Journal of Informetrics 14 (2020) 101040

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

An alternative topic model based on Common Interest Authors for topic evolution analysis

Sukhwan Jung [a], Wan Chul Yoon [b,*]

[a] Department of Knowledge Service Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
[b] Department of Computer Science, University of South Alabama, 307 N University Blvd, Mobile, AL 36688, United States

ARTICLE INFO

Article history:
Received 21 September 2019
Received in revised form 26 March 2020
Accepted 27 March 2020

Keywords:
Topic modeling
Bibliographic network
Topic evolution
Scientometric

ABSTRACT

Topic modeling methods aim to extract semantic topics from unstructured documents, and topic evolution is one of its branches seeking to analyze how temporal topics in a set of documents evolve and has shown successful identification of *content transitions* within static topics over time; yet, the inherent limitations of topic modeling methods inhibit traditional topic evolution methods from highlighting *topical correlations* between different, dynamic topics. The authors propose an alternative topic modeling method conscious of the topical correlation in the academic domain by introducing the notion of the common interest authors (CIA[1]), defining a topic as a set of shared common research interests of a researcher group. Publication records related to the Human Computer Interaction field were extracted from the Microsoft Academic Graph dataset, with *virtual reality* as the target field of research. The result showed that the proposed alternative topic modeling is capable of successfully model coherent topics regardless of the topic size with only the meta-data of the document set, indicating that the alternative approach is not only capable of allowing topic correlation analysis during the topic evolution but also able to generate coherent topics at the same time.

1. Introduction

Scientific research can be viewed as discovering new knowledge from the mine of knowledge. Scientific researchers write research articles to introduce new knowledge to the field of research, digging further from the mined shafts with high yield to maximize efficiency. Each publication of the research article expands the current limit of our knowledge; newly mined knowledge is connected to the background research already extracted in the vicinity. This makes understanding the boundary of research fields as an integral quality of any researcher, which is becoming more of a challenging task in this information-overloaded era; all researchers have limited information intake capability and most if not all research fields are swarmed with more research articles one can read at any given time, which is the equivalent of an individual miner unable to keep up with the current state of the massive network of mines. Topic modeling methods aid such individuals by mimicking the review process of human researchers, automatically extracting semantic themes, or topics, from a set of unstructured documents. Resulting topic models are word-cooccurrence based statistical models designed to extract latent semantic themes from a document collection with the use of natural language processing (NLP) methods. Research articles

* Corresponding author.
E-mail addresses: shjung@southalabama.edu, wcyoon@kaist.ac.kr (W.C. Yoon).
[1] CIA: Common Interest Authors

https://doi.org/10.1016/j.joi.2020.101040
1751-1577/

are summarised as a set of distinct topics within the domain knowledge to reveal the development of distinct research fields in a form of topic popularity, with the assumption that the existing topics mark the boundary of the research field. Results of topic modeling visualize the mine of knowledge by marking all shaft ends within the mine, outlining previously mined knowledge.

Understanding the boundaries of research fields is not sufficient for being a good researcher, however, as a result of a publication can vary greatly and there are multiple possibilities of a new knowledge having low to no impact due to various reasons. Knowledge outside the boundary could be unexplored because previous mining attempts have failed, existing shafts leading to the location could have become obsolete while competing against other branches of the field, or the expected yield from it holds no practical and theoretical importance to the field. Knowing the productivity and popularity of knowledge from the mining shaft leading up to the boundary is, therefore, also crucial for researchers to produce good quality research publications. Topic evolution extracts and compares topic models in subsequent years to analyze the evolution of topics over time, providing up-to-date changes in research trends. Topic evolution in research publications allows researchers to automatically extract changes in a topic, or a branch of research, in the form of changes in the word vector representing the topic over the years.

Unlike mining materials, however, finding knowledge is more complicated. More often than not, new knowledge requires the amalgamation of multiple different branches of existing knowledge with varying degrees, from as low as algorithmic variants to as high as incorporating wholly different knowledge domains. The mines of knowledge are ultimately all interconnected with a multitude of complex interactions. Traditional topic evolution approaches are limited in understanding such behaviors as inherent limitations borne from the topic modeling techniques hinder the comparison of multiple dynamically evolving topics. Firstly, they suffer the limitation of NLP methods imbedded into the topic modeling methods. Their inherent need for the large corpus makes the existing techniques useful in predicting the overall trend of the field but makes them inefficient where available data is scarce. They also suffer from ambiguity, language variability, synonyms, grammatical errors and other problems (Resnik, 1999). Secondly, topic evolution analysis based on a language model similarity measure is limited to the *content transition* within a single strand of similar topics. Words and their popularities are the whole constructors for topics in traditional approaches, and this forces the language model similarity measure to be used to link topics over time for topic evolution analysis. The language similarity measure is an overlapping measure where one topic can be similar to many topics with an undefined similarity sum with which the ratio of topic transference comparison cannot effectively be made. This causes only a single interpretation of the similarity value available, where a set of topics from a document collection is bound to be disparate. This representation is however inaccurate in ascertaining an actual set of topics shared by the members of the community, where topics are distinguished not by the representative labels of the topics but by their goals and predecessors. Utilizing a single similarity measure for topic comparison causes other hinderances, as well. The traditional approach cannot distinguish changes in the representative terms of a topic from changes in the topic itself as both are measured with a single similarity value. Such lack of distinction is overlooked in the traditional topic evolution where the content transition is the primary concern, as the two are bound together in a topic comparison. This is not the case for the more advanced evolutionary events such as merge and split, however, where the changes in the representation of an ongoing topic and the changes in the topic itself can occur separately. The introduction of the author-based topic modeling method allowed the proposed method to bypass such limitations of traditional text-based topic modeling approaches and enabled enhanced topic evolution analysis without needing to address the issues such as word sense disambiguation. The proposed method allowed such distinction by providing an additional author-based similarity measure, conscious of the topic history and is independent of the topic words.

The authors proposed an alternative topic modeling using author-to-author links from bibliographic networks. Bibliographic graphs represent how research efforts are connected to build a research domain, including co-author, co-citation, bibliographic coupling, author-citation, and co-word links each representing collaboration effort between researchers, recognition of shared research topics between researchers, coupling strength in the shared research topics between researchers, acknowledgment of researchers in the research topics, and research topic similarities between researchers. The authors proposed a common interest author group (CIA group), a group of closely related researchers with direct and indirect research collaborations formed around common research interests, specific visions, directions, and purposes. Such themes of CIA groups change over time as members' common research interests and their membership evolve. Temporal representations of CIA groups generate topic models, or CIA topics, which is a set of commonly used keywords among the group members.

Fig. 1 shows a graphical overview of the proposed method. The proposed method starts by filtering keyword-specific active authors from the dataset by their relevance and connectivity with a series of user-input measure thresholds. The system then lets the user view *co-authors*, *author-citations*, *co-citations*, *bibliographic coupling*, and *co-word* networks between selected authors to allow interactive multigraph clustering. The resulting list of disconnected author subgraphs, or CIA groups, are distinctive sub-networks of the researchers connected by different publication activities. CIA topics represent CIA groups and are extracted from the meta-data of the time-constrained publications written by the CIA group members in the form of a word-frequency vector. The CIA topics capture the common research interests shared by a group of authors, which are not necessarily bound by a specific set of words. Topic modeling by the CIA group enables correlation analysis for dynamically changing research interests as CIA groups act as a transitional medium between CIA topics, where the traditional document-based topic evolutions only detect transitions in the representation of pre-defined topics.

Section 2 reviews the related works on topic modeling with regard to topic evolution as well as background research for the proposed method. Section 3 details the proposed method and experimentation where the proposed method was tested

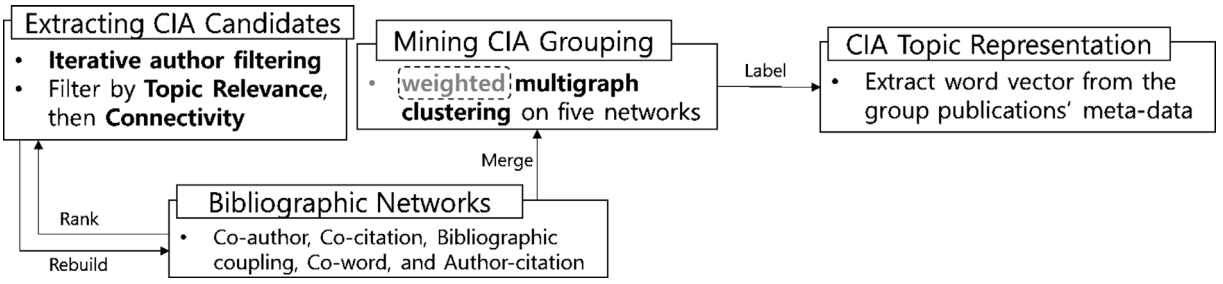

Fig. 1. An overview of the proposed method.

on 712k HCI articles from the Microsoft Academic Graph dataset, with 170k authors and 303k keywords connected to them. The experiment results proved that it is feasible to generate coherent CIA topic models with the proposed method while providing an author-based topic similarity measure, as shown in section 4.

