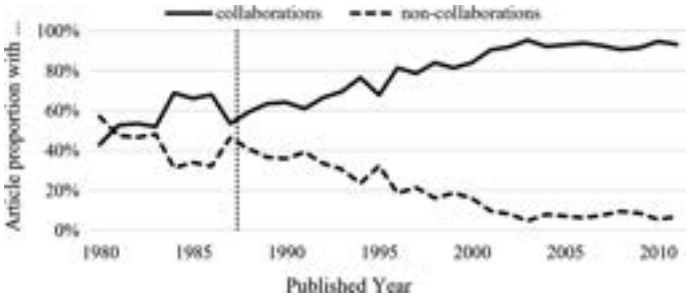


Fig. 3 Selected AMiner articles in NLP-IE domains show an increasing number of collaborations through co-authorship

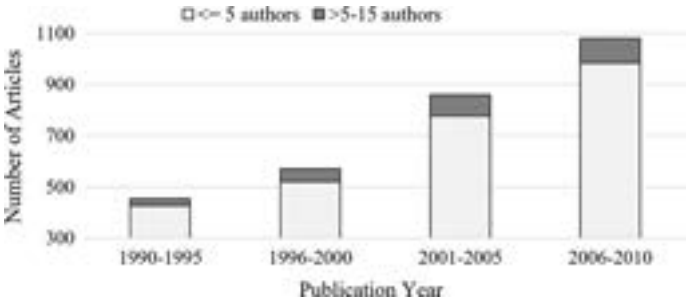


Fig. 4 Increasing number of co-authors in collaborations from selected AMiner articles in NLP-IE domains

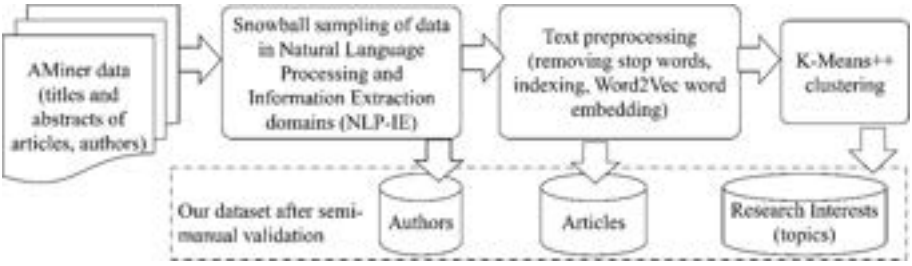


Fig. 5 Preprocessing for AMiner dataset of NLP-IE domains

Descriptions in details about those steps are clarified in the subsequent subsections. All data preparations were performed using Python with primary libraries of Natural Language Toolkit (www.nltk.org) for preprocessing in stop words removal, Scikit-learn (scikit-learn.org) for clustering, in addition to Gensim (radimrehurek.com/gensim) for word embedding and Palmetto for coherence measures of clustered words (palmetto.aksw.org).

Data description

Our experiments work on one of popular and open access scholarly datasets (Xia et al. 2017) containing publication metadata in the fields of Computer Science. The dataset was collected by using web harvesting mechanisms from academic search and mining system, recently called as AMiner and has been discussed in many expert finder related researches (Ortega 2014; Tang 2016; Kong et al. 2019). AMiner data was initially collected and started from an academic search system ArnetMiner as the works of Chinese researchers in extraction and mining of academic social networks (Tang et al. 2007, 2008). AMiner gave profiles on research activities to improve scholarly recommendations (Tang et al. 2010; Datta et al. 2017; Ayaz et al. 2018). Data entities of AMiner database are articles, authors who write articles, citations, and relations between article-author-citation. AMiner domains of Natural Language Processing and Information Extraction (NLP-IE) become our designated testbed because of its sufficiently representable size. Selecting the NLP-IE testbed makes easier semi-manual validation and collaboration pattern examination to avoid bias in the reported results later.

AMiner has defined a benchmark list of authors available for expert finder researches (Deng et al. 2012). Snowball sampling to generate our dataset was initiated from AMiner benchmark list of 70 NLP-IE experts. After that, the data selection process included other authors in AMiner database who at least have seven publications with the initial experts. There was standardization shortcoming in AMiner database because of the harvesting process like duplication, missing abstract text or only listing some of the authors. Semi manual cleaning processes were necessary to be conducted because of web data extraction issues. Then, the following step was to select a number of publications. As a result, our experiments were performed on the final NLP-IE dataset with 212 researchers and ± 4600 articles in total. Those articles were divided into four data groups of 5-year intervals. Period-1 contained any published metadata until 1995 with the earliest year was around 1980, Period-2 for metadata between 1996 and 2000, and accordingly for Period-3 (2001–2005) and Period-4 (2006–2010). Collaboration tendency of those periods is shown in the following figures. Figure 3 displays the decreasing proportion number of articles by single authors (non-collaborations). Starting around 1990, the collaboration number has presented such steep increment in Fig. 3. Therefore, we emphasize the tendency for intensified collaboration number in the period after 1990 through the number of co-author in Fig. 4.

Clustering for topic identification

This paper discusses the dynamics of bipartite author-topic networks which are constructed from scientific co-authored article metadata. The metadata does not have predetermined categories of scientific disciplines and subcategories of scientific fields. The categories can roughly represent the content of articles. We used the term of topics instead of categories of scientific disciplines in this study. Therefore, no pre-determined categories in this study is equivalent to topic absence. Substituting the topic absence, clustering approach was often applied to automatically generate subject areas or topics of the articles which can be replicated for article texts with any Latin alphabet based languages (Purwitasari et al. 2016). The text preprocessing in substitution step followed the standard practices of Word2Vec word embedding (Mikolov et al. 2013) to enhance the cluster results and get more suitable keywords. Based on prior practices,

text preprocessing also removed words with certain conditions (Fig. 5). They are words with three characters or less, stop words, and words that appeared in fewer than ten articles from the texts of NLP-IE dataset. After word indexing process from the articles and removing words with certain conditions, there were 3500 remaining candidate keywords which were only ± 25% representation from the initial words. Previous works suggested 200 dimensions for embedding process (Mikolov et al. 2013). Hence, the subsequent step was to create a 3500 × 200 feature matrix of keywords. It means that each word has 200 values as the relation weights. Then the embedding process took out words with an average value which were less than 0.04. The average value of a word was determined from dividing all its 200 weights evenly.

Comparison between supervised method with human assistance like predefined topics, and unsupervised approach in topic modeling, such as clustering for science mapping, is difficult to argue since each approach has its own superiority (Suominen and Toivanen 2016). However, mutations of research problems leading to the emergence of multidisciplinary research that poses co-author influence in the expansion of research interest, as discussed here, make the clustering approach more fitting in this study. Therefore, KMeans++ was applied to cluster the remaining keywords after removing unnecessary words for topic mapping. The clustering results were evaluated with various K values using Silhouette scores as indicators to measure the cluster goodness as shown in Table 1. KMeans++ provided better initial seeding than state-of-the-art KMeans (Aubaidan et al. 2014; Purwitasari et al. 2011).

By using KMeans++, simulations with cluster number $K = \{100, 50, 30\}$ were carried on to determine the appropriate number. We also used some algorithms for comparison as shown in Table 1. KMeans provides satisfactory results when data within clusters are circularly placed or hyper-sphere in higher dimensions (Hornik et al. 2012). Meanwhile, Gaussian Mixture (GM) supposedly can take data with different shapes. Since articles placed on hyper-dimensions are unlikely in the form of hyper-sphere, we carried on clustering with GM involving the mixture of multiple Gaussian distributions from data features as a comparison. GM simulations applied small number of features extracted with typical methods of Feature Agglomeration or Principal Component Analysis. Although GM simulations had better Silhouette scores, it is unlikely for real texts to have two features. From the KMeans++ results, less cluster number was suggested although the Silhouette scores were not satisfactory with less than 0.5.

Table 1 Various simulations to determine cluster number K (with bold values refer to the best clustering parameters for each K)

K	Silhouette score Avg.	Algorithm	Input matrix	Feature
100	0.135	KMeans++	± 3500 × 100	Document frequency, DF with 100 embedding dimensions for fast results
	0.364	Gaussian mixture	± 3500 × 2	Feature agglomeration (2)
	0.256	Gaussian mixture	± 3500 × 2	Principal component analysis, PCA (2)
50	0.115	KMeans++	± 3500 × 100	Document frequency
30	**0.179**	KMeans++	± 3500 × 100	Document frequency
	0.177	Gaussian mixture	± 3500 × 2	Feature agglomeration (2)
30	0.651	Gaussian mixture	± 600 × 2	Feature agglomeration (2) Notes: only using title texts

Therefore, we cautiously continued on other clustering scenarios which leaded to better results of higher Silhouette score (>0.5) by only using title texts and made a smaller feature matrix (Purwitasari et al. 2018a).

Finally, the value of K was 30 groups of clustered words designated as topics $T = \{t_1 \ldots t_{30}\}$. We checked the coherence of clustered words $\text{coh}(t_k)$ to assure topic interpretability using Wikipedia corpus for topic learning (Röder et al. 2015). Based on Pearson correlation, the four coherence levels are "satisfactory" for $0.28 \leq \text{coh}(\cdot) < 0.3$, "average" for $0.3 \leq \text{coh}(\cdot) < 0.4$, "fairly good" for $0.4 \leq \text{coh}(\cdot) < 0.5$, and "good" for $\text{coh}(\cdot) \geq 0.5$. Higher values indicated better topic interpretability. The clustering results had 5 topics assessed for coherence level of "good", 12 topics for "fairly good", 11 topics for "average", and 2 topics for "satisfactory". It meant that the cluster results can represent topics for substituting scientific fields.

Model specification

The proposed model utilizes entities of researchers as authors and articles as the output of their social relationship organized into an academic social network (Kong et al. 2019). The proposed model contains co-author (one-mode) networks and author-topic (two-mode) networks and their relationship are presented in Fig. 6. The mapping of the entities of author and topic with a case of article-A and article-C mapped into topic T1 is illustrated in Fig. 6. Even though author-*i,* author-*j*, and author-*k* are co-authors of article-A, the case only reveals the relation of two authors to T1 because the similarity value between author-*k* and topic T1 is low. Researcher interests are indirectly established from mapped topics as presented in Fig. 7, which also displays other processes to specify and apply the proposed model for confirming our hypotheses. Model.A accommodates collaboration dynamics while focusing on behaviors of the researchers in interest changes. Model.B and Model.C become the comparison models.