2. Related work

2.1. Topic modeling for topic evolution

Topic modeling provides a statistical approach to discovering *topics*, which are latent semantic structures occurring within a given corpus in the form of word and popularity clusters based on the statistical distribution and co-occurrences of words. Latent Dirichlet Allocation(LDA) (Blei et al., 2003) is one of the oldest and simplest topic models where latent topics within a document collection are learned by iteratively assigning word-topic links with word co-occurrences between documents. Topics, defined as word distributions over a corpus dictionary, are then assigned to each document (Steyvers & Griffiths, 2007). This leads to a three-level hierarchy; a document consists of a set of topics each built with a set of words. The majority of topic modeling methods are extensions of LDA model. An author-topic model (Rosen-Zvi et al., 2004) is an extension of the LDA model randomly distributing words to authors to generate an author model which is then combined with the topic model to generate an author-topic model. Bernoulli process topic (BPT) (Guo et al., 2014) is a framework utilizing LDA, where the document collection is extended with the citation relationship from the original corpus. On top of the authorship and citation relationships, studies in the field also include methods to filter out only useful technologies with patent co-citation frequencies (Kay et al., 2014), acronym usage patterns and co-authorship of research papers. There are several topic modeling approaches depending on the network structures of texts, where the topic modeling task becomes the identification of clusters from complex networks built from the input documents. Document-focused topic modeling is built around the citation relationship between documents (Silva et al., 2016), while the word-focused topic modeling is built from various word relationships such as word-adjacency (Amancio, 2015), syntactic similarity (Ferrer & Cancho et al., 2004), and semantical similarity (de Arruda et al., 2016). The word embedding is a different approach to language modeling, assigning numerical context to words (Levy & Goldberg). Numerical vectors are assigned to the unique words in the given document collection and the words positioned close to one another in the multi-dimension space are defined to have similar semantic meaning or context. The probabilistic language models such as Word2Vec were enhanced by the recent development of neural networks, outperforming its n-gram based predecessors (Bengio et al.; Mikolov et al.; Mikolov et al., 2013). These topic modeling approaches successfully modeled known topics with high performances while sharing the same limitation of traditional text-based topic modeling; they cannot distinguish changes in the representative terms of a topic from changes in the topic itself as their topic models are only represented by words.

Topic evolution is a research field focusing on identifying the evolution of topics from a sequentially ordered document collection. Dynamic topic models (Blei & Lafferty, 2006) is one of the initial approaches to topic evolution focusing on identifying content transition of a chained topic from a corpus with fixed timeslots. *K* dynamic topics are chained over the time slices, showing approximated changes in word frequencies. An iterative topic evolution learning framework (He et al., 2009) leveraged citation relationships between documents in order to better understand the evolution of topics. Technology forecasting (Porter & Detampel, 1995) is a field of research aiming to predict the characteristics of technology in the future. Various techniques from simple extrapolation to organization management (Battistella, 2014) and fuzzy NLP (Newman et al., 2014) are used to identify and predict changes in technology indicators(Bongers & Torres, 2014). Studies in the field also include methods to filter out only useful technologies with patent co-citation frequencies (Kay et al., 2014), acronym usage patterns, and co-authorship of research papers.

Topic modeling conscious to the time constraints utilize non-discretized time intervals instead (Castano et al., 2017); Topic Detection and Tracking (TDT) (Fiscus & Doddington, 2002) is a multi-site research project aiming to predict and follow novel topics in a stream of text data (Allan et al., 1998), where the data size and delay intolerance resulted in a demand for memory-efficient text clustering strategy (Zhong, 2005). Manually assisted technology trend analysis was done to identify the roots of new technologies with their projected impacts (Segev et al., 2013), while the semantic smoothing model is extended to enhance the text clustering quality from the stream of large-scale texts (Liu et al., 2008). A similar approach was

tried with multiple data sources to predict technology trends while showing that different data sources exhibit different forecast speeds (Segev et al., 2015), while inter- and intra- text relationships in two news sources are analyzed to identify changes in the popularity of topics both local and common topics (Hong et al., 2011).

Burst term detection is a form of time-dependent topic modeling approach under the assumption that the emergence and dissipation of topics are signaled by a "*burst of activity*" (Kleinberg, 2003); the infinite-state automaton over a continuous data stream captures the transitions between low-activity and high-activity state words, enabling the temporal analysis of word trends. Burst term detection, in conjunction with keyword co-word analysis, allows multi-dimensional exploration of the research front in question (Li & Chu, 2017). Identifying burst patterns in co-occurrences, topics over time (TOT) jointly model word and timestamp distribution over non-discretized time resulting in a beta distribution of topics over time (Wang & McCallum, 2006). These methods detect topic evolution based on the changes in the projection of connected topics over time, or the changes of the word co-occurrence patterns within time-spanning topics. While they are capable of analyzing the changes in the topic trends, they are not suitable to analyze more complex topic evolution events such as merge and split as the topics are analyzed independently.

The evolutionary relationships between topics over time are analyzed in order to capture simultaneous merge and split. The topics are formed as content-based document clusters, and evolutionary transitions are detected when the content dissimilarity between clusters across different time intervals stays below a threshold (Gaul & Vincent, 2017). Availability to multiple connections between topics allowed merge and split detection, but the nature of distance measure prohibited the detection of the topic correlation ratio between themes (Mei & Zhai, 2005). One of the more recent approaches to the merge and split detection in topic evolution identifies topic correlations with two sets of topics with LDA (Chen et al., 2017); corpus-level topics as time-spanning global topics and temporal topics as timeslot-specific local topics. Local topics at each timeslot are linked to the global topics with cosine similarity measure and merging and splitting of global topics are defined as decreased and increased number of local topics connected to the same global topics in the future. Cross-citations between topic pairs' member documents were also tried for connecting topics over time, also limited to content transition identification (Jo et al., 2011). A similar approach was tried with the word embedding approach. Dynamic embedding based on the generalized word embedding analyzed the changes in the semantic usage of words over time within a given corpus (Rudolph & Blei, 2018) by linking word embedding in different time slices with static *context vectors*. Both approaches were able to detect the topical transition over time, capturing how the representations of given topics evolve over time. They are, however, inefficient in tracking more complex topic evolution events such as merging and splitting. While these approaches allow the detection of merging and splitting of time-spanning topics and their transitional ratio at the specific temporal time, the use of the predetermined global topics prohibits 1) merging and splitting identification for time-spanning topics, and 2) introduction of new topic chains over the timeslots. The same limitation applied to the topic prediction based on the topic co-occurrence network based on the LDA topic models (Menenberg et al., 2016), where changes in the type of the predefined topics were predicted.

2.2. Bibliographic networks

Bibliographic data analysis allows researchers to understand the past and current trends of research topics. One such data is a bibliographic network, where the nodes denote entities such as authors and publications and links denote relationships between them (Batagelj & Cerinšek, 2013). The co-author network and citation network are among the most common types of bibliographic networks dealing with directly stated relationships, each representing different types of academic interactions. Co-author network represents the scientific collaboration between researchers and citation network represent the endorsement of scientific research, respectively shown as the links between authors and publications (Ding, 2011). Collaboration and endorsement are considered major topics in bibliometrics, and there are a number of research utilizing both networks for citation-aided author clustering (Papalexakis et al., 2013), citation-weighted co-author network generation (Börner et al., 2005), bibliometric index incorporating co-authorship into traditional h-index based on citation counts (Ausloos, 2012), and so on.

Similarity-based bibliographic networks, on the other hand, are a set of networks derived from direct links between authors and publications, where various types of author similarities are shown as links between authors. Author co-citation network (White & McCain, 1998), bibliographic coupling network (Kessler, 1963), and co-word networks (Callon et al., 1983), each represent the different type of similarities between authors. A co-citation network represents author pairs endorsed in the same publications as similar authors with a frequency of common endorsement between two authors as link weight, whereas a bibliographic coupling network represents an inverse relationship between authors representing common endorsement given by author pairs as links. Co-word network is one of the more heavily connected author networks where author pairs sharing common keywords on their publication records are linked together. Such networks based on co-occurrences are generally analyzed for the identification of the indirect multidisciplinary relationship between authors (Yan & Ding, 2012).

2.3. Multigraph clustering

Multigraph clustering focuses on subgraphs extraction from multiple graphs. A multigraph or multi-layer graph is a type of graph that allows multiple edges between any pair of nodes, with different types of links forming a layer of the multigraph

over common entities often represented by a shared set of nodes. A more sophisticated and rare type of multigraph involves layers of graphs each with different nodes that need to undergo a node mapping process to connect the layers (Kim & Lee, 2015), which is not discussed in this paper.

Modularity based community structure in multiscale/multiplex networks is found by linking each graph slices with inter-slice edges and run conventional community detection algorithms (Mucha et al., 2010). Communities found at each one-dimensional graphs are then used to form meta-communities, and similarities between each dimension are measured by information variation between meta-communities (Rocklin & Pinar, 2013). Different characteristics in different dimensions are ill-considered in both methods, however, as link sparsity of multigraph layers do not necessarily correlate to their information value regarding the underlying common community structure (Paul & Chen, 2016). Link Matrix Factorization model is proposed where each graph dimension is approximated by a factor each connected – or linked – by a common vector (Tang et al., 2009), and a probabilistic generative model based on the variational Bayesian approach generated is used to match genes to the metabolism category (Shiga & Mamitsuka, 2012). Inspired by the spectral clustering algorithm, joint matrix factorizations and graph regularization are used to build joint eigenvectors of graph Laplacian matrices on which k-means clustering algorithms are used (Dong et al., 2012). This is later expanded to incorporate Grassman manifold theory to add subspace analysis capability to multigraph clustering where the influence of the relationships between the individual graph layers can be differentiated (Dong et al., 2014). Traditional clustering algorithms can also be applied to multigraphs instead of purposefully designed multigraph clustering algorithms by first flattening it into a single graph. Differential flattening (Kim & Lee, 2015) utilizes multi-objective optimization to maximize a clustering coefficient of the flattened graph on which multiple single-layer clustering algorithms were run. This approach, while introduces higher performance, suffers a scalability problem.