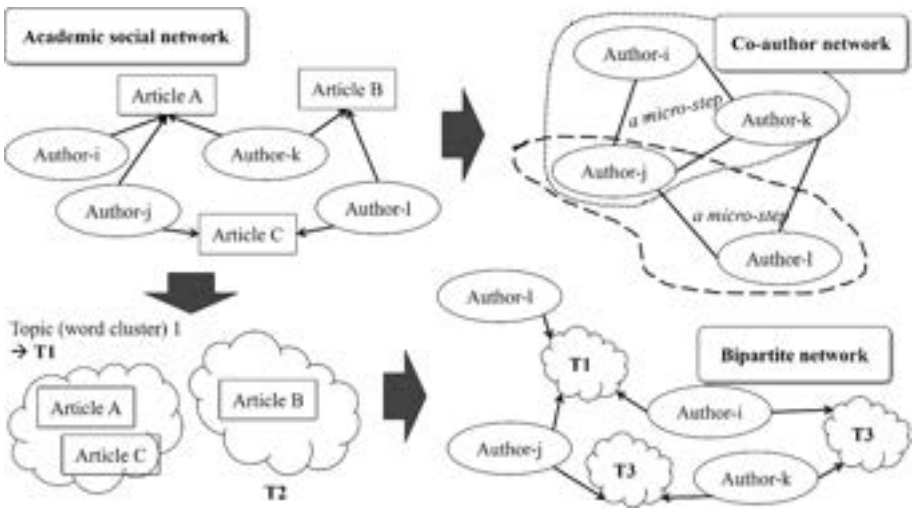


Fig. 6 Academic social networks used in the proposed model specification

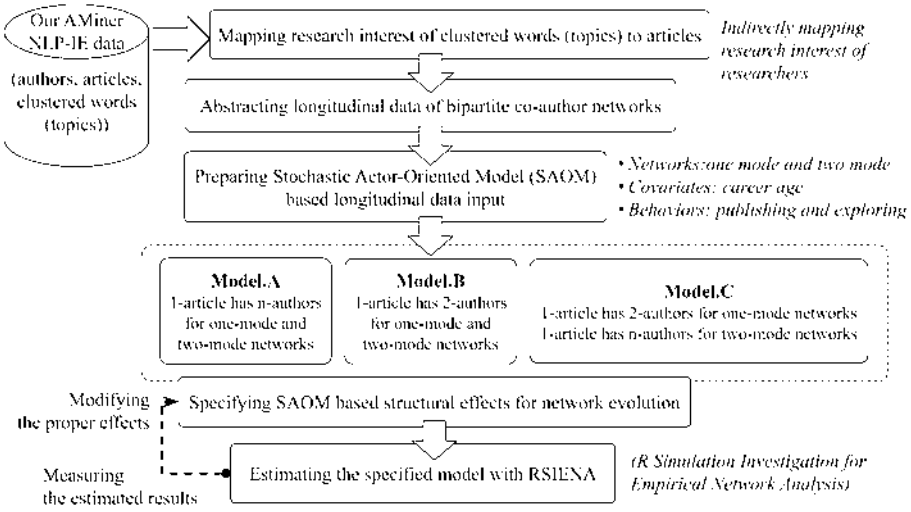


Fig. 7 Steps used to evaluate the proposed hypotheses

The proposed model focuses on the influence of co-authors for interest changes. Therefore, researchers turn to be the focal points which is termed as an actor-oriented model. Thus from the researcher perspective, we selected a prevalent Stochastic Actor-Oriented Model (SAOM) for modeling the network evolution to comprehend the differences between evaluation periods (Snijders 2001; Snijders et al. 2010). Our model specification from SAOM inspected the hypotheses by using its implementation of RSIENA (R Simulation Investigation for Empirical Network Analysis) (cran.r-project.org/package=RSiena).

Stochastic actor-oriented model (SAOM) for the proposed model

SAOM is a multinomial probability model for predicting changes of tie formation in network evolution model (Snijders 2001). As a statistical model of network evolution, SAOM requires at least two or more networks. The first observed network $X_{(1)} = X_{ij(1)}$ is taken as the starting point for the simulation. The observed network at time-m is represented as an adjacency matrix of binary values with $X_{ij(m)} = 1$ if there is a tie between actor-i and actor-j. Changes from a network $X_{(m)}$ to subsequent network $X_{(m+1)}$ are the results of a series of confounding mechanisms where an actor at a given moment does a micro-step or an action related to one tie formation. Micro-step indicates creating, dropping, or keeping relation to an actor which may not be responded. RSIENA models *network change processes* and *attribute change process* through simulation. The first process of network change is about tie formation, whereas the second one of attribute change is about actor changes in characteristics and behaviors (Steglich et al. 2010). SAOM specification for network change processes includes some interdependence structural effects as functions that depend on the actors. Comprehensive overviews of structural effects can be looked up in RSIENA manual (www.stats.ox.ac.uk/~snijders/siena).

In this paper, network change processes were used to validate mechanism of transitive closure with co-authors of co-authors become co-authors as well in the first hypothesis. Then, attribute change process supported the second hypothesis of researchers' interest with characteristics of actors or researchers become covariates and their researchers'

interest become behaviors. Covariates were used to explain the changes in network or behavior. Thus, there are changes of behavior variables on the observed networks. These changes occurred due to the influences of researchers or co-authors in tie formation. RSIENA requires some inputs to estimate change values that affects tie formation in SAOM based network evolution model. Co-author networks and author-topic networks became the RSIENA network input (Snijders et al. 2013). We defined the inputs of networks, covariates and behaviors with size and description shown in Table 2.

Preparing one-mode co-author networks

SAOM actors have directed relations, while the co-author networks contain undirected ties or dyads. As a result, in an article with more than two authors, more than one co-authorship reciprocal dyads are formed at a given moment. For example Article-A with three authors in Fig. 6 has six directed dyads for relations of author-*i*·author-*j*, author-*i*·author-*k* and author-*j*·author-*k*. Therefore, individual initiative in the co-authoring case must have reciprocal confirmation which implies transitivity in a situation of more than two authors. These also occurred when the involved actors mutually agree for tie formation such as trade networks of countries (Manger et al. 2012) or co-author networks of researchers (Ferligoj et al. 2015).

To show SAOM special reciprocity case of co-author networks, the experiments considered an alternative micro-step definition in constructing the RSIENA network input. We used the term of network data (ND) for co-authorship dyads. Data ND1 employed all authors in one article, while data ND2 only used two authors for each article to reduce transitivity situation. The experiments used three panel-waves of Period-1 until Period-3 for both network data ND1 and ND2 (Table 2). Period-1 became the starting point of RSIENA simulation and the panel waves of Period-2 and Period-3 were used to estimate the changes in the network evolution.

Preparing two-mode author-topic networks

Research interest or topics are referred as changing variables because of the social influences of researchers. Relations between researchers and topics are displayed in bipartite networks of binary with value 1, which means at least there is one article created by a researcher and mapped to a particular topic. There are three bipartite networks for Period-1 until Period-3 in Table 2.

Before constructing bipartite author-topic networks, titles and abstracts of articles were preprocessed to obtain clusters of words. Since articles do not have pre-determined categories, the next process was labeling articles with topics from the clustering results. A topic is actually a cluster of words, so one topic can be assumed as one article. Therefore, topic labeling of one article requires calculating similarities between two articles which carried out through computing Cosine similarity between article-topic vectors (Manning et al. 2008). The dimension of article vector and topic vector was the number of keywords after preprocessing. Element values within the vectors were term weights which represent the frequency of keywords. Cosine similarity measured the closeness of an article and a topic using the computation of term weights. We labeled articles with two or three most related topics based on higher Cosine similarity values. When the similarity difference between the second topic and the third topic had less than 0.001, it was necessary to do the third labeling process. Notably, a topic or a cluster of words had no specified label. Therefore,

Table 2 RSIENA data input for bipartite author-topic networks

No	RSIENA data input	Description
1	One mode network data (ND), X size 212 × 212	Co-authoring between researchers in three panel waves (ND1: consider ≥ 2 authors, ND2: consider only 2 authors) binary values, $x_{ij} = x_{ji} = 1$ means that author-i and author-j are co-authors
2	Two-mode (bipartite) network, W size 212 × 30	The relation between researchers and topics in three panel waves binary values, $w_{ih} = 1$ means that author-i has at least one article mapped to topic-h
3	Individual covariate (IC) size 212 × 1(career age)	Age for starting a career in publication, constant in all observations, encoded values: 1–5 with the average value is 3. (1: ≥ 2010, 2:2000–2009, 3:1990–1999, 4:1980–1989, 5: < 1980)
4	Behavior data (BD) publishing level size 212 × 3	Values of publishing level are extracted from articles without topic concern. The graded values of publishing level are: 1: publishes at least one article in a year 2: publishes at least one article per semester (6 months) 3: publishes at least more than two articles per semester 4: publishes at least one article in every other month
	Exploring level size 212 × 3	Values of exploring level are extracted from articles with topic concern. The graded values of exploring level are: 1: at most one new topic in each year during 5-years period 2: at most two new topics in each year during 5-years period 3: at least three new topics in each year during 5-years period The new topic is counted with at least one published article exist

we manually identified topic label by selecting frequent keywords from the titles of articles within a cluster for validation purpose.

After indirectly mapping topics to authors, the main objective of abstracting step in Fig. 7 is to create bipartite networks of author-topic, in addition to create co-author networks. Since there were four data groups from Period-1 to Period-4, accordingly, the abstracting step produced a set of one co-author network and one bipartite network for each data group. The experiments only used three sets of both networks from the first three data groups. A co-author network of one data group represented co-authorships occurred during certain five-year period with the bipartite network showed authors and their interests from the topic mapping step.

Defining variables based on SAOM

RSIENA covariate input is an independent variable used to explain the changes of networks or behaviors. The earliest year of publications or career age was used to control individual covariates in academic ability because researchers with more expertise tend to be older. Career age as constant RSIENA covariate input was encoded into graded values of 1–4 with 1 is assumed as the most expert (Table 2) because the researcher has the earliest published article. Our behavior input was assumed to have relations with the evolution of the network data between panel-waves, while the covariate changes had other unrelated reasons.

In the proposed model, research interest of co-authors as changing variables is represented in behavior data of publishing and exploring which become RSIENA behavior input. The behavior data is in the form of graded values where researchers with higher values indicate to have high academic level or more expertise. Both behavior data demonstrate researchers' activities whether their frequentness in publishing and their ability in exploring varied topics. We defined the terms of "publishing level" and "exploring level". The publishing level refers to the behavior of a researcher in publishing scientific articles. Higher level of publishing behavior indicates that the researcher has more expertise since the person is frequently writing. On the other hand, the exploring level refers to the behavior of researcher in exploring new topics. Researcher with higher level of exploring behavior means that the person likes to keep up to date with research trends which usually refers to numerous topics. Since there is skewness in the distribution of publication number (#pubs), the values of publishing level are corrected with formula $\ln(1+\#\text{pubs})$ and rounded to nearest integer values. For example, one researcher has 12 published articles in Period-1 is mapped to have a publishing behavior of level-3 and categorized as a thriving expert. Thus, the researcher is assumed to write with some co-authors at least more than two articles per semester on average in a period of 5 years. Publishing level of a researcher is coded into 1–4 based on the published articles without considering information of mapped topics of the articles. Definition of publishing level has minimum level of 1 ("at least once a year") until the highest expertise level of 4 ("at least once every other month") in Table 2.