Frequent Subgraph Mining aims to extract frequently occurring patterns within a number of smaller graphs or within one larger graph. The method starts by assigning a single vertex to a pattern. Each pattern is matched across the list of graphs, and patterns with frequency less than the minimum support – or frequency – threshold are discarded. Each iteration expands the remaining patterns by replacing a pattern with all possible combinations of a pattern and one of its neighboring nodes, and iterations end once all possible patterns have been viewed, or no expansion is possible (Cook & Holder, 1993). The research field stemmed from the well-studied frequent pattern mining techniques, but experience higher overheads because of the graph representation of input data; candidate generation is harder because of the exponential number of subgraph combinations available, and isomorphism checking is also much more resource-intensive. A vertical search scheme is proposed to identify and reduce the number of redundant candidates proposed at each iteration (Huan et al., 2003). Edge weight-based subgraph importance measure is also proposed, but this could lead to the patterns having a non-monotonic property where a pattern could be not frequent while its super-graphs are, requiring further research on the iteration process (Jiang et al., 2010). Quasi-clique based subgraph mining is proposed to deal with variable support values, overcoming limitations of the traditional total quasi-clique method where the quasi-clique much be found from all multigraph layers to be distinguished. Variable support allows partial subgraph matching from the layers instead.

Tensor decomposition aims to reduce multi-way indices into the elementary operations between simpler forms of tensors. Its applicability to various fields such as machine learning, signal processing, and natural language processing led it to be one of the highly sought-after research fields even with the NP-hard nature of some of its tasks which are overcome by series of heuristics and guided disciplines (Sidiropoulos et al., 2017). Implementing spectral algorithms for tensor decomposition resulted in polynomial-time computations for sparse dictionary learning problem, as fast as the non-tensor based approaches while being more robust and easy to implement (Schramm & Steurer, 2017). Application of tensor decomposition extended to the multilayer clustering in the resent research aiming to deal with the noisy real-life networks by differentiating groups of network layers sharing common structures (Chen et al., 2019). The rank-terms decomposition is used to identify partitions of network groups sharing the common node structures embedded in the adjacency tensors, resulting in superior effectiveness and robustness for both the real-world networks and synthetic datasets compared to existing tensor-based single-layer clustering methods. The symmetric non-negative factorization can be extension three-way data optimizing multigraph clustering into a single process (Zhang et al., 2013).

All multigraph clustering methods mentioned in this section are all applicable to the proposed method. Pattern mining approach is used in the paper as it suffers lessor scalability problems on multigraphs with an identical node set; layer-spanning nodes do not require inter-layer connections eliminating the need for isomorphism checking.

3. Methods

he proposed method aims to identify topic models based on common interest authors group (CIA group) within a user-specified field of research and is divided into three parts. 1) Extracting CIA candidates filters out authors with little to no impact on the targeted research field by identifying a lack of active contributions and collaborations to the field. 2) Mining CIA groups run multigraph clustering on five bibliographic networks built with the candidate authors in order to identify a series of CIA groups that are distinctive sub-networks of researchers connected by multiple publication activities. 3) CIA topic representation labels such author groups with their shared research interests with the meta-data of the time-constrained publications written by the CIA group members in the form of a word-frequency vector. Topic modeling by CIA group enables correlation analysis for dynamically changing research interests as CIA groups act as a transitional medium between CIA topics, where traditional document-based topic evolutions only detect transitions in the representation of pre-defined

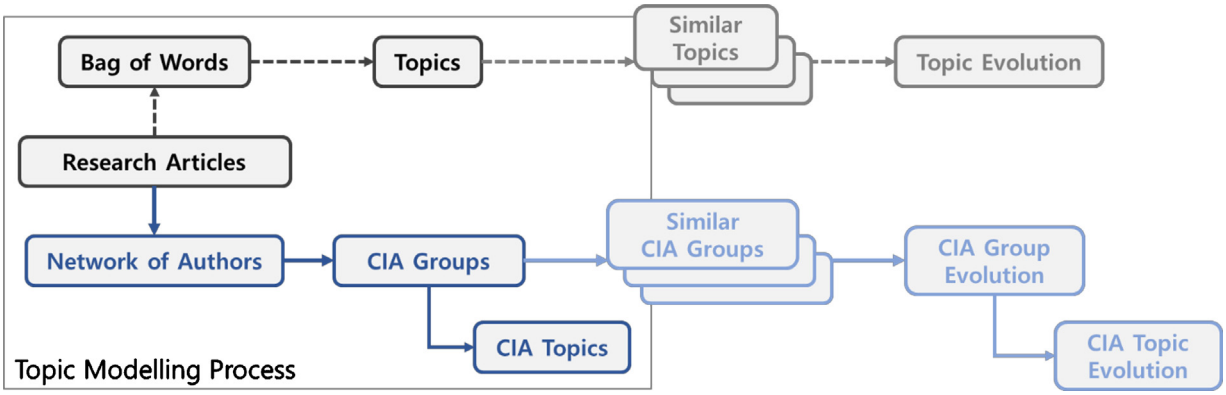

Fig. 2. Comparison of an alternative topic evolution approach from the proposed method (top) against the traditional topic evolution approach (bottom).

Table 1
Node labels used in initial database.

Label	Count	Attribute	Attribute Description
(:Paper)	712,228	pId	The publication identifier from MAC dataset
		title.	Title of the publication
		year	Year of the publication
(:Author)	170,060	name	Name of the author
		citeCount	Citation count of the given author
		hIndex	hIndex of the given author
(:Word)	303,045	label	The label of the word used in either keywords or FOS field

Table 2
Edge types used in the initial database.

Type	Count	Description
(:Author) –[:AUTHORS]->(:Paper)	346,304	Authoring relationship between authors and publications
(:Paper)-[:CITES]->(:Paper)	1,935,659	Citation relationship between publications
(:Paper)-[:HAS_KEYWORD]->(:Word)	1,329,424	Connection between Publications and their keywords

topics. Fig. 2 shows the overall generation process of the traditional and the proposed topic modeling methods and how topic evolutions are analyzed on their results.

3.1. Dataset

Microsoft Academic Graph (MAG) was extracted to prepare an experiment dataset. MAG is a heterogeneous publication graph created by Microsoft containing over 210 million publications and 254 million authors. Re-launched in 2016, MAG is a relatively new academic publication search engine but is deemed competitive to existing search engines such as Google Scholar and Scopus (Hug et al., 2017). MAG contains publication records, citations between them, and other meta-data such as their keywords, venues, institutions as well as fields of study[2] . The field of study is a categorization of publication records built from an iterative graph link analysis and entity filtering which are confirmed by scanning through there meta-information such as title, keywords, and abstracts. Currently, the MAG dataset holds 229,634 Fields of Study labels.

Neo4j was used to store the extracted dataset. Neo4j is an open-source graph database management system focused on relationships between entities and is considered the most popular graph database according to DB-Engines ranking of graph DBMS[3] . Data is stored in the form of nodes with labels and edges with types, each with attributes to store relevant information. A sample relationship (:Person {name: 'Gildong')-[e:LIKES]->(:Food {ingredient: 'Rice') says a *person* named Gildong *likes* Rice-based *food*. Table 1 and Table 2 show the node labels and the edge types used to represent the HCI-related publication records.

While identifying different authors shown as the same name and collating multiple names as a single author both positively affect the quality of citation network analysis, author disambiguation is not done on the dataset for the following reasons. The lack of known author clusters in the training set renders the unsupervised learning approach inappropriate to the general bibliographic dataset. The unsupervised learning approach is also not generalizable as it is sensitive to the

[2] https://academic.microsoft.com/
[3] https://db-engines.com/en/ranking/graph+dbms

domain of the research field in question, from the co-authoring pattern to the ratio of specific nationality of authors (Louppe et al., 2016; Strotmann & Zhao, 2012; Torvik & Smalheiser, 2009).

3.2. Extracting CIA candidates

Let $A \in a$ as a set of authors, $P \in a$ as a set of papers, P^y as a set of papers at year y, and K as a union set of keywords and field of studies. The series of graphs shown below represent publication records for authorship, keyword usage, and citation. $G^t = (U, V, E)$ denotes a bipartite graph G with type t, where U is a source node set and V is a target node set. $E = (u, v, w)$ denotes the edge set from $u \in U$ to $v \in V$ with edge weight w.

- $G^{author} = (A, P, E^{author})$ where $E^{author} = \{a_i, \mathrm{p}_j, w_{i,j}^{author} | a_i \in \mathrm{A}, \mathrm{p}_j \in P, 0 \leq w_{i,j}^{author} \leq 1\}$.
- $G^{keyword} = (P, K, E^{keyword})$ where $E^{keyword} = \{p_i, k_j | p_i \in \mathrm{P}, k_j \in K\}$.
- $G^{cite} = (P, P, E^{cite})$ where $\mathrm{E}^{\mathrm{cite}} = \{p_i, p_j | p_i, p_j \in P\}$.

A CIA group is a collection of researchers closely working together on the subtopics within the given research field. The authors should be relevant to the target research field in order to generate meaningful topics, and they should have well-established relationships with other researchers to show they work together. Both relevance and connectivity of the field-related authors were considered in identifying candidate authors for such groups as a result.

Field-related papers $P_y^{related}$ are any publications made within a buffered period *[y-b, y]* that received non-zero citations from any year, with the given research field q as one of its keywords or FOS. Active authors A_y^{active} are their authors who have also published in that period *[y-b, y]*, who contributed to the field by publishing P_y^{active} and receive recognition through citation from $P_y^{related}$ during the time. The outcome is a list of authors A_y^{active} who are core to the given research field at year y.