As the next RSIENA behavior input, values of exploring levels for each researcher in three panel-waves are extracted from articles with regard to their mapped topics. The first value of exploring levels is obtained from Period-1, while the second and third values are obtained from differences between Period-1 to Period-2 and Period-2 to Period-3. The first exploring value of a researcher represents the number of mapped topics from his or her published articles in Period-1 which is transformed then with formula $\ln(1+\#\text{topics})$. The

second exploring value of a researcher is obtained from his or her published articles in Period-2 with different mapped topics that do not exist in Period-1. This situation demonstrates that the researcher takes a new interest and publishes an article for it. Similar to the first exploring value, the number of those mapped topics is transformed with the same formula to be the second exploring value. Then, the third exploring value is calculated by using articles form Period-2 to Period-3. Graded levels and their descriptions for both behaviors are defined by analyzing the selected dataset.

Our hypotheses rely on the assumption in which the more researchers publish, the more expertise they get. Previous studies also considered citations as scientific excellence indicator for researchers (Ferligoj et al. 2015). Our study did not consider citation number because of the data constraint since AMiner only automatically collects the recent article citations. In short, Aminer data does not contain chronological information of citations, which is necessary for network evolution analysis. The proposed model does not use information on current citations to estimate collaboration dynamics in the past. Regardless, we still needed Aminer recent citations as reliable evidences for indicating the expertise of researchers.

Apart from technical difficulties on parsing affiliation texts, especially in articles with too many authors, we did not use affiliation value because Aminer only collects the recent affiliation. Aminer collecting mechanism of the recent affiliation is already similar to the collection process of the recent citations. AMiner overrides values of citations and affiliation of each researcher with the most recent ones. Nevertheless, approximately 69% of the researcher proportion in our AMiner NLP-IE dataset affiliates to universities in the United States of America (USA). This occurrence possibly due to the researchers who used to be either staff members, former students or collaboration participants of the universities. In short, the benefit of the proposed design as shown in Fig. 7 would reduce any dependencies on other data collections aside from publication metadata.

Specifying effects based on SAOM

RSIENA estimated parameters of the specified models in three networks: Period-1, Period-2, and Period-3. The changes of parameter values in Period-2 and Period-3 were conditioned from Period-1. The estimated values are parts of selected RSIENA effect functions and used to reproduce the data distributions of network and behavior at subsequent moments in the network evolution. Two types of effect functions for RSIENA estimate processes are rate function and evaluation function. Subsequent data changes of network and behavior at Period-2 and Period-3 were indicated with the changes values of the rate functions. These rate functions showed the expected frequencies for a researcher to have opportunities to change a tie in network or behavior data. On the other hand, evaluation functions related to the selection of co-authors to determine which change a researcher (ego) makes to other researchers (alters). We iteratively specified RSIENA effect functions in the proposed model and dropped non-significant effects, which were not our primary interest and not discussed further in this paper. The significance of RSIENA effect functions refers to the t-statistics of estimates divided by the standard error from RSIENA results.

We argued that social mechanisms of the proposed models between two networks in consecutive time points are selection and influence. Selection mechanisms leads to tie formation within networks and influence mechanisms leads to changes in researchers' interest. Both mechanisms are represented with RSIENA effect functions as shown in Table 3. There were three models in the experiments to support the proposed hypotheses. Model.A assumed that

Table 3 RSIENA effect functions for the specification of the proposed models

Effects	Descriptions
Relate to co-author networks (undirected relations between researchers)	
Transitive triads (*transtriads*)	Positive estimate indicates cyclical pattern among researchers Negative estimate indicates co-authorships have hierarchical relations. Thus researchers do not seek co-authors in cyclical pattern
Knowing the popularity of alters based on degrees (*inPop*)	In-degree popularity or known as degree of alter is similar to out-degree activity since co-author networks are undirected Positive estimate supports the Matthew effect of "the richer gets richer" which translated as popular researchers tend to collaborate more
From topic agreement in the bipartite network (*from*)	Positive estimate indicates researchers with similar topic interests are most likely having co-authorship relations Negative estimate indicates researchers, who have dissimilar interests but possibly related, tend to collaborate. However, it needs further investigation
Based on the covariate values of career age (sim*X*, ego*X*, alt*X*)	Positive estimate of sim*X* indicates the researcher tendency to work with co-authors who have the same level of starting publication year Positive estimate of ego*X* indicates senior researchers who have higher values of career age tend to initiate more collaboration. Notes, ego*X* and alt*X* have similar effects because of undirected co-author networks. In case of alt*X*, it means that senior researchers tend to receive more collaboration
Based on the behavior values for selection/influence (ego*X*, alt*X*)	The effects of sender ego*X* and receiver alt*X* examine selection and influence mechanisms for tie formation in co-author networks based on behavior data of different periods
Relate to bipartite networks (directed relations of researchers to topics)	
From co-authorship leads to topic agreement (*to*)	Positive estimate indicates the interest expansion of researchers because of co-authorships
Relate to behavior data	
Behavior adaptations (avSim)	Positive estimate indicates the preference of an ego being similar or adapted to his/her alters for certain behavior

one article has more than two authors (ND1) while Model.B and Model.C assumed that one article is represented with two randomly selected authors (ND2). However, the bipartite networks in Model.C were constructed from articles written by authors in ND1.

Results and analysis

We analyzed 212 researchers with their publication metadata harvested by AMiner during approximately 20 years of the observation period. Figure 8 illustrates two of four network data of ND1 in Period-1 ($\leq$1995) and 15 years later in Period-4 (2006–2010). A network

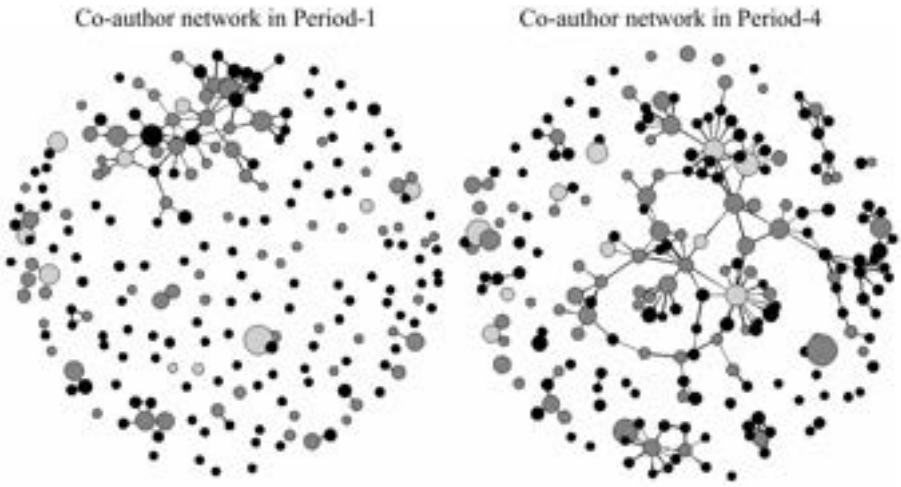


Fig. 8 Illustrations of the 15 years differences between Period-1 and Period-4 from AMiner NLP-IE dataset

tie exists when two researchers have at least one joint-publication. Node size denotes the publication number of each researcher in a certain period. The number of ties or co-author relations has increased from Period-1 to Period-4, which demonstrates the differences of network densities between two periods. Node color illustrates the effects of the collaborative nature between researchers derived from citations. A node in "white" color expresses the researcher who has the least expertise, because his or her articles are less frequently cited. Taking account on the article citations, four colors for the researcher node are distinguished according to the graded levels of "white", "light grey", "grey", and "black". The carefully selected AMiner researchers are productive and having expertise in the specified domains. Accordingly, after 15 years most researchers have advanced their careers and had more frequently cited articles with the evidence of larger and darker researcher nodes in Period-4 compared to Period-1.

Data descriptives

Node positions of researchers in two networks of Period-1 and Period-4 displayed in Fig. 8 were not fixed. Nevertheless, network characteristic values such as density, average node degree and total number of ties listed in Table 4 confirmed the influences of co-authorship increment. The average degree of NLP-IE researchers did not change much over time

Table 4 Network characteristics from AMiner NLP-IE dataset

	Network density	Average degree	Ties	0 => 1, 1 => 0, 1 => 1
Period-1	0.005	1.104	117	–
Period-2	0.007	1.575	167	110, 60, 57
Period-3	0.010	2.160	229	140, 78, 89
Period-4	0.010	2.198	233	128, 124, 105

from Period-1 to Period-4. The degree finding showed that in around 1995 within Period-1, NLP-IE researchers tended to maintain the same collaborations with 1-2 co-authors based on the *Average Degree* value 1.104. Then, after 2010 in Period-4, the researchers were more likely to have close works with 2–3 co-authors. Despite the small-scale number, there were strong links between co-authors because the researchers were already chosen for the empirical setting. They should have at least seven co-authored articles with NLP-IE experts during 15 years. The explainable situation of strong links can be continual collaboration between supervisors and doctoral students even after graduation (Shibayama 2019). The repeated co-authorships may be accounted for the attractiveness in building a research portfolio of the former students for advancing their scientific career. The networks from Period-2 to Period-4 presented evidence of maintaining co-authorships by 1 =>1 tie forming type with the changes of 57 ties to 105 ties. Co-author networks in our empirical setting were not dense based on a limitation number of formed ties. However, the degree distribution and frequency number of researcher ties of the sparse networks in Fig. 9 still followed a power-law and thereby confirming the small-world structure (Ferligoj et al. 2015).

Aside of descriptive statistics of researcher networks, this section points out structural features of co-author and bipartite networks like researcher career age. Substituting for seniority level, the individual attribute of career age indicates starting point of each researcher to publish an academic article. For example, according to the definitions in Table 2, career age = 5 is intended for researchers with more expertise because they began writing earlier than others. Consequently, the younger ones are expected to have career age = 1. The AMiner NLP-IE testbed was resulted from snowball sampling. For that reason, the chosen researchers are argued to actively participate in their career advancement to some extent. Thus, the researchers are less likely to have inactive status. With the average value of career age = 3 from the data finding, it shows that most of the researchers start publishing around 1980–2000. Although, there are some senior researchers who begin publishing before 1980 or others who start after 2000, but their combined proportions show less than 10%.