With query term q, year y, and buffer b, field-related publications and active authors can be shown as

$$P_y^{related} = \left\{ p_i | \exists p_j \in P, \exists p_i \in P_{[y-b,y]}, k_l = q, e_{i,j}^{cite}, e_{j,l}^{keyword} \right\}$$

$$A_y^{active} = \left\{ a_i | \exists a_i \in A, \exists p_j \in P_y^{related}, \exists p_l \in P_{[y-b,y]}, e_{i,j}^{author}, e_{i,l}^{author} \right\}$$

$$P_y^{active} = \left\{ p_i | \exists p_i \in P_{[y-b,y]}, \exists a_j \in A_y^{active}, k_l = q, e_{j,i}^{author}, e_{i,l}^{keyword} \right\}$$

Researchers relevant to the research field actively work in the field, referencing field-related publications as related works while receiving noticeable amounts of recognition from colleagues as well. Field relevance $a_i^{relevance}$ of an author a_i at year y is measured by the weighted sum of three measures; frequency of papers authored (activeness), citations received (importance), and citations given (cohesion).

$$a_i^{relevance} = \sum\nolimits_{n \in I} w_n^{relevance} \times Filter_n^{relevance}(a_i) \; where \; I = \left\{ activeness, \; importance, \; cohesion \right\} \; and \sum\nolimits_{n \in I} w_n^{relevance} = 1$$

$$Filter_{activeness}^{relevance}(a_i) = \exists a_i^{active} \exists p_j^{active} \forall e_{i,j}^{author} \left(\left| p_j \right| \right)$$

$$Filter_{importance}^{relevance}(a_i) = \exists a_i^{active} \exists p_j^{active} \exists p_l^{active} \forall e_{i,j}^{author} \left(\left| e_{l,j}^{cite} \right| \right)$$

$$Filter_{cohesion}^{relevance}(a_i) = \exists a_i^{active} \exists p_j^{active} \exists p_l^{active} \forall e_{i,j}^{author} \left(\left| e_{j,l}^{cite} \right| \right)$$

H-index is a widely used author metric measuring both the author's productivity and academic impact. It is defined as the maximum number of publications h an author has published that has at least h citations. Similar measures are used to measure the connectivity of active authors measuring the maximum number of neighbors n an author has on a bibliographic network that has at least n frequencies. Connectivity $a_{i,y}^{connectivity}$ of an author a_i at year y is measured by the weighted sum of neighbor size and the H-index-like measure from bibliographic networks aiming to identify those with a large number of frequently connected colleagues. H-index-like measure for active authors in each bibliographic graph is calculated with the following formulas.

$$a_i^{connectivity} = \sum\nolimits_{n \in I} w_n^{connectivity} \times Filter_n^{connectivity}(a_i) + \sum\nolimits_{n \in N} w_{n,Hindex}^{connectivity} \times Filter_{n,Hindex}^{connectivity}(a_i)$$

Table 3
Types of Bibliographic Networks Used.

Bibliographic network	Property represented by the network
Co-author	Collaboration efforts between researchers.
Co-citation	Recognition of sharing research topics between researchers.
Bibliographic Coupling	Coupling strengths in the shared research topics between researchers.
Author-citation	Acknowledgments of researchers to the research topics.
Co-word	Similarities of research topics between researchers.

Table 4
Definition of the edge set for five bibliographic networks.

Edge set	**Definition**
E_y^{coAuth}	$\left\{\left(a_i, a_j, w_{i,j}^{coAuth}\right) a_i, a_j \in A_y^{active}, \forall p_l e_{i,l}^{author} e_{j,l}^{author}, w_{i,j}^{coAuth} = \lvert p_l \rvert\right\}$
E_y^{coCite}	$\left\{\left(a_i, a_j, w_{i,j}^{coCite}\right) a_i, a_j \in A_y^{active}, \forall p_l e_{i,l}^{author} \forall p_m e_{j,m}^{author} \exists p_n e_{n,l}^{cite} e_{n,m}^{cite}, w_{i,j}^{coCite} = \lvert p_m \rvert\right\}$
$E_y^{bibCouple}$	$\left\{\left(a_i, a_j, w_{i,j}^{bibCouple}\right) a_i, a_j \in A_y^{active}, \forall p_l e_{i,l}^{author} \forall p_m e_{j,m}^{author} \exists p_n e_{l,n}^{cite} e_{m,n}^{cite}, w_{i,j}^{bibCouple} = \lvert p_m \rvert\right\}$
$E_y^{authCite}$	$\left\{\left(a_i, a_j, w_{i,j}^{authCite}\right) a_i, a_j \in A_y^{active}, \forall p_l e_{i,l}^{author} \forall p_m e_{j,m}^{author}, w_{i,j}^{authCite} = \left\lvert e_{l,m}^{cite} \right\rvert\right\}$
E_y^{coWord}	$\left\{\left(a_i, a_j, w_{i,j}^{coWord}\right) a_i, a_j \in A_y^{active}, \forall p_l e_{i,l}^{author} \forall p_m e_{j,m}^{author} \forall k_n e_{l,n}^{keyword} e_{m,n}^{keyword}, w_{i,j}^{coWord} = \left\lvert k_m \right\rvert\right\}$

where $BN = \left\{coAuth,\ bibCouple, authCite,\ coCite,\ coWord\right\}$ and $\sum_{n \in N} w_n^{connectivity} + \sum_{n \in N} w_{n,Hindex}^{connectivity} = 1$..

$$Filter_n^{connectivity}(a_i) = \exists a_i^{active} \forall a_j^{active} \sum_j \left| e_{l,j}^n \right| Filter_{n,Hindex}^{connectivity}(a_i) = \max_j \min\left(rankedCount_n(a_i, j), j\right)$$

where $rankedCount_n(a_i, j) = \exists a_i^{active} \exists a_j^{active} \left| e_{l,j}^n \right|,\ rankedCount_n(a_i, j) \geq rankedCount_n(a_i, j+1)$

Filtering options are used to filter A^{active} by applying per-option weight and overall author percentage threshold.

$$A_y^{active} = \left\{ a_i | \exists a_i \in A_y^{active}, a_i^{filterType} \geq a_{i+1}^{filterType}, i \leq round\left(\left|A_y^{active}\right| \cdot AuthorPercentageThreshold\right)\right\}$$

$$P_y^{active} = \left\{ p_i | \exists p_i \in P_y^{active}, \exists a_j \in A_y^{active}, e_{j,i}^{author} \right\} \text{ where } filterType = \{relevance, connectivity\}$$

Bibliographic networks represent different types of relationships between authors. While not directly representing a topical similarity between authors, these networks contain information on the shared research topics as authors connected within such networks are connected through the research activities they represent. Five different bibliographic networks were used to represent different aspects of the author relationships. Table 3 shows the types and represented properties of bibliographic networks used in the proposed method, and Table 4 includes definitions for the edge set of the five bibliographic networks $G_y^n = \left(A_y^{active}, E_y^n\right),\ n \in N$.

Author relevancy and connectivity are independent of each other and can be used to filter CIA candidates with different perspective, allowing variable weight parameters to tailor the filtering process to the specific research patterns in the given domain; for example, number of publications would be more meaningful in the mathematical theory field where publications are more scares, while the low tendency for co-authoring result in such relationship result in co-authorship being more important than other fields. While the process is open for iterative filtering with varying parameters, however, the experiment is done to validate the possibility of topic modeling based on the relationship between common interest authors and its application for topic evolution. Parameters are not fitted to the target research field, and the default, uniform values were used instead. Extracting CIA candidates was done sequentially, first by relevance $a_i^{relevance}$ then by connectivity $a_i^{connectivity}$. Equal weights were used in the experiment to both relevance and connectivity measures, 1/3 for Activeness, Importance, and Cohesion each and 1/10 for degrees from the five networks and their H-index variants each. 50 % of authors were filtered out at each iteration, resulting in 25 % of authors remaining. Table 5 shows the experimental conditions used for the experiment.

3.3. Mining CIA groups

A multigraph G_y^{multi} was generated by merging all bibliographic networks for the common set of nodes A_y^{active}.

$$G_y^{multi} = (A_y^{active}, \{E_y^{coAuth}, E_y^{authCite}, E_y^{coCite}, E_y^{bibCouple}, E_y^{coWord}\})$$

Table 5
List of experimental conditions used for Extracting CIA candidates.

Condition	Value	Reasoning
Starting year	1998	Some of the recently born research fields such as 'Big Data' start to disappear from the publication records over 20 years.
Finishing year	2018	It is the most recent year available in data.
Year buffer	10	Selected allow smooth topic transition between snapshots.
No Inactive authors	True	An author must have non-zero publications during the timeslot.
No Irrelevant authors	True	An author must have non-zero citations received during the timeslot.
Yearly Citation threshold	0	Any citation at all validates the relevance of authors.
Bibliographic Networks Used	List	$I = \{coAuth,\ bibCouple,\ authCite,\ coCite,\ coWord\}$
Author Relevance Weight	List	$w^{relevance} = \{0.\dot{3}, 0.\dot{3}, 0.\dot{3}\}$.
Author Connectivity Weight	List	$w^{connectivity} = \{0.1, 0.1, 0.1, 0.1, 0.1\}$, $w^{connectivity}_{Hindex} = \{0.1, 0.1, 0.1, 0.1, 0.1\}$.
Author Percentage Threshold	0.5	Half of the authors were filtered out at each iteration.

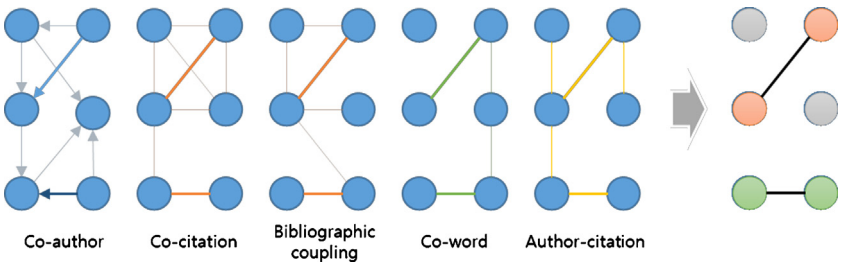

Fig. 3. Visualization of frequent subgraph mining with 100 % support used on the proposed method.

Table 6
List of experimental conditions used for Mining CIA Groups.

Condition	Value	Reasoning
Minimum CIA group size	3	Subgraphs with 1 or 2 authors as their members were considered too small and were discarded
Bibliographic Networks Used	List	$\{coAuth,\ bibCouple,\ authCite,\ coCite,\ coWord\}$

Multigraph clustering was done with a total overlap policy, a frequent pattern mining approach with 100 % support. Total overlap policy used on G_y^{multi} allowed filtering of non-overlapping links, resulting in a non-directional graph G_y with one edge type, as shown in Fig. 3.

$$G_y = \left(A_y^{active}, E_y^{overlap}\right)\ where\ e_{i,j}^{overlap} = \left\{ a_i, a_j, \sum_m \frac{w_{i,j}^m}{\max(w^m)} w_{i,j}^m > 0 \right\}$$

Resulting graph G_y is a disconnected graph with a number of connected sub-graphs each representing a distinctive collection of authors working heavily with each other. Size discrepancies are modulated by dividing the largest sub-graph with a Louvain clustering algorithm and discarding sub-graphs with node size smaller than the minimum CIA group size. Each of the resulting subgraphs is designated as a CIA group. Table 6 shows the experimental conditions used for the experiment.

Fig. 4 shows the difference between CIA groups and clusters from a co-author network. Dots in the graph represent authors positioned by the OpenOrd layout algorithm (Martin et al., 2011) with co-authorship clusters, where authors within each cluster are visually clustered together. Red lines indicate connections within CIA groups, which included many long lines indicating the presence of inter-cluster edges within CIA groups. This was because CIA groups were built with multi-dimension research activities. There were however many CIA groups found within clusters as well. Fig. 5 shows randomly selected CIA groups over the same author layout with colored nodes represent authors in CIA groups by color, revealing that many of the authors in the same CIA group were positioned near each other where the author positions were decided based on the clustering result from co-author network.

3.4. CIA topic representation

A CIA topic is a labeling of the common research interest shared by members of a CIA group. Keyword and FOS usage frequencies of a CIA group were used to build a CIA topic for this research. This was to ensure that the format of CIA topics followed the format of topics from the traditional methods while maintaining the proposed method free from the use of natural language processing. Both the keyword and FOS are specific terms designed to represent the given research fields

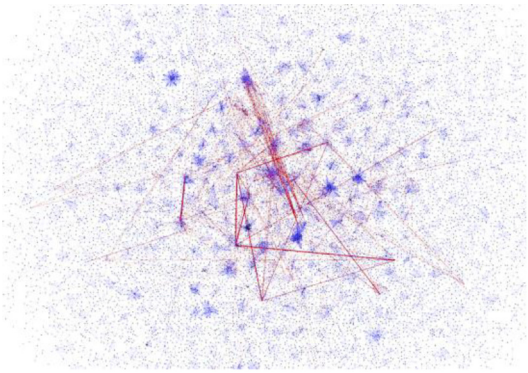

Fig. 4. Visualization of author links within CIA groups on a clustered co-author network.

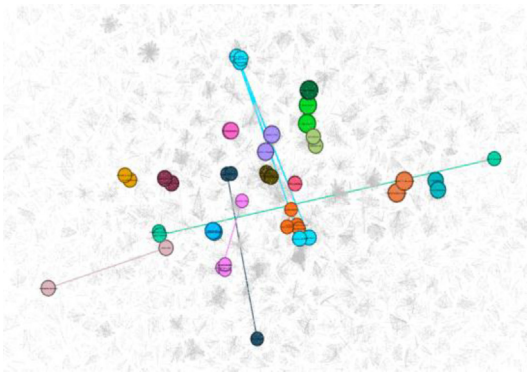

Fig. 5. Visualization of 18 CIA group samples on a clustered co-author network.

Table 7
List of experimental conditions used for CIA Topic Representation.

Condition	Value	Reasoning
Number of CIA topics	10	A preliminary result showed top ten topics comprise more than 80 % of the CIA group members.
Number of words per topic N	(Ausloos, 2012; Blei & Lafferty, 2006; Dong et al., 2012)	# of Words per a CIA topic, ten as the standard value, added half/double values for comparison.

and topics, removing the necessity of topic word extraction. Many of the language ambiguities borne from sentences such as syntactic, discourse, or pragmatic ambiguities (Anjali & Babu) are disregarded because of the structured nature of input data, while semantic ambiguity is not addressed as the absence of text-based modeling in the proposed method. The words are not used in the modeling process hence no word sense disambiguation is required to distinguish words with different contexts (Resnik, 1999); CIA topics are modeled by the CIA groups and the words are only used to represent their topics in the textual format. A CIA topic aims to identify the collective interest of a group not a sum of the group members, hence any publication authored by multiple members of the group is only counted once to remove duplicates. Table 7 shows the experimental conditions used for the experiment.

For *i*th CIA group in year *y* $ciaGroup_y^i$, CIA topic $topic_y^i$ is calculated as

$$topic_y^i = \left\{ w_y^{i,k}, f_y^{i,k} | w_y^{i,k} \in words, p_j \in P_y^i, f_y^{i,k} = \left| \exists e_{j,k}^{keyword} + \exists e_{j,k}^{fos} \right| \right\} where f_y^{i,l} \geq f_y^{i,l+1} > 0$$

where the document set of the CIA group P_y^i is defined as

$$P_y^i = \left\{ p_j | p_j \in P_{y-b\sim y}, a_k \in ciaGroup_y^i, e_{k,j}^{author} \right\} \quad ()$$

3.5. Connected CIA group generation

Disconnected subgraphs found from section 3.3 represent a snapshot of author relationships in a given year. CIA group behaviors in previous years need to be identified in order to perform classifications on states of CIA groups in the future. CIA groups are found yearly to generate snapshots of graph $G = \{C_i, C_{i+1}, \ldots, C_{i+m}\}$ where m is the total number of snapshots. CIA groups, once found, are then connected over the years with the CIA group similarity measure. Based on a community similarity score (Hopcroft et al., 2004), Connected CIA similarity measure for CIA groups C_y^i and C_{y+d}^j is defined as:

$$\text{sim}\left(C_y^i, C_{y+k}^j\right) = min\left(\frac{\left|\left(C_y^i - connected_{y+k}\right) \cap C_{y+k}^j\right|}{C_y^i}, \frac{\left|\left(C_y^i - connected_{y+k}\right) \cap C_{y+k}^j\right|}{C_{y+k}^j}\right) > \theta, k \geq 1, \theta \leq 0.5$$

where C_y^i stands for a community i at year y. Two communities at year y and year $y + k$ are defined as *similar* when the similarity value exceeds θ, a minimum similarity threshold which cannot exceed 0.5 in order to allow multiple CIA groups to be *similar* to a single CIA group in the different timeslot. Year distance k decides the yearly difference between connected CIA groups. $k = 1$ results in returning CIA group connections only between neighboring years. CIA group is not required to be active at every year and can resurface after a period of dormancy with fewer research activities among them, and $k \geq 1$ allows connecting CIA groups over non-neighboring timeslots. A set of common authors from already connected CIA groups were excluded in finding additional CIA groups to prevent recurring authors in a single connected CIA group to inflate the similarities between them, allowing multiple connections between CIA groups in the same year as well as in multiple years tracking authors in the original CIA group.

Connected CIA group $CC_{event}^{i,y}$ is a set of sequentially ordered CIA groups originated from a CIA group C_y^i satisfying a specific evolutionary event.

$$CC_{survival}^{i,y_1} = \left(C_{y_1}^i, C_{y_2}^j, \ldots, C_{y_m}^j\right) \; where \; y_i < y_{i+1}$$

$$CC_{merge}^{i,y} = \left(\left\{C_{y_j}^j, C_{y_l}^l, \ldots, C_{y_m}^m\right\}, C_y^i\right) \; where \; \left\{y_j, y_l, \ldots, y_m\right\} < y$$

$$CC_{split}^{i,y} = \left(C_y^i, \left\{C_{y_j}^j, C_{y_l}^l, \ldots, C_{y_m}^m\right\}\right) \; where \; y < \left\{y_j, y_l, \ldots, y_m\right\}$$

4. Evaluations

4.1. Extracting CIA candidates

Degree of topical autocorrelation in each of the bibliographic network was analyzed by comparing the keyword similarities $sim\left(G_y^{multi}\right)$ between author pairs in each layer of the multigraph G_y^{multi}; an autocorrelation is a characteristic of many graph data where the linked instances are more likely to have similar values in their attributes (Rossi et al., 2012). Analysis of keyword vector similarities between author pairs randomly selected from different layers and random author pairs in timeslot y for all 21 iterations were run. Two hundred keywords most frequently used by the CIA group candidates were used for computational efficiency. Authors from all bibliographic layers exhibited significantly higher (p-value = 0.0000) topical autocorrelation than random pairs of similarity 0.4847, with $sim\left(G_y^{authCite}\right) = 0.6312$, $sim\left(G_y^{coCite}\right) = 0.6305$, $sim\left(G_y^{bibCouple}\right) = 0.6310$, $sim\left(G_y^{coAuth}\right) = 0.6510$, and $sim\left(G_y^{coWord}\right) = 0.6256$. This validated the premise of correlation between author relationships and degree of topic sharing.