Works on Slovenian researchers incorporated SAOM micro-step of all co-author dyads in single publication to construct co-author networks (Ferligoj et al. 2015). The

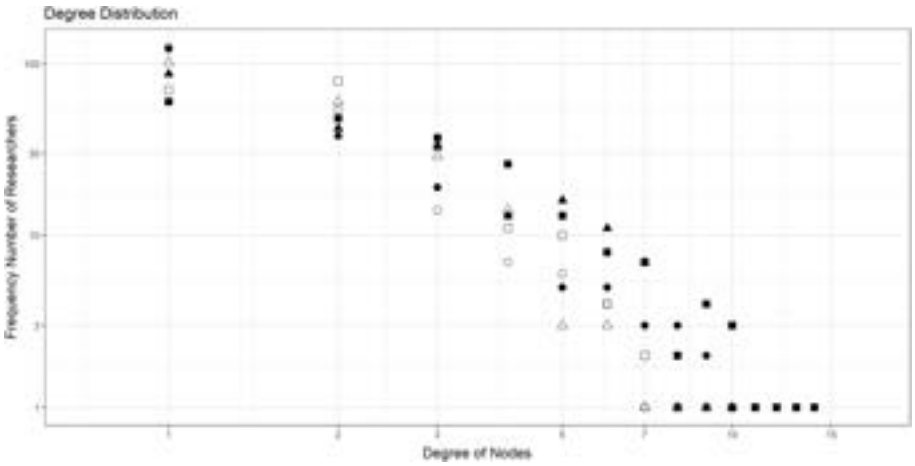


Fig. 9 Degree distribution of AMiner NLP-IE dataset in three periods of both co-author networks (ND1 {Period1 = filled circle, Period2 = filled triangle, Period3 = filled square}, ND2 {Period1 = open circle, Period2 = open triangle, Period3 = open square})

Table 5 Descriptive for RSIENA networks constructed from AMiner NLP-IE dataset

	Network descriptive	Period-1	Period-2	Period-3
1	ND1 (micro step ≥ 2 authors): Jaccard coefficient	–	0.25	0.29
2	ND2 (micro step 2 authors): Jaccard coefficient	–	0.23	0.26
3	ND1 (micro-step ≥ 2 authors): avg. clustering coefficient	0.64	0.61	0.62
4	ND2 (micro step 2 authors): avg. clustering coefficient	0.21	0.23	0.20
5	ND1 (micro step ≥ 2 authors): average degree	1.10	1.58	2.16
6	ND2 (micro step 2 authors): average degree	0.76	0.92	1.29
7	Bipartite Network (BN): Jaccard coefficient (for ND1)	–	0.14	0.18
8	Bipartite Network (BN): average degree (for ND1)	6.19	6.82	8.53

proposed model adopts that assumption as well and defined the empirical model as ND1. As an extension, our model also randomly selects one dyad to represent a tie of two researchers as co-authors in single publication, which we defined as ND2. Table 5 shows Jaccard coefficients to measure network stability between observations which displays an intermediate size for both types of network data (ND1 and ND2). According to RSIENA guidelines, networks with Jaccard values < 0.2 have some difficulties in estimation process. Elements of networks, ND1 or ND2, have (edge) value = 1 if two researchers at least have co-authored one article. The experiments inspected other networks of ND1 and ND2 with different condition as well, where the value becomes 1 if there are at least two articles co-authored by the researchers. In short, the threshold value ≥ 2. However, RSIENA results of the co-author networks with new threshold had infinitesimal values of density and constant authorship rate which gave the impression of no evolutionary process. Thus, the networks with new threshold were discarded and not discussed in this paper because the work objective was observing the evolution of co-author networks.

Elements of bipartite networks have value = 1 if an author at least writes one article in a specific topic at a particular period. The average degrees of bipartite networks slightly increased over time which revealing the interest of researchers to become somewhat expanded with at least 1–2 new topics (i.e. from 6.19 to 8.53). The expansion can be caused of co-author interactions which is confirmed by high correlation of average degrees, $r_{xy\cdot\mathrm{deg}\cdot\mathrm{ND1-BN}} = 0.98$, in co-author (one-mode) networks and author-topic (two-mode or bipartite) networks from row-5 ($x_{\mathrm{deg}\cdot\mathrm{ND1}} = \{1.10, 1.58, 2.16\}$) and row-8 ($y_{\mathrm{deg}\cdot\mathrm{BN}} = \{6.19, 6.82, 8.53\}$) (Table 5). Similar to different condition for ND1 and ND2 with threshold value ≥ 2, we also set up different bipartite networks in which a researcher is mapped to certain topic if the person published ≥ 2 articles on the topic at a particular period. However, the similar situation of RSIENA infinitesimal estimates occurred on those bipartite networks. Therefore, we discarded those bipartite networks with threshold value ≥ 2.

Other descriptive information are the average degrees in network data and the corresponding clustering coefficients. The average degrees indicated that researchers have 1–2 co-authors in average. The clustering coefficient expressed the closure existence of co-authors of co-authors becoming co-authors from triadic effects. Both descriptive values of average degree and clustering coefficient for each period did not change much over time. Those rather constant descriptive values as seen in row-3 and row-4 (Table 5) were understandable because maintaining collaboration works required efforts,

especially with worldwide partners. The explanatory reason suggested that researchers widening their collaboration networks through co-authors from various affiliations rather than expanding their fields. As stated by relatively constant values, the co-author number of a researcher was fairly kept on the same, but the co-authors themselves do not necessarily the same. The correlations of both descriptive are negative plus moderate with the values of

- $r_{xy\cdot\text{clust}\cdot\text{deg}\cdot\text{ND1}} = -0.61$
  From row-3 ($x_{\text{clust}\cdot\text{ND1}} = \{0.64, 0.61, 0.62\}$) versus row-5 ($y_{\text{deg}\cdot\text{ND1}} = \{1.10, 1.58, 2.16\}$)

- $r_{xy\cdot\text{clust}\cdot\text{deg}\cdot\text{ND2}} = -0.53$
  From row-4 ($x_{\text{clust}\cdot\text{ND2}} = \{0.21, 0.23, 0.20\}$) versus row-6 ($y_{\text{deg}\cdot\text{ND2}} = \{0.76, 0.92, 1.29\}$)

We argued that the efforts of researchers are more focused in expanding research interest instead of collaborating with new co-authors. The smaller values of clustering coefficients in ND2 are normal since there is only one dyad of two researchers accounted for each micro-step of single publication.

In short, the aforementioned data descriptive of both networks, co-author and bipartite, showed sufficient collaboration dynamics on real sampled data of AMiner NLP-IE. Higher and positive correlations on average degrees for both networks ($r_{xy\cdot\text{deg}\cdot\text{ND1}-\text{BN}} = 0.98$) revealed a relation between the expansions of research interest and the slightly increased co-authorships. As a consequence, the relation supported the first hypothesis of transitive closure.

Evaluation functions

Our empirical setting investigated the co-author factor for selection and influence on one-mode and two-mode networks. Table 6 listed some evaluation functions derived from a combination of RSIENA effects on Model.A to support the investigations and their association with our proposed hypotheses. In short, the functions consider covariate and behavior values of researchers by taking account on a scenario of ego-alter. Career age stands as the covariate aspect, whereas publishing level and exploring level as the behavior aspects.

Table 6 Evaluation functions to observe selection and influence effects on one-mode and two-mode networks of Model.A

Eq.	Function based on RSIENA effects	Confirmed hypothesis
(1)	Co-author Selection a. Career age similarity: ego, alter, similarity, ego × alter $f_i^{\text{sel}}\left(x, \text{cov}_{\text{start}\cdot\text{pub}}\right)$	*H1* Bipartite author-topic networks based on topic interests demonstrate transitive closure and researcher preferences in forming cliques
(2)	b. Publishing behavior: ego × alter $f_i^{\text{sel}}\left(x, \text{beh}_{\text{pub}}\right)$	*H2* Behavior values from bipartite author-topic networks based on topic interests are associated with experience such that researchers incline to form ties with others in looking for supervision aspect
(3)	c. Exploring behavior: ego × alter $f_i^{\text{sel}}\left(x, \text{beh}_{\text{exp}}\right)$	
(4–5)	Co-author Influence a. Publishing behavior $f_i^{\text{inf}}\left(x, \text{beh}_{\text{pub}}\right)$	
(6–7)	b. Exploring behavior $f_i^{\text{inf}}\left(x, \text{beh}_{\text{exp}}\right)$	

Table 7 Descriptive values of covariates and behaviors from RSIENA output file with one-mode networks of Model.A as input

No	Symbol	Description	Value	Data finding
From all researchers				
1	$\bar{v}$	Mean for covariate value of career age	3.127	Most researchers were starting to publish after 1990
2	$\bar{z}_{\text{pub}}$	Mean for behavior value of publishing level	1.630	Most researchers at least published two articles per year
3	$\bar{z}_{\text{exp}}$	Mean for behavior value of exploring level	1.540	Most researchers at least explored two new topics per year
From co-author networks in all panel waves (Period-1, Period-2, Period-3) With similarity variable = 1 if two researchers of a dyad have the same value				
4	$\widehat{\text{sim}^{v}}$	Similarity mean for career age	0.735	At least 70% co-author pairs have similar career age
5	$\widehat{\text{sim}^{z_{\text{pub}}}}$	Similarity mean for publishing level	0.640	At least 60% co-author pairs have similar publishing behavior
6	$\widehat{\text{sim}^{z_{\text{exp}}}}$	Similarity mean for exploring level	0.670	At least 65% co-author pairs have similar exploring behavior

Before estimating the results of specified effect functions, RSIENA obtained several descriptive values for all input networks as shown in Table 7. For example, from a descriptive value $\widehat{\text{sim}^{z_{\text{pub}}}}$, it can be inferred that the number of co-author pairs who have similar publishing level is lower than the pairs with similar career age.

Those values from Table 7 become parameters for the evaluation functions, along with the results of RSIENA estimates which are listed in Table 8. RSIENA guidelines suggest estimation results of SAOM based models should have convergence ratio of less than 0.25. Moreover, the guidelines mention that t-ratios for all estimates of specified effect functions are around 0.1 in absolute value. Nevertheless, the results of RSIENA estimates are unstandardized, such that the estimates of different effect functions are not directly comparable.

Co-author networks without behavior influences

Estimates of Model.A and Model.C in the empirical models have reasonable results according to the RSIENA guidelines as listed in Table 8 with convergence ratios of 0.19 and 0.17. The rate function of co-author networks exhibited structure changes from the significant values of network rates for the empirical models. The changes indicated author enthusiasm for widening their networks through co-authorships. Taking account on *Rate period 1* estimate of Model.A, the value 1.995 suggested that at least there are two opportunity chances for a researcher to reconsider tie formation within a span of 5 years (between Period-1 to Period-2). However, the evaluation function of co-author networks based on *Degree* estimate in Model.A has a negative value -3.349, which supported an association to the first hypothesis. This evaluation signified that even though the researchers in Model.A were motivated to have knowledge sharing through making new interactions, they were more confident to connect with co-authors of the existing co-authors who have already collaborated before. The author inclination is strongly confirmed with cyclic relations from the positive estimates of Transitive triads with value 2.084. For example, there is a situation with three author_i, author_j, author_k and x_{ij} has became co-authors. According to the first hypothesis of transitive closure, the evaluation function for author_i to partner with author_k based on estimates of *Degree* and *Transitive triads* in Model.A is going to give a benefit of $-3.349+2.084\sim-1.23$. The final negative benefit indicates a researcher has other reasons to collaborate with a new co-author aside of preceding relations signified by Transitive triads effect. Some of those other reasons are evaluated with co-author selection function based on career age similarity.