4.2. Mining CIA groups

The degree of shared interest in the whole author bases was evaluated by analyzing the topic similarities between authors in all CIA groups. Author similarities from individual bibliographic networks were also calculated for comparison. The co-word network was excluded from the comparison as it is a near-complete graph with poor clustering results. One thousand author pairs were randomly selected yearly for 20 iterations to generate 20,000 random author pairs across CIA groups and bibliographic clusters, excluding the year 2018 as most of the bibliographic data used in the iteration is from the previous years. Random selection was used as a baseline, randomly selecting author pairs among those related to the target research field. Ranked frequencies of FOS and keywords assigned to the author's publications published within the given timeslot were extracted as author topics with topic size N = (Ausloos, 2012; Blei & Lafferty, 2006; Dong et al., 2012) and then compared using a cosine similarity measure. A list of *domain words* used to at dataset extraction and target research field selection was selected to be filtered out, which are *human–computer interaction*, *human-computer interaction*, *human computer interaction*,

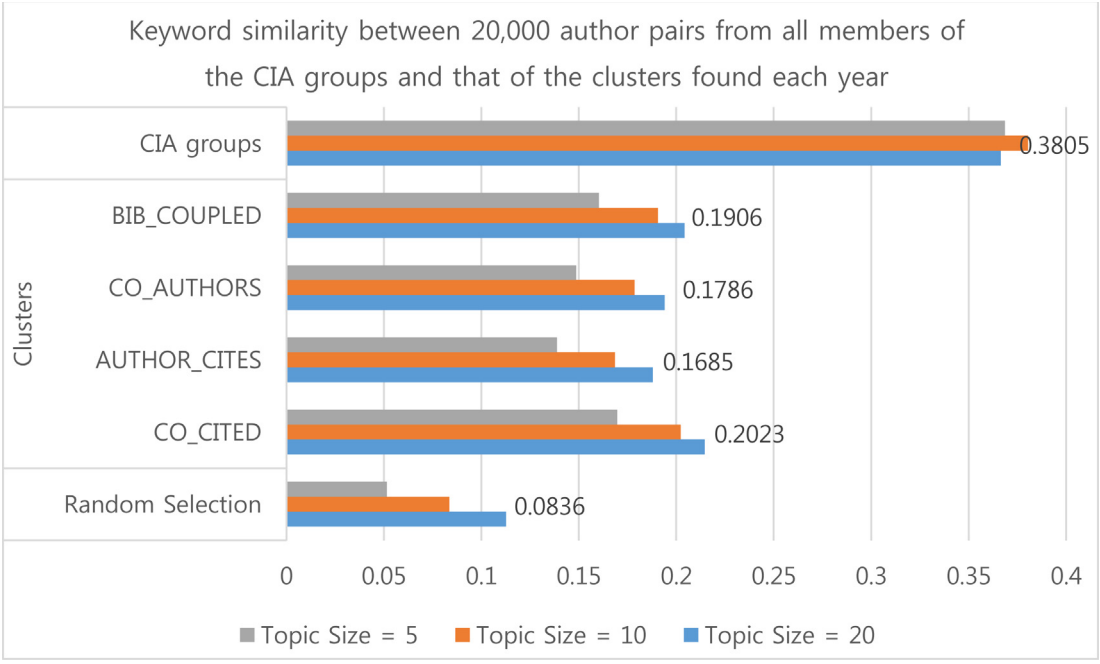

Fig. 6. Average topic similarities between author pairs in the same timeslot for different author group generation methods, CIA groups, bibliographic clusters, and random selection.

hci, and *virtual reality*. Domain word filtering aims to isolate the innate similarities between authors and to keep the mining CIA groups result from having its values boosted by them.

The outcome as shown in Fig. 6 revealed that author pairs in CIA groups have higher topical similarities compared to author pairs in all four types of bibliographic clusters as well as the baseline random selection. Author pairs from CIA groups had on average 2.09 times higher topical similarities against author pairs from clusters, with zero p-values for all three topic sizes.

The authors' topic similarities correlated to the number of topics used in the four bibliographic clusters and a random selection baseline, which was an expected behavior as a larger number of words per topic leads to a higher chance of two topics sharing them. CIA groups instead showed the highest author similarities when topic size is ten, which is not in sync with such observations. This suggested that authors forming the CIA groups share a smaller number of common vocabularies somewhere between 5–20 and the research interests of CIA group members are more concentrated compared to that of the authors from the bibliographic clusters. Higher topical similarities between authors in CIA groups led to higher topical similarities between CIA groups, as well, where the difference was significant with p = 0. The proposed method is shown to be capable of overcoming the limitations of a single evaluative measure in the traditional topic modeling, incorporating the inherent common interest shared by the target field instead of forcing the focused document collection to generate a textually disparate topic set.

LDA topic models were extracted to analyze the difference in the content of the topics generated by the proposed method. Document set $P^{y-b\sim y}$ for year y was used for each iteration. MAG includes abstracts in its dataset, hence abstracts were used as document texts. Texts were first tokenized into any alphanumeric words and stemmed using an implementation of Porter's algorithm (Porter, 1980) in NLTK[4] for its long history, and then any unigram, bigram, and trigram words with more than one instances in the abstracts of the given publication with word distances no longer than five were found using Gensim[5] allowing up to a couple of words between them. *Noun*, *verb*, *adj*, and *adv* were extracted using lemmatization with the spaCy[6] English model. Implementation of LDA modeling algorithm in Hoffman et al. 2010 (Hoffman, 2010) was then applied to generate ten topics with topic size N=(Ausloos, 2012; Blei & Lafferty, 2006; Dong et al., 2012) words with 100 iterations of the EM step each with 50 passes over the dataset or the relative improvement L of the document-topic distribution reaches below 0.001. The number of CIA groups generated per timeslot varies, and CIA topics of the ten largest CIA groups at each iteration were selected for further analysis.

3rd, 6th, and 9th CIA and LDA topics from $y = 2017$ were extracted as sample topics in Table 8, showing words unique within each topic type in **bold** and words common to all three sample topics in *italic*. CIA topics share four keywords

[4] https://www.nltk.org/
[5] https://pypi.org/project/gensim/
[6] https://github.com/explosion/spaCy

Table 8
CIA topic and LDA topic samples from year 2017.

	CIA_03	CIA_06	CIA_09	LDA_03	LDA_06	LDA_09
1st	*computer science*	augmented reality	*simulation*	display	use	가상현실
2nd	*multimedia*	*computer science*	*multimedia*	collabor	applic	다양한
3rd	*simulation*	*multimedia*	*computer science*	virtual_ hand	interact	디지털
4th	artificial intelligence	*simulation*	mixed reality	latenc	develop	사용자의
5th	virtual environment	interaction technique	artificial intelligence	stem	virtual_ realiti	이러한
6th	*computer graphics (images)*	user interface	engineering	box	environ	가존의
7th	visualization	mobile computing	pervasive computing	puzzl	technolog	대한
8th	collaboration	visualization	ubiquitous computing	threshold	system	증강현실
9th	avatars	*computer graphics (images)*	*computer graphics (images)*	head_ coupl	user	한다
10th	haptic interfaces	information interfaces and presentation	mobile computing	hmd	experi	garden

computer science, multimedia, simulation, and *computer graphics (images)* describing their common interest in the target field of research while displaying varying priorities toward them with the different word orders. Three words artificial intelligence, visualization, and mobile computing shared by two topics indicate the shared interests among different CIA groups. The words unique to the CIA topics show the unique interest of the CIA groups they represent; CIA_03 specializes in the **collaboration** in the **virtual environment** with **avatars** using **haptic interfaces**, for example. LDA topics are much more distinct, not showing any word sharing between three sample topics. While it is easier to discern the uniqueness of a given topic, it is harder to visualize the possible transitions and correlations between different LDA topics unlike the CIA topics where the commonalities and differences in the topics are visible on the words representing them.

Coherencies of CIA topics were analyzed to identify the degree of vocabulary sharing within the CIA group members. The topical coherence measure based on word co-occurrences within the given document set (Newman et al., 2010) was shown to have a higher correlation to expert-rated topic qualities than existing measures such as the topic size and the held-out likelihood, which is considered the highest form of golden data in measuring topical quality in the shortage of ground truth (Mimno et al., 2011). CIA topics and LDA topics share the same structure despite their origins hence the topical coherence measure for traditional topic models was used here. Topical coherence is defined as

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 0.001}{D\left(v_l^{(t)}\right)}$$

where $V^{(t)}$ represents a set of ordered words with size M $\left\{v_1^{(t)}, v_2^{(t)}, \ldots, v_M^{(t)}\right\}$, and $D\left(v_l^{(t)}\right)$ represents the document frequency of word type $v_l^{(t)}$, and D $\left(v_m^{(t)}, v_l^{(t)}\right)$ represents the co-document frequency of word types $v_m^{(t)}$ and $v_l^{(t)}$. Constant 0.001 in the formula is the smoothing factor added to avoid calculating log(0).

It is crucial to make a distinction between the topical coherence as a topic quality measure and the topical coherence as a measure of topic coherence; the definition of a *topic* is different in CIA topics. Unlike the traditional topic models with the goal of generating distinctive topics, CIA topics aim to identify topics of distinctive CIA groups and not necessarily distinctive themselves. The quality of CIA topics lies in the accurate identification of author groups with distinct interests, not in the accurate representation of distinct word co-occurrences. Generation of golden data through expert-annotation is impractical for the large-scale and evolving dataset, and lack of validated author-based topic quality measure prohibits a systematic analysis or performance comparisons against other state-of-the-art topic models. The comparison is done against LDA as it is the most robust and widely accepted traditional topic model, showing the CIA topics are coherent compared to the traditional baseline.

Coherences of topics from CIA groups, bibliographic clusters, and LDA topics are shown in Fig. 7 with an absolute logscale, where higher coherence is represented by values close to 0. Topical coherence for all four types of bibliographic clusters showed similar results while significantly less coherent compared to the CIA topics with p-values 1.14×10^{-144}, 1.52×10^{-187}, and 1.64×10^{-201} respectively for topic size N = 5,10,20. One possible reason for the low performance of the LDA topics is the use of abstract instead of the full-text in the topic modeling process, as a document set with longer texts generates better topics. Wang et al., 2017 (Wang et al., 2017) for example extracted topic models with -30 topical coherence from a scholarly dataset NIPS with topic size 10. The proposed method generated topics with a single-digit coherences, however, showing that the proposed method generated coherent topics even compared against topic models from a full-text dataset.

The necessity of a multi-layer approach is analyzed by comparing CIA topics representing CIA groups from different combinations of bibliographic networks; this includes all five networks, four networks each excluding co-author and co-word network to remove direct and indirect author relationships from the combination, and individual networks. Fig. 8

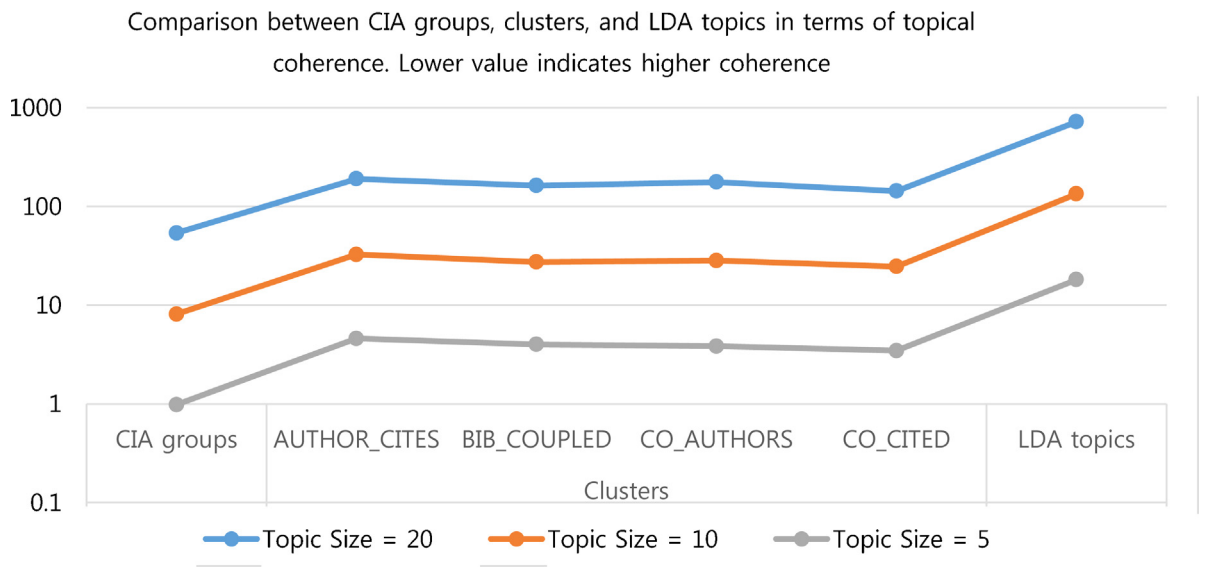

Fig. 7. Coherences of CIA groups, bibliographic clusters, and LDA topics with varying topic size, using absolute values for logscale representation.

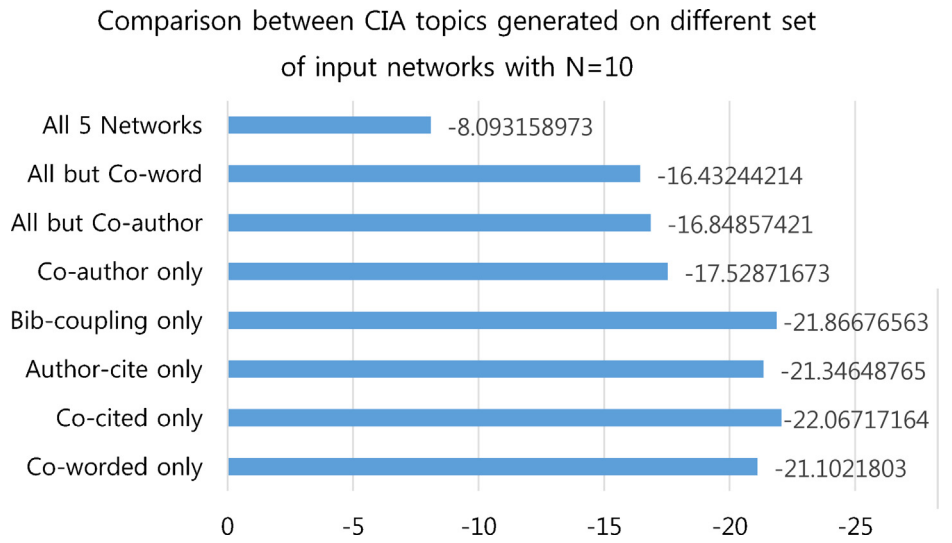

Fig. 8. CIA topic coherences from different combinations of bibliographic networks, with topic size 10.

shows CIA topics from all networks are the most coherent with topic size 10. CIA groups without co-word and co-author networks were more coherent than that of CIA topics from only a co-author network, which was the highest among the CIA topics from individual networks. It is worth noting that the poorly performing four individual networks were combined to produce topics of higher quality than the co-author network which distinctively is better than the rest in the quality of generated topics. The differences were statistically significant with $p = 3.65 \times 10^{-69}$. CIA groups with four networks showed minimal differences in topical coherence; different combinations of bibliographic networks were able to capture similarly meaningful CIA groups. This implies that while the result of individual networks may vary, combinations of multiple author relationships can lead to more coherent topic generation.

4.3. Showcasing CIA topic evolutions

The goal of the current paper is to propose an author-based topic modeling method conscious of the topical correlation in the academic domain. CIA-based topic evolutions are done to showcase the merge and split evolution analysis capabilities of the proposed method; the overall flow between multiple CIA groups reveals how the researchers, and their topics, have merged and split over time.

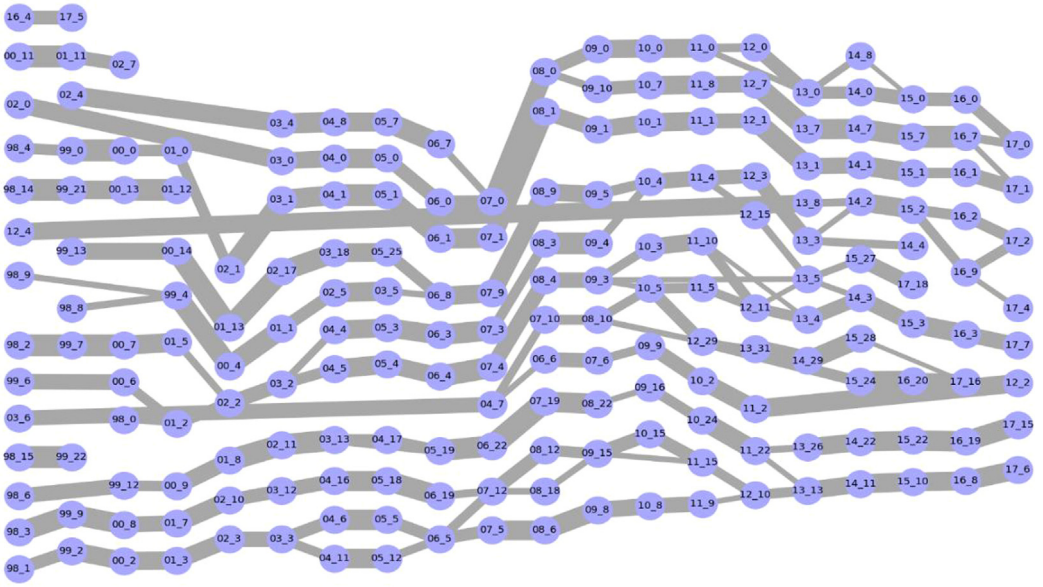

Fig. 9. Connected CIA groups with $\theta = 0.2$.

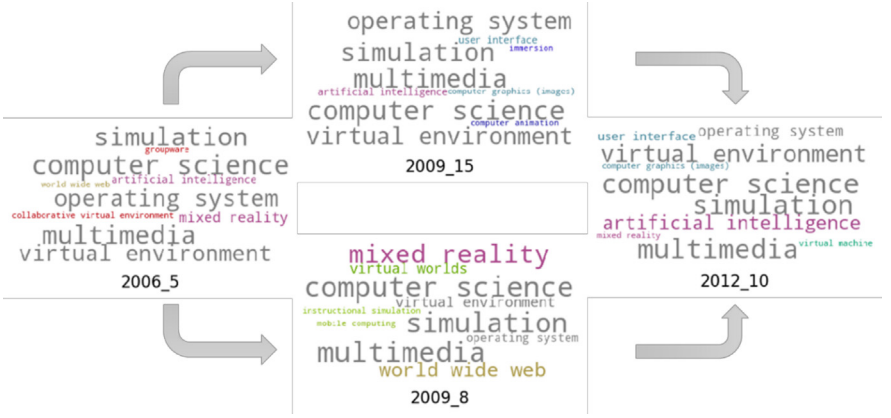

Fig. 10. An example of topic evolution with prolonged topic separation before re-emergence.

Fig. 9 is a visualization of the connected CIA groups including merge and split events with $\theta = 0.2$, where the nodes are CIA groups represented as the *yyyy_id* format with the first two letters removed for spacing purposes. The edges connect *similar* CIA groups; the edge width is linearly normalized so that the CIA groups with similarity one have a width 75 % of the node diameter. The disconnected chains indicate that the CIA group at the end of the chain shares no more than one-fifth of the authors with any other CIA groups in the future. This is rarely the case in the figure, and many groups are connected in long chains with frequent interconnections while maintaining the *main* branch in most cases. Multiple groups including *2012_03* and *2015_09* are simultaneously merging and splitting, indicating pivoting topics in the timeline. Duplicate connections within a single chain can also be observed. A path *2011_00–2013_00* indicates the delayed author convergence, where a minor portion of authors was dormant for a year before re-joining the initial CIA group. This is a rare occurring pattern in the figure; many researchers, while working on various research topics, do not halt research activities for a prolonged period before returning to the field. Connection with multiple year difference is also rare in the figure for the same reason.

The topical evolutions in the case of a topic separated for multiple years before re-merging into one are shown in Fig. 10, with the start and end of topic separation as well as one topic in each of the branches. All four topics share *computer science, multimedia, simulation, operating system, and virtual environment* as their major words depicting their overall shared interest in the domain field, marked as grey colors. *Mixed reality* was inherited by the main branch *2009_08* with more focus, which was shifted away when the two branches re-merged at *2012_10*. The opposite occurred with *artificial intelligence*, where the word survived through the sub-branch and received more focus when merged. Two branches, while similar, had distinct interests affecting the resulting topic differently. *2009_15* newly added *user interface* and *computer graphics (images)* to the merged topic indicating newly introduced interests of the CIA group. *2009_08* does not pass *world wide web* from the

Table 9
The occurrences of topic evolution events found for each global topic for 20 iterations, with topic size ten and similarity threshold 0.05, 0.25, and 0.45.