Covariate of career age signifies starting publication period and becomes researcher preference in the first hypothesis. There was strongly significant evidence in forming new ties based on the estimate of career age similarity effect in Model.A ($t = 2.973/1.012 = 2.94 \geq 3.5$)and Model.C ($t = 4.32 \geq 3.5$). We defined an evaluation function based on career age similarity $f_i^{\text{sel}}\left(x, \text{cov}_{\text{start}\cdot\text{pub}}\right)$ (1). The function presents the likelihood of author_i as ego to form a tie based on covariate value of author_j as alter. The function demonstrates preferences with several significant effects of ego, alter, similarity, and ego×alter. Because co-author networks are undirected, the alter effect is substituted with the ego effect. The evaluate function $f_i^{\text{sel}}\left(x, \text{cov}_{\text{start}\cdot\text{pub}}\right)$ uses some values listed in Table 7. Value of $\Delta_v = 5 - 1 = 4$ was taken from the covariate of career age with values 1…5 (Table 2). Formulae (1) used RSIENA estimates of ego effect $\beta_{\text{ego}} = 0.226$ and similarity effect $\beta_{\text{sim}} = 2.973$.

Table 8 RSIENA results for the proposed models constructed from AMiner NLP-IE dataset

RSIENA model for one and two mode networks	Model.A: ND1 for co-author and author-topic			Model.B: ND2 for co-author and author-topic			Model.C: ND2 for co-author and ND1 for author-topic		
	par	SE	sig.	par	SE	sig.	par	SE	sig.
Co-authorship network									
Rate period 1	1.995	0.326	***	1.050	0.175	***	0.875	0.134	***
Rate period 2	2.756	0.616	***	1.309	0.210	***	1.156	0.143	***
Endogenous effects									
Degree, $\beta_{\text{deg}}^{\text{coauthor}}$	−3.349	0.653	***	−3.665	0.201	***	−3.984	0.327	***
Transitive triads	2.084	0.360	***	1.905	0.313	***	2.322	0.500	***
In degree Popularity	*−0.006*	*0.048*		0.198	0.060	**	0.183	0.107	†
Covariate of career age effects									
Ego, β_{ego}	*0.226*	*0.232*					0.925	0.323	**
Similarity, β_{sim}	2.973	1.012	**				3.667	0.848	***
Ego × Alter, $\beta_{e\times a}$	−0.816	0.246	**				−0.849	0.192	***
Behavior effects									
ego × alter publishing, $\beta_{e\times a}^{\text{pub}}$	*0.247*	*0.265*					−1.128	0.382	**
ego × alter exploring, $\beta_{e\times a}^{\text{exp}}$	−0.734	0.437	†				*−0.086*	*0.269*	
Mixed effect									
From topic agreement (bipartite)	*−0.043*	*0.355*							
Bipartite network									
Rate period 1	29.636	2.40	***	13.992	0.99	***	29.582	1.829	***
Rate period 2	44.766	10.88	***	*61.141*	*15.86*		44.941	5.584	***
Mixed effect									
Out degree, $\beta_{\text{outdeg}}^{\text{bipartite}}$	−0.555	0.023	***	−1.076	0.019	***	−0.550	0.019	***
Co-authorship to topic agreement	*0.012*	*0.033*					*0.001*	*0.045*	
Co-evolution behavior: publishing									
Rate period 1	2.008	0.262	***	2.136	0.343	***	2.242	0.324	***

Table 8 (continued)

RSIENA model for one and two mode networks	Model.A: ND1 for co-author and author-topic			Model.B: ND2 for co-author and author-topic			Model.C: ND2 for co-author and ND1 for author-topic		
	par	SE	sig.	par	SE	sig.	par	SE	sig.
Rate period 2	2.307	0.395	***	4.189	0.784	***	4.040	0.741	***
Behavior dynamics									
Linear, β^{pub}_{linear}	*0.083*	*0.108*		−0.242	0.075	**	−0.376	0.096	***
Quadratic, β^{pub}_{quad}	*−0.012*	*0.063*		*0.022*	*0.088*		0.149	0.072	*
Average similarity, β^{pub}_{avSim}	5.715	2.638	*				2.658	1.275	*
Co-evolution behavior: exploring									
Rate period 1	4.691	1.272	***	3.400	0.707	***	4.944	0.834	***
Rate period 2	4.527	0.985	***	8.912	3.463	**	5.237	0.904	***
Behavior dynamics									
Linear, β^{exp}_{linear}	*−0.188*	*0.164*		−0.303	0.057	***	−0.105	0.059	†
Quadratic, β^{exp}_{quad}	*0.136*	*0.137*		*0.018*	*0.036*		*−0.028*	*0.048*	
Average similarity, β^{exp}_{avSim}	*7.803*	*5.526*					4.930	1.491	**
Convergence ratio	0.19, all t-ratios ≤ |0.1|			2.81, all t-ratios < |1.9|		0.17, all t-ratios < |0.1|		
1.7 ≤ t-statistic < 2.0	†$p < 0.1$		Highly suggestive significant				italic:	not significant	
2.0 ≤ t-statistic < 2.5	*$p < 0.05$		Weakly significant						
2.5 ≤ t-statistic < 3.5	**$p < 0.01$		Moderately significant				t-stats.	= |par/s.e.|	
t-statistic (stats.) ≥ 3.5	***$p < 0.001$		Strongly significant						

$$
\begin{aligned}
f_i^{\text{sel}}\left(x, \text{cov}_{\text{start-pub}}\right) &= \beta_{\text{ego}}\left(v_i - \bar{v}\right) + \beta_{\text{ego}}\left(v_j - \bar{v}\right) + \beta_{\text{sim}}\left(1 - \frac{\left|v_i - v_j\right|}{\Delta_v} - \widehat{\text{sim}^v}\right) \\
&\quad + \beta_{ex\alpha}\left(v_i - \bar{v}\right)\left(v_j - \bar{v}\right) \\
&= 0.23\left(v_i - 3.13\right) + 0.23\left(v_j - 3.13\right) \\
&\quad + 2.97\left(1 - \frac{\left|v_i - v_j\right|}{4} - 0.73\right) - 0.82\left(v_i - 3.13\right)\left(v_j - 3.13\right)
\end{aligned} \tag{1}
$$

The older alter is a researcher who starts before 1980 while the younger one starts after 1990. Without concerning significance of the estimates, the evaluation function (1) was supported by ego effect which close to 0 ($\beta_{\text{ego}}^{\text{age}} = 0.226$) and similarity effect ($\beta_{\text{sim}}^{\text{age}} = 2.973$) which rather dominant. Thus, Fig. 10 indicates a stronger preference for researchers as ego to be linked with alters who have similar career age with around 3–7 years in different ($v_j^{\text{age}} = \{3, 4\}$) or seniors who have more writing experience at least 10 years before themselves ($v_j^{\text{age}} = 5$). The indication supports the concept of knowledge transfer or academic mentoring in collaborations of certain domains such as the supervision aspect (Shibayama 2019). The benefits in Fig. 10 are all positive values which means start publishing is one of supporting factors to collaborate. If we used Transitive triads from previous example with $v_i^{\text{age}} = 4$ and $v_j^{\text{age}} = 3$, then the result was positive, $-3.349 + 2.084 + 1.24 \sim 0.01$. The findings in terms of starting publication indicate researchers have a preference to work with friends (co-authors) of their co-authors, signified by Transitive triads effect, who are peers or seniors.

Bipartite networks without behavior influences

Previous discussions focused on the covariate of career age and its implication on the collaboration dynamics of co-author networks. Apart of co-author networks, our study also considered author-topic or bipartite networks. RSIENA results on the dynamics of bipartite networks showed that the rate functions revealed higher chances for scholars to reorganize their research career. We specified bipartite networks were about relations between researchers and research topics which had been set to 30 topics. The term of

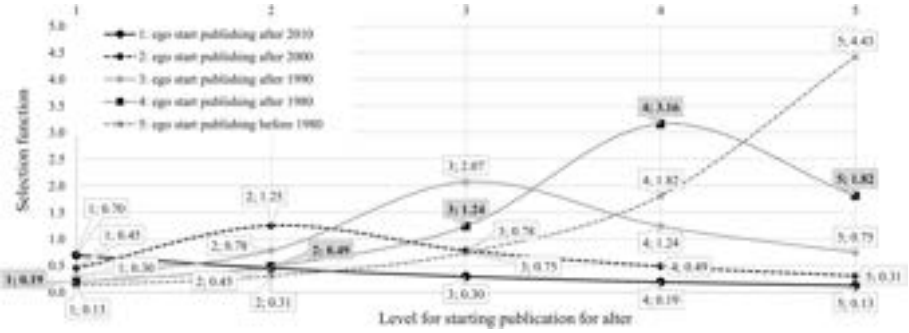


Fig. 10 Plot of the log-odds for co-authorship selection with respect to starting publication covariate from AMiner NLP-IE dataset

reorganized career was meant to the possibilities of ties connecting to certain topics, thus widening the interest of researchers. Table 8 presented the estimate *Rate period 1* on Model.A with value 29.636 which we rounded to 30. The estimate of *Rate period 1* meant that there are nearly 30 opportunities for researchers to change, withdraw or maintain the same topics on their first 5 years of collaborations. The fact that the estimate *Rate period 2* had higher value of 44.766 was consistent with more possibilities to change topic on the next 5 years for the researchers, which was nearly 45 chances.

Researcher enthusiasm to have new topic interests is even higher than the enthusiasm to connect to new co-authors. The findings are consistent with RSIENA estimates from *Out Degree* value (mixed effect in bipartite networks) $\beta_{\text{outdeg}}^{\text{bipartite}} = -0.555$ which is larger than *Degree* value (endogenous effect in co-author networks) $\beta_{\text{deg}}^{\text{coauthor}} = -3.349$. Further finding showed that collaborations do not necessarily make researchers to have the same interest to their co-authors. The evidence was on the insignificant value 0.012 for the estimate of "co-authorship to agreement" effect. Even more, the co-authors do not always have similar topic interest based on insignificant and negative estimate of "from topic agreement" -0.043. Mixed topics could suggest inter-departmental research collaborations in universities (Purwitasari et al. 2017).

Even with hesitance, having influenced by co-authors, researchers may explore new topics. The same NLP-IE domain allows them to explore on the fusions of their previous interests and the co-authors' interest. Notably, including the probable fusion topics, our experiments interpreted 30 fine-grained subjects as defined from the preliminary clustering process. The supporting evidence of topic advances depicted in Fig. 11 with the popularities of 30 topics ordered by the least number of followers in Period-1. The term topic follower represents a researcher who publishes articles on that particular topic. The illustration suggested there were growths of some topics such as "collaboration filtering" topic (T22) and "software agents" topic (T23). As mentioned before, the topic labels were manually defined for validation purposes. The search results of Google Scholar have also elucidated the findings on the topic growths. Topic T22 in 1995 (Period-1) had around 10 thousand articles and increased to ten times more in 2015 (Period-4), while topic T23 even superior from around 400 thousand articles to almost two million articles.