Survive				Merge				Split			
Thres	0.05	0.25	0.45	Thres	0.05	0.25	0.45	Thres	0.05	0.25	0.45
LDA_01				LDA_01				LDA_01			
LDA_02	7	2		LDA_02	1			LDA_02	3		
LDA_03	7	4	4	LDA_03	1			LDA_03	3	1	
LDA_04	5	1		LDA_04	1			LDA_04	1		
LDA_05				LDA_05				LDA_05			
LDA_06	1			LDA_06				LDA_06			
LDA_07	20	20	20	LDA_07	9	8	11	LDA_07	6	7	6
LDA_08	20	18	1	LDA_08	8	6		LDA_08	6	3	
LDA_09				LDA_09				LDA_09			
LDA_10	20	20	10	LDA_10	6	7	3	LDA_10	7	5	1

Table 10
Top five words of sample global LDA topics related to the merge and split events, never used topics shown in *italic* and highly used topics shown in **bold**.

	1st	2nd	3rd	4th	5th
LDA_01	*user_experi*	*cultur*	*citi*	*healthcar*	*male*
LDA_09	*campu*	*artwork*	*context_awar*	*tablet*	*elderli*
LDA_07	use	interact	system	user	design
LDA_10	simul	train	system	model	virtual

original topic to the merged topic reflecting the lost interest in the word over the years. *Groupware* and *collaborative virtual environment* in red and *virtual machine* in green each represent the unique interest of *2006_05* and *2012_10* not shown at the intermediate topics. The result suggests that the topic correlation can be observed by the proposed methods detecting merge and split between CIA groups over time, an overview of how topics correlate over time can be generated based on the outcome.

Such correlation analysis is hard to achieve with the traditional language-based topic evolution. The work of Chen et al., 2017 (Chen et al., 2017) was tried for comparison. LDA topics are found in corpus-level as time-spanning global topics and temporal topics as timeslot-specific local topics. Ten topics were found for both the topic sets in order to match the number of topics per year with CIA topics. Ten global topics are compared against ten local topics for each year y using cosine similarity measure with the same similarity threshold θ used for the CIA group similarity measures, resulting in the number of local topics $localTopicCount_y^g$ in year y similar to the specific global topic g. A global topic g with $0 < localTopicCount_y^g < localTopicCount_{y+1}^g$ and local topics connected to it were considered as *splitting* at year y, while g with $0 < localTopicCount_{y+1}^g < localTopicCount_y^g$ and similar local topics were considered as *merging*. $0 < localTopicCount_y^g = localTopicCount_{y+1}^g$ indicate unchanged number of similar local topics and was considered as *surviving*.

The majority of evolutionary events from LDA topics were observed on only a few global topics. With all nine different similarity thresholds [0.05, 0.10, . . ., 0.45], three global topics share on average 89 % of split events and 96 % of merge events while the remaining seven global topics were rarely if at all contributed to any of the events. Table 9 summarizes the number of the evolutionary events observed per global topics over 20 iterations; four out of ten global topics experience only experienced one survival events over 20 years. Table 10 exemplifies how such topics are different from highly used global topics. The generic terms used in LDA_07 and LDA_10 explain why such topics were connected to many different local topics as they are in essence the background materials for the domain of HCI. LDA_01 and LDA_09, on the other hand, are specific topics but their themes were unmatched to any of the local topics rendering them irrelevant to topic evolution analysis. These topics were found to be distinctively popular in the time-spanning dataset but were distant from the temporal topics, showing the temporal insensitivity of the method

5. Conclusion

The traditional topic modeling methods define the topics as distinct co-occurring word sets in the target document collection, failing to accurately reflect how the actual *topics* are formed and shared by the authors of those documents. A topic is not a keyword, but a concept structured within the mind of the authors, and an author-based topic modeling alternative is proposed where topics are defined as the set of common research interests exhibited by the members of the target research community. Topics are represented with the meta-data of the documents, eliminating the need for NLP and therefore a large text dataset in the proposed method as well. The proposed method was tested with HCI-related research articles from Microsoft Academic Graph with *virtual reality* as the topic of interest in 21 10-year span timeslots from 1988 to 2018. The experiment showed that the CIA-based topic modeling is a valid topic modeling alternative in generating a coherent and descriptive set of topics while retaining its focus on the authors. The CIA topics incorporated the inherent common interest shared by the members of the target research field interacting through several bibliographic links, which were combined to generate the most coherent topics. The CIA topic modeling is done in the premise that the group of creators

– authors – have a multitude of relationships around the shared common interests – topics, and it can be generalized not only to the other research domains but also to the other document types sharing the similar author relationships. For example, the same approach can be conducted with books, patents, and news articles to identify the topical evolution in the industry, academia, and public with different focus and time sensitivity (Segev et al., 2015). Application on the social network services would also be possible in some limited scenarios where the users share the common topics, such as in a disaster situation. Instead of identifying the overall interest over the region or the event and adapting to the changes (Segev, 2009), the proposed method would offer a more targeted trend tracking, allowing the aid workers to better adapt to the needs of different refugee groups.

The proposed method is a base for an enhanced topic evolution analysis, as well. Topic modeling based on the author groups can differentiate the changes in the group members from shifts in their interests which the traditional topic modeling methods cannot; the proposed method allows the distinction between such different perspectives on the topic evolution. The proposed method also allows divisive quantification of topic *flow* between multiple topics, allowing more complex topic evolution than the *content transition* where a single topic survived through with evolving word sets; CIA groups act as a transitional medium between CIA topics, enabling *correlation analysis* for dynamically changing research interests merging and splitting at the same time. The author-based topic similarity will be unaffected by the topic size as the CIA topics are focused on the concepts shared by the group members which are not dictated by the degree and size of the words shared.

Author contributions

Sukhwan Jung: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Wan Chul Yoon: Conceived and designed the analysis; Performed the analysis; Wrote the paper.

Declaration of interest

None.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.joi.2020.101040.

References

Allan, J., et al. (1998). *Topic detection and tracking pilot study final report. Proceedings of the broadcast news transcription and understanding workshop (Sponsored by DARPA). (Feb. 1998).*

Amancio, D. R. (2015). A complex network approach to stylometry. *PloS One*, *10*, e0136076 http://dx.doi.org/10.1371/journal.pone.0136076

Anjali, M.K. and Babu, A.P. Ambiguities in Natural Language Processing. International Journal of Innovative Research in Computer and Communication Engineering. 392–394.

Ausloos, M. (2012). *A scientometrics law about co-authors and their ranking. The co-author core. arXiv:1207.1614 [physics]. (Jul. 2012).*

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, *96*, 845–864. http://dx.doi.org/10.1007/s11192-012-0940-1

Battistella, C. (2014). The organization of Corporate Foresight: A multiple case study in the telecommunication industry. *Technological Forecasting and Social Change*, *87*, 60–79. http://dx.doi.org/10.1016/j.techfore.2013.10.022

Bengio, Y. et al. A Neural Probabilistic Language Model. 19.

Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models. Proceedings of the 23rd international conference on machine learning (New York, NY, USA, 2006).* pp. 113–120.

Blei, D. M., et al. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research : JMLR*, *3*, 993–1022.

Bongers, A., & Torres, J. L. (2014). Measuring technological trends: A comparison between U.S. And U.S.S.R./Russian jet fighter aircraft. *Technological Forecasting and Social Change*, *87*, 125–134. http://dx.doi.org/10.1016/j.techfore.2013.12.007

Börner, K., et al. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, *10*, 57–67. http://dx.doi.org/10.1002/cplx.20078

Callon, M., et al. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, *22*, 191–235. http://dx.doi.org/10.1177/053901883022002003

Castano, S., et al. (2017). Exploratory analysis of textual data streams. *Future Generation Computer Systems*, *68*, 391–406. http://dx.doi.org/10.1016/j.future.2016.07.005

Chen, B., et al. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, *11*, 1175–1189. http://dx.doi.org/10.1016/j.joi.2017.10.003

Chen, Z., et al. (2019). Tensor decomposition for multilayer networks clustering. *Proceedings of the AAAI Conference on Artificial Intelligence AAAI Conference on Artificial Intelligence*, *33*, 3371–3378. http://dx.doi.org/10.1609/aaai.v33i01.33013371

Cook, D. J., & Holder, L. B. (1993). Substructure discovery using minimum description length and background knowledge. *The Journal of Artificial Intelligence Research*, *1*(1993), 231–255. http://dx.doi.org/10.1613/jair.43

de Arruda, H. F., et al. (2016). Topic segmentation via community detection in complex networks. *Chaos an Interdisciplinary Journal of Nonlinear Science*, *26*, 063120 http://dx.doi.org/10.1063/1.4954215

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*, 187–203. http://dx.doi.org/10.1016/j.joi.2010.10.008
Dong, X., et al. (2014). Clustering on multi-layer graphs via subspace analysis on Grassmann Manifolds. *IEEE Transactions on Signal Processing*, *62*, 905–918. http://dx.doi.org/10.1109/TSP.2013.2295553
Dong, X., et al. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing*, *60*, 5820–5831. http://dx.doi.org/10.1109/TSP.2012.2212886
Ferrer, I., Cancho, R., et al. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*, 051915 http://dx.doi.org/10.1103/PhysRevE.69.051915
Fiscus, J. G., & Doddington, G. R. (2002). *Topic detection and tracking evaluation overview. Topic detection and tracking.* pp. 17–31. Boston, MA: Springer.
Gaul, W., & Vincent, D. (2017). Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification*, *11*, 159–178. http://dx.doi.org/10.1007/s11634-016-0241-2
Guo, Z., et al. (2014). A two-level topic model towards knowledge discovery from citation