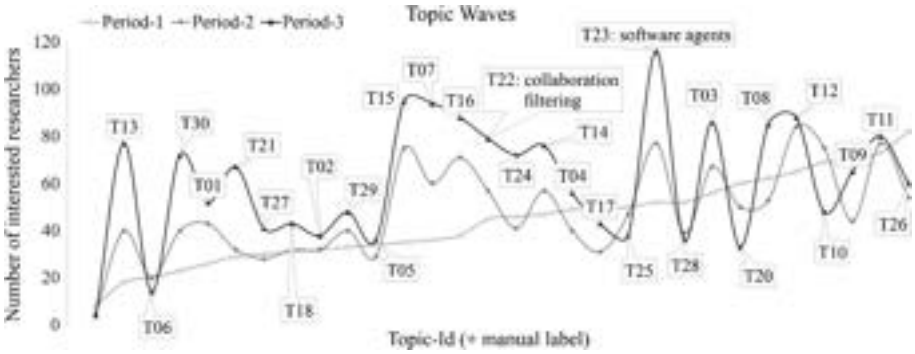


Fig. 11 Topics with increasing popularities over time from AMiner NLP-IE dataset

Behavior influences on the networks

The collaboration dynamics of one mode and two mode networks indicate chances for researchers in expanding their interest. The influence of co-authors can become essential roles in the interest expansion. We examined behaviors of researchers in publishing and exploring related to co-author selection proses, thereby suggesting that co-author tie formation would be more beneficial in interest expansion. Those four evaluation functions derived from two behavior effects are already mentioned in Table 6.

- *Publishing Behavior* of one researcher indicates the frequency of his or her published articles regardless the topics during each period of five years.
- *Exploring Behavior* in a period indicates the frequency of articles being published in different topics than before compared to the previous period.

Publishing behavior has four graded levels, while exploring behavior has three graded levels (Table 2).

We defined behavior effects of linear and quadratic in the experiments of Model.A to support the evaluation functions. RSIENA results for co-author networks and behavior co-evolution showed the effect of a behavior (linear effect) and the effect to of a behavior on itself (quadratic effect). For example, the quadratic effect of publishing demonstrates whether a researcher who often publishes articles is going to be more actively writing in the future.

We fit ego × alter effect to the evaluation function $f_i^{\text{sel}}\left(x, \text{beh}_{\text{pub}}\right)$ for co-author selection based on publishing behavior without being concerned to the significance of RSIENA estimates (2). The function presents the likelihood of author_i as an ego to form a tie based on publishing behavior value of author_j as the alter. The function uses some parameters from RSIENA output file (Table 7) and ego × alter effect ($\beta_{e\times a}^{\text{pub}} = 0.247$) which closes to 0. By using formulae (2), the illustration for selection preference is inferred in Table 9 showing upper triangular values and disregarding dark cells, since author relation is undirected. Cell values in grey area of each row display the attractiveness of alters to co-author with egos based on behavior values. The cell values are obtained from formulae (2) with combinations of publishing behavior values for ego z_i^{pub} and alter z_j^{pub}.

Table 9 Evaluation function results based on publishing behavior

Based on publishing behavior	Selection attractiveness				Influence attractiveness			
An alter with z_j^{pub}	alt_1	alt_2	alt_3	alt_4	alt_1	alt_2	alt_3	alt_4
An ego publishes at least … (z_i^{pub})								
One article in a year (ego_1)	0.10	−0.06	−0.21	−0.37	1.15	−1.26	−1.28	−1.33
One article per semester (6 months) (ego_2)		0.03	0.13	0.22		0.73	0.71	0.66
More than two articles per semester (ego_3)			0.46	0.80			2.69	2.65
One article in every other month (ego_4)				1.39				4.64

$$f_i^{\text{sel}}\left(x, \text{beh}_{\text{pub}}\right) = \beta_{e\times a}^{\text{pub}}\left(z_i^{\text{pub}} - \bar{z}_{\text{pub}}\right)\left(z_j^{\text{pub}} - \bar{z}_{\text{pub}}\right) = 0.25\left(z_i^{\text{pub}} - 1.63\right)\left(z_j^{\text{pub}} - 1.63\right) \tag{2}$$

An ego who publishes less, $z_i^{\text{pub}} = 1$, with the values of ego_1 row in *Selection attractiveness* of Table 9, has a somewhat lower attraction to productive alters with higher level of publishing behavior. The probability value of $\text{ego}_1 \times \text{alt}_2 = \text{e}^{-0.06} = 0.94 < 1$ showed that co-authoring between $\text{ego}_1 \times \text{alt}_2$ is less likely happened. Lower numbers with the negative sign of the attractiveness values for ego_1 implies less chances of co-authorship. However, the collaboration chance is higher if both authors stand on the same stage, inferred from their publishing frequency of one article in a year, $\text{ego}_1 \times \text{alt}_1 = \text{e}^{0.10} = 1.11$. It may be caused by the fact that the experiments used AMiner NLP-IE dataset with most of them were experts (Deng et al. 2012) and had some international collaborators. However, the probability of middle experts ($z_i^{\text{pub}} = 2, 3$)) to co-author with researchers who have more experience is high. Therefore, the finding suggests either ego or alter is at least publishing more than one article in a year ($z_i^{\text{pub}} \geq 2, z_j^{\text{pub}} \geq 2$). The probability to connect for an ego $z_i^{\text{pub}} = 3$ with more active alter of $z_j^{\text{pub}} = 4$ is $\text{e}^{0.80} = 2.23$ times as high as the probability of no forming ties at all. The judgments are consistent with the values from rows of ego_2 and ego_3 in *Selection attractiveness*, suggesting the more prepared the researchers are for international collaborations (Iglič et al. 2017). In short, for ego who regularly publishes ($z_i^{\text{pub}} \geq 3$), the level of publishing behavior of alters hardly matters.

We also fit the same ego × alter effect to the function $f_i^{\text{sel}}\left(x, \text{beh}_{\text{exp}}\right)$ (3), which presents the likelihood of author_i as an ego to form a tie based on exploring behavior value of author_j as the alter.

$$\begin{aligned} f_i^{\text{sel}}\left(x, \text{beh}_{\text{exp}}\right) &= \beta_{e\times a}^{\text{exp}}\left(z_i^{\text{exp}} - \bar{z}_{\text{exp}}\right)\left(z_j^{\text{exp}} - \bar{z}_{\text{exp}}\right) \\ &= -0.73\left(z_i^{\text{exp}} - 1.54\right)\left(z_j^{\text{exp}} - 1.54\right) \end{aligned} \tag{3}$$

The results of formulae (3) coincided with the previous finding on publishing behavior with computed values for the attractiveness of ego × alter based on exploring behavior are listed in Table 10. The findings concluded that researchers will get more benefit if their co-authors have different gap in exploring level. The probability of an ego who focuses on one new topic ($z_i^{\text{exp}} = 1$) to work with an alter who explores three topics ($\text{e}^{0.63} = 1.88$) is almost doubled, compared to the working probability with another alter who just interested in two topics $z_j^{\text{exp}} = 2$ ($\text{e}^{0.17} = 1.19$). The most attractive alters have the least topics to explore. There are two reasonable situations for alters who explore a few topics. With the situation of younger researcher, alter can have more responsibilities in the experimental works. On the other hand, if the alter is more senior, exploring fewer topics means that the researcher has become the

Table 10 Evaluation function results based on exploring behavior

Based on exploring behavior	Selection attractiveness			Influence attractiveness		
An alter with z_j^{exp}	alt_1	alt_2	alt_3	alt_1	alt_2	alt_3
An ego explores … (z_i^{exp})						
At most one new topic in each year ego_1	−0.29	0.17	0.63	2.25	−1.98	−1.72
At most two new topics in each year ego_2		−0.10	−0.37		1.74	2.00
At least three new topics in each year ego_3			−1.38			5.71

expert. All the same, both situations make the mentoring process practically possible (Shibayama 2019). Evidence in the row of ego_1 in Table 10 with positive values of *Selection attractiveness* are consistent with the expert situation. Therefore, the finding suggests to work with more focused researcher who only explores one new topic in a year.

Both evaluation functions on co-author selections based on behaviors (2) and (3) have comparable effects from the difference between the highest and lowest values of the functions.

The difference of (2) is taken from $\left|\text{ego}_4 \times \text{alt}_4\right| + \left|\text{ego}_1 \times \text{alt}_4\right| = 1.39 + 0.37 = 1.76$ (Table 9).

The difference of (3) is taken from $\left|\text{ego}_1 \times \text{alt}_3\right| + \left|\text{ego}_3 \times \text{alt}_3\right| = 0.63 + 1.38 = 2.01$ (Table 10).

Thus, the final evaluation functions for co-author selection with default *Degree* effect and aforementioned *Transitive triad* effect plus one of behavior effects are as follows.

Considering publishing behavior, the function shows a positive result, $-3.349 + 2.084 + 1.76 = 0.495$

Then from exploring behavior, the function also shows a positive result, $-3.349 + 2.084 + 2.01 = 0.745$

Positive result of an evaluation function on co-author selection denotes a collaboration on co-authoring. We concluded that the preferred alters would be researchers who publish often (the selection attractiveness value ≥ 0.46 from Table 9) and explore more (the selection attractiveness value ≥ 0.63 from Table 10), which again solidifies the mentoring process (Shibayama 2019).

Aside from selection impact, other behavior related impact is co-author influence. Following RSIENA guidelines, we specified two evaluation functions using co-author influence from publishing behavior (4–5) and exploring behavior (6–7). The objective is to know how likely an ego changes his or her behavior based on the behavior of alters. Both functions apply the effects of behavior trend (linear), behavior on the behavior itself (quadratic) and average similarity to represent social relationship between researchers. The number of co-authors and the range of behavior differences are required to denote the social relationship issue. For a given researcher x_i, the number of co-authors is symbolized as x_{i+}. Then, the value $\Delta_{z\cdot\text{beh}}$ represents the range of certain behavior by max-value minus min-value, such as $\Delta_{z\cdot\text{pub}} = 4 - 1 = 3$ and $\Delta_{z\cdot\text{exp}} = 3 - 1 = 2$.Moreover, values on the following functions are taken from Tables 7 and 8.

$$f_i^{\text{inf}}\left(x, \text{beh}_{\text{pub}}\right) = \beta_{\text{linear}}^{\text{pub}}\left(z_i^{\text{pub}} - \bar{z}_{\text{pub}}\right) + \beta_{\text{quad}}^{\text{pub}}\left(z_i^{\text{pub}} - \bar{z}_{\text{pub}}\right)^2 + \beta_{\text{avSim}}^{\text{pub}} \frac{1}{x_{i+}} \sum_j x_{ij} \left(\frac{\left|\bar{z}_{\text{pub}} - z_j^{\text{pub}}\right| - \left|z_i^{\text{pub}} - z_j^{\text{pub}}\right|}{\Delta_{z\cdot\text{pub}}} \right) \tag{4}$$

$$f_i^{\text{inf}}\left(x, \text{beh}_{\text{pub}}\right) \cong 0.08\left(z_i^{\text{pub}} - 1.63\right) - 0.01\left(z_i^{\text{pub}} - 1.63\right)^2 + 5.72\left(\frac{\left|1.63 - z_j^{\text{pub}}\right| - \left|z_i^{\text{pub}} - z_j^{\text{pub}}\right|}{3} \right) \tag{5}$$

$$f_i^{\text{inf}}\left(x, \text{beh}_{\text{exp}}\right) = \beta_{\text{linear}}^{\text{exp}}\left(z_i^{\text{exp}} - \bar{z}_{\text{exp}}\right) + \beta_{\text{quad}}^{\text{exp}}\left(z_i^{\text{exp}} - \bar{z}_{\text{exp}}\right)^2 + \beta_{\text{avSim}}^{\text{exp}} \frac{1}{x_{i+}} \sum_j x_{ij} \left(\frac{\left|\bar{z}_{\text{exp}} - z_j^{\text{exp}}\right| - \left|z_i^{\text{exp}} - z_j^{\text{exp}}\right|}{\Delta_{z\cdot\text{exp}}} \right) \tag{6}$$

$$f_i^{\text{inf}}\left(x, \text{beh}_{\text{exp}}\right) \cong -0.19\left(z_i^{\text{exp}} - 1.54\right) + 0.14\left(z_i^{\text{exp}} - 1.54\right)^2 + 7.80 \left(\frac{\left|1.54 - z_j^{\text{exp}}\right| - \left|z_i^{\text{exp}} - z_j^{\text{exp}}\right|}{2} \right) \tag{7}$$

The evaluation functions of co-author influence on behavior (5) and (7) are used to obtain the values of Influence attractiveness as presented in Tables 9 and 10 for all combinations of ego × alter. The values are computed with an assumption that all co-authors have the same behavior values. Cell values at the diagonal position in Tables 9 and 10 are the highest, which were obtained from a scenario where ego and alter have the same behaviors, $z_i^{\text{pub}} = z_j^{\text{pub}}$ or $z_i^{\text{exp}} = z_j^{\text{exp}}$. However, the social pressure of alters seemed to have a role in changing the behavior of ego. The interpretation occurred because the *Average Similarity* effect on both evaluation functions is rather dominant compared to the estimates of *Linear* and *Quadratic* effects which are close to 0. However, inactive alters ($z_j^{\text{pub}} = 1$ or $z_j^{\text{exp}} = 1$) have no persuasion power of any kind to more active ego ($z_i^{\text{pub}} > 1$ or $z_i^{\text{exp}} > 1$), which is signified by the negative sign in the values of Influence attractiveness (for the odds $e^{-\text{val}} < 1$) for both behaviors.

The persuasion power of active alters has increased and is consistent with an exponential model of $0.039e^{1.972z_j^{\text{pub}}}$ with a condition of $z_i^{\text{pub}} > 1$ and $z_i^{\text{pub}} \geq z_j^{\text{pub}}$. In a case of alter with $z_j^{\text{pub}} = 2$, the influence values for $\text{ego}_2, \text{ego}_3, \text{ego}_4$ are $0.73, 0.71, 0.66$ and make the odds of $e^{0.73} \sim 2.1, e^{0.66} \sim 1.9$. With the exponential model, the persuasion power of alter $z_j^{\text{pub}} = 2$ equals to $0.039e^{1.972\times2} \sim 2.01$. The power of alter $z_j^{\text{pub}} = 2$ stays the same for all active ego ($z_i^{\text{pub}} = \{2 \ldots 4\}$). It can be interpreted that the possibility of the ego changes his or her publishing habit is two times higher compared to no changing situation with co-authors who have one article per semester. If the co-authors are more active, the possibility has increased to 14 times higher for $z_j^{\text{pub}} = 3$ and more than 100 times higher for $z_j^{\text{pub}} = 4$. The findings show the strength of social pressure of the co-authors. If ego and alters have the same publishing behavior values, the influences are mutual since the researchers strengthened one another, with the highest value at diagonal position as evidence. This result confirmed the perspectives in building good habit for self-improvement as researchers by connected to supporting co-authors (Datta et al. 2017) which makes another reinforcement of the mentoring process (Shibayama 2019).

For the final evaluation function on co-author influence based on exploring behavior, there is only one notable case in which alters have higher values than the ego ($z_i^{\text{exp}} = 2$ and $z_j^{\text{exp}} = 3$). That particular case has the odds of $e^{2.00} \sim 7.4$ indicating the possibility of the ego expands his or her exploring habit is seven times higher compared to no changing situation. The influencers are co-authors who keeping up to date with the latest research trends and suggesting stronger motivations for replicating their behavior as parts of the mentoring process (Shibayama 2019).

Discussions

This study investigates to what degree co-author relations affecting the changes of interest and behavior of the researchers. Summaries of our findings are displayed on Fig. 12. Aside of our mechanisms in data preparation, the topic approaches based on clustering in current studies are similar to other works for recommending academic collaborators (Kong et al. 2017). Our mechanisms can be applied to any article metadata regardless disciplines on a cold-start situation with researcher expertise scarceness or topic absence (Zhu et al. 2017; de Siqueira et al. 2018). Current studies used AMiner dataset which was controlled to include only mid-researchers and co-authors who publish more than the inactive ones in NLP-IE domains. Automatically harvested metadata in AMiner dataset are journal articles and conference papers. We have selected a number of co-authors which was shown through data descriptive of co-author average degree with longitudinal analysis. However our results were comparable to the studies of Slovenian co-author networks, especially in the field of computer science, which had more systematized conditions, i.e. well-maintained topics (Ferligoj et al. 2015) and the studies that only employed metadata of certain journal articles (Abbasi et al. 2011).

We firstly confirmed transitive closure on the proposed stochastic agent-based model before arguing preferences and behaviors as the contributing factors in selecting and influencing co-authors. The proposed model with topic absence confirmed transitive closure from the networks of co-authors and author-topic. Our model is comparable to related models of co-author networks with standardized scientific disciplines or without topic absence issue (Wen and Junping 2014; Ferligoj et al. 2015).

The intention to get author property from article metadata without manual data collection (Siciliano et al. 2018) made our studies use career age covariate (Ferligoj et al. 2015) to indicate researcher experience. Coupling with transitive closure (RES1 in Fig. 12), the career age preference in our findings suggests the concept of knowledge transfer or academic mentoring in co-authoring collaborations such as the supervision aspect (Shibayama 2019). The findings demonstrated that researchers are likely to collaborate more with co-authors who have closer years in starting publication (peers) or who have higher values of starting publication levels (seniors) (RES2 in Fig. 12). The findings established our hypotheses in investigating bipartite networks with the perspective that researchers can be the influencer for expanding research interest. The influencers have similar functions of core authors in other studies on collaboration dynamics (Abrahams et al. 2019). By

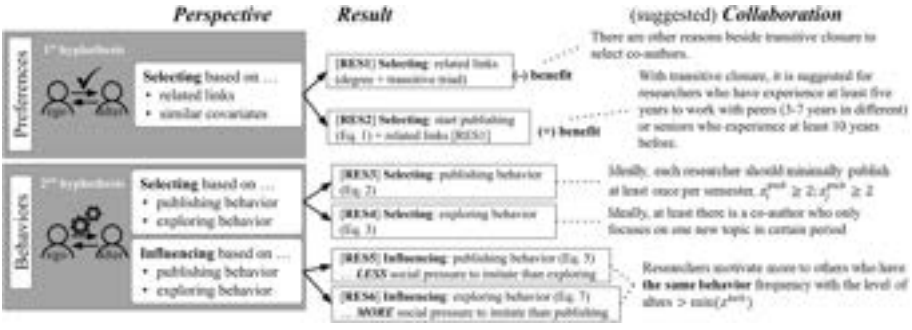


Fig. 12 The results of specified model for identifying collaboration dynamics with the influences of interest changes

co-authoring with active peers or senior mentors, researchers are exposed to have slightly different topic interests, such as collaborations across disciplines (Iefremova et al. 2018; Purwitasari et al. 2017), which happened in our experiments since clustered words as topics in NLP-IE domains are interrelated. Evidence of the expansion extent of research interest was stated in the data descriptive results of bipartite networks through average degrees, besides the correlations for one and two mode networks.

Researchers who want to build their scientific career, usually the middle experts, are attracted to others with certain academic level (RES3 in Fig. 12). Co-authors can be less experienced but productive or have much higher expertise for mentoring (RES4 in Fig. 12) as edge factor in the collaborations (Shibayama 2019). There is higher collaboration chance for co-authorships if all authors stand on the same stage, which is inferred from their publishing behavior. Whatever stages of the co-authors, their helps are advantageous in research career advancement (Iglič et al. 2017). In short, researchers prefer to collaborate with co-authors of their co-authors who are peers or seniors but do not necessarily have similar interest. The peer-or-senior status could refer to starting publication year, publishing frequentness or exploration motives.

Thus, the levels of exploration and exploitation regarding research topics (Siciliano et al. 2018) should be carefully maintained (RES5–RES6 in Fig. 12). We defined the exploration as exploring behavior level, while the exploitation as publishing behavior level. In author-keyword networks (Wang et al. 2018), researchers who have slightly changed interest or focusing on fusion topics can change the network models. However in our proposed model, the change issue is controllable, since the keywords are treated as topics in a broader context.

Because of interest changes, behavior levels of researchers are changed, which eventually affecting their positions as the influencer in expanding interest. Behavior scenarios took part not only in selecting co-authors but also in influencing them. Thus, preferences and behaviors indicate strong predictors in tie formation of co-author network and author-topic network. A researcher and the co-authors can support and strengthened one another if there is less discrepancy in their publishing behavior levels. Another tie forming factor relies on the researchers who become the influencers for interest changes because of high levels of exploring behavior. Although co-author behaviors could be one type of social pressure, the researchers who have conformed tendencies would consider collaborations through co-authorship as self-improvement for advancing their career.

Furthermore, we have pointed out that by understanding researcher behaviors, there will be some known insights as guidance in creating policies. For example, a funding tendency to one research team with members at least some seniors who have minimal article numbers, and more of peers who at least have annually published one article in the last 5 years.

Conclusions

In this paper, we empirically investigated driving factors in collaboration dynamics using longitudinal data of co-author (one-mode) networks, author-topic (two-mode) networks, and behaviors related to the co-authorships. We processed metadata of co-authored articles from selected researchers of AMiner database to construct the networks, and extract author properties as inputs in the proposed stochastic agent-based models.

Besides the empirical discussion with research interest change perspective, our findings stated the properties of authors were the driving factors, such as starting publication

year and author behaviors in publishing as well as exploring. Some researchers were likely to have more collaborations if they had minimal years in publishing experience, i.e. have closer years (peers) or higher values (seniors) in starting publication. Behavior findings on researchers as ego-alter combinations provided collaboration evidence as well when they were fitting on certain condition.

The driving factors of publishing experience or research-related behaviors, to some extent, reflect key points for recommending co-author candidates with high academic level in collaborations. We are going to employ those factors for constructing researcher profiles in automatically specified scientific fields within our aforementioned framework as the defining parameters to improve expert recommendation.

Acknowledgements This work as parts of a dissertation about scholar profiles in expert recommendation system was funded by the Indonesia Endowment Fund for Education (LPDP in Indonesian) with the grant number PRJ-4228/LPDP.3/2016 of the LPDP Doctoral Scholarship Programme fiscal year 2017–2020. Some sections of the manuscript was prepared during September-December 2018 in University of Groningen, the Netherlands under Enhancing International Publication (EIP or PKPI in Indonesian) Program by Ministry of Research, Technology and Higher Education of the Republic of Indonesia (RISTEKDIKTI in Indonesian). Furthermore, this research was also partially funded by RISTEKDIKTI under World Class Universities (WCU) Program managed by Institut Teknologi Bandung, Indonesia in 2019.

Compliance with ethical standards

Conflict of interest The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K. J. R. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics, 89*(2), 687. https://doi.org/10.1007/s11192-011-0463-1.

Abrahams, B., Sitas, N., & Esler, K. J. (2019). Exploring the dynamics of research collaborations by mapping social networks in invasion science. *Journal of Environmental Management, 229,* 27–37. https://doi.org/10.1016/j.jenvman.2018.06.051.

Amjad, T., Daud, A., & Song, M. (2018). Measuring the impact of topic drift in scholarly networks. In *Companion Proceedings of the The Web Conference 2018* (pp. 373–378). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3186358.

Aubaidan, B., Mohd, M., & Albared, M. (2014). Comparative study of K-means and K-Means ++ clustering algorithms on crime domain. *Journal of Computer Science, 10*(7), 1197–1206. https://doi.org/10.3844/jcssp.2014.1197.1206.

Ayaz, S., Masood, N., & Islam, M. A. (2018). Predicting scientific impact based on h-index. *Scientometrics, 114*(3), 993–1010. https://doi.org/10.1007/s11192-017-2618-1.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries, 17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy, 33*(4), 599–616. https://doi.org/10.1016/j.respol.2004.01.008.

Datta, S., Basuchowdhuri, P., Acharya, S., & Majumder, S. (2017). The habits of highly effective researchers: An empirical study. *IEEE Transactions on Big Data, 3*(1), 3–17. https://doi.org/10.1109/TBDATA.2016.2611668.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems, 23*(6), 615–625. https://doi.org/10.1016/j.knosys.2010.04.008.

de Siqueira, G. O., Canuto, S., Gonçalves, M. A., & Laender, A. H. F. (2018). A pragmatic approach to hierarchical categorization of research expertise in the presence of scarce information. *International Journal on Digital Libraries*. https://doi.org/10.1007/s00799-018-0260-z.

Deng, H., Han, J., Lyu, M. R., & King, I. (2012). Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 71–80). https://doi.org/10.1145/2232817.2232833.

Ebadi, A., & Schiffauerova, A. (2015). On the relation between the small world structure and scientific activities. *PLoS ONE, 10*(3), e0121129. https://doi.org/10.1371/journal.pone.0121129.

Ferligoj, A., Kronegger, L., Mali, F., Snijders, T. A. B., & Doreian, P. (2015). Scientific collaboration dynamics in a national scientific system. *Scientometrics, 104*(3), 985–1012. https://doi.org/10.1007/s11192-015-1585-7.

Fu, T. Z. J., Song, Q., & Chiu, D. M. (2014). The academic social network. *Scientometrics, 101*(1), 203–239. https://doi.org/10.1007/s11192-014-1356-x.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102.

Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012). Spherical k-means clustering. *Journal of Statistical Software Articles, 50*(10), 1–22. https://doi.org/10.18637/jss.v050.i10.

Hou, H., Wang, C., Luan, C., Wang, X., & Zhuang, P. (2013). The dynamics of scientific collaboration networks in scientometrics. *COLLNET Journal of Scientometrics and Information Management, 7*(1), 121–140. https://doi.org/10.1080/09737766.2013.802627.

Iefremova, O., Wais, K., & Kozak, M. (2018). Biographical articles in scientific literature: Analysis of articles indexed in Web of Science. *Scientometrics, 117*(3), 1695–1719. https://doi.org/10.1007/s11192-018-2923-3.

Iglič, H., Doreian, P., Kronegger, L., & Ferligoj, A. (2017). With whom do researchers collaborate and why? *Scientometrics, 112*(1), 153–174. https://doi.org/10.1007/s11192-017-2386-y.

Jung, J. J. (2015). Big bibliographic data analytics by random walk model. *Mobile Networks and Applications, 20*(4), 533–537. https://doi.org/10.1007/s11036-014-0555-2.

Kong, X., Jiang, H., Wang, W., Bekele, T. M., Xu, Z., & Wang, M. (2017). Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics, 113*(1), 369–385. https://doi.org/10.1007/s11192-017-2485-9.

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

Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. *Journal of Informetrics, 6*(4), 639–644. https://doi.org/10.1016/j.joi.2012.06.006.

Li, H., An, H., Wang, Y., Huang, J., & Gao, X. (2016). Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network. *Physica A: Statistical Mechanics and its Applications, 450,* 657–669. https://doi.org/10.1016/j.physa.2016.01.017.

Liang, W., Jin, Q., Lu, Z., Wu, M., & Hu, C. (2016). Analyzing of research patterns based on a temporal tracking and assessing model. *Personal and Ubiquitous Computing, 20*(6), 933–946. https://doi.org/10.1007/s00779-016-0965-1.

Lin, S., Hong, W., Wang, D., & Li, T. (2017). A survey on expert finding techniques. *Journal of Intelligent Information Systems, 49*(2), 255–279. https://doi.org/10.1007/s10844-016-0440-5.

Manger, M. S., Pickup, M. A., & Snijders, T. A. B. (2012). A hierarchy of preferences: A longitudinal network analysis approach to PTA formation. *Journal of Conflict Resolution, 56*(5), 853–878. https://doi.org/10.1177/0022002712438351.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Meho, L. I. (2019). Using Scopus's CiteScore for assessing the quality of computer science conferences. *Journal of Informetrics, 13*(1), 419–433. https://doi.org/10.1016/j.joi.2019.02.006.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems—Volume 2* (pp. 3111–3119). http://dl.acm.org/citation.cfm?id=2999792.2999959.

Oliveira, M., Curado, C., & Henriques, P. L. (2018). Knowledge sharing among scientists: A causal configuration analysis. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2018.12.044.

Ortega, J. L. (2014). AMiner: Science networking as an information source. In J. L. Ortega (Ed.), *Academic search engines* (pp. 47–70). Oxford: Chandos Publishing. https://doi.org/10.1533/9781780634722.47.

Purwitasari, D., Fatichah, C., Arieshanti, I., & Hayatin, N. (2016). K-medoids algorithm on Indonesian Twitter feeds for clustering trending issue as important terms in news summarization. In *Proceedings of 2015*

international conference on information and communication technology and systems, ICTS 2015 (pp. 95–98). https://doi.org/10.1109/ICTS.2015.7379878.

Purwitasari, D., Fatichah, C., Purnama, I. K. E., Sumpeno, S., & Purnomo, M. H. (2017). Inter-departmental research collaboration recommender system based on content filtering in a cold start problem. In *2017 IEEE 10th international workshop on computational intelligence and applications, IWCIA 2017—proceedings* (Vol. 2017-Decem). https://doi.org/10.1109/IWCIA.2017.8203581.

Purwitasari, D., Fatichah, C., Sumpeno, S., & Purnomo, M. H. (2018a). Ekstraksi Ciri Produktivitas Dinamis untuk Prediksi Topik Pakar dengan Model Discrete Choice. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *7*(4), 418–426.

Purwitasari, D., Ilmi, A. B., Fatichah, C., Fauzi, W. A., Sumpeno, S., & Purnomo, M. H. (2018b). Conflict of interest based features for expert classification in bibliographic network. In *2018 IEEE international conference on computer engineering, network and intelligent multimedia, CENIM 2018—proceedings*.

Purwitasari, D., Priantara, I. W. S., Kusmawan, P. Y., Yuhana, U. L., & Siahaan, D. O. (2011). The use of Hartigan index for initializing K-means++ in detecting similar texts of clustered documents as a plagiarism indicator. *Asian Journal of Information Technology, 10*(8), 341–347. https://doi.org/10.3923/ajit.2011.341.347.

Renoust, B., Melançon, G., & Viaud, M.-L. (2014). Entanglement in multiplex networks: Understanding group cohesion in homophily networks. In R. Missaoui & I. Sarr (Eds.), *Social network analysis—Community detection and evolution* (pp. 89–117). Cham: Springer. https://doi.org/10.1007/978-3-319-12188-8_5.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on web search and data mining* (pp. 399–408). New York, NY, USA: ACM. https://doi.org/10.1145/2684822.2685324.

Shibayama, S. (2019). Sustainable development of science and scientists: Academic training in life science labs. *Research Policy, 48*(3), 676–692. https://doi.org/10.1016/j.respol.2018.10.030.

Siciliano, M. D., Welch, E. W., & Feeney, M. K. (2018). Network exploration and exploitation: Professional network churn and scientific production. *Social Networks, 52,* 167–179. https://doi.org/10.1016/j.socnet.2017.07.003.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology, 31*(1), 361–395. https://doi.org/10.1111/0081-1750.00099.

Snijders, T. A. B., Lomi, A., & Torló, V. J. (2013). A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. *Social Networks, 35*(2), 265–276. https://doi.org/10.1016/j.socnet.2012.05.005.

Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks, 32*(1), 44–60. https://doi.org/10.1016/j.socnet.2009.02.004.

Steglich, C., Snijders, T. A. B., & Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology, 40*(1), 329–393. https://doi.org/10.1111/j.1467-9531.2010.01225.x.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476. https://doi.org/10.1002/asi.23596.

Tang, J. (2016). AMiner: Toward understanding big scholar data. In *Proceedings of the ninth ACM international conference on web search and data mining* (p. 467). New York, NY, USA: ACM. https://doi.org/10.1145/2835776.2835849.

Tang, J., Yao, L., Zhang, D., & Zhang, J. (2010). A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data, 5*(1), 2:1–2:44. https://doi.org/10.1145/1870096.1870098.

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. In *Proceedings of the 2007 seventh IEEE international conference on data mining* (pp. 292–301). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/ICDM.2007.30.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998). New York, NY, USA: ACM. https://doi.org/10.1145/1401890.1402008.

Wang, B., Bu, Y., & Huang, W. (2018). Document- and keyword-based author co-citation analysis. *Data and Information Management, 2*(2), 70–82. https://doi.org/10.2478/dim-2018-0009.

Wen, L., & Junping, Q. (2014). Semantic information retrieval research based on co-occurrence analysis. *Online Information Review, 38*(1), 4–23. https://doi.org/10.1108/OIR-11-2012-0203.

Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data, 3*(1), 18–35. https://doi.org/10.1109/TBDATA.2016.2641460.

Zhu, J., Zhang, J., Zhang, C., Wu, Q., Jia, Y., Zhou, B., et al. (2017). CHRS: Cold start recommendation across multiple heterogeneous information networks. *IEEE Access, 5,* 15283–15299. https://doi.org/10.1109/ACCESS.2017.2726339.

Affiliations

Diana Purwitasari[1,2] · Chastine Fatichah[2] · Surya Sumpeno[1,3] · Christian Steglich[4] · Mauridhi Hery Purnomo[1,3]

Chastine Fatichah
chastine@if.its.ac.id

Surya Sumpeno
surya@ee.its.ac.id

Christian Steglich
c.e.g.steglich@rug.nl

Mauridhi Hery Purnomo
hery@ee.its.ac.id

1 Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

2 Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

3 Department of Computer Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

4 Department of Sociology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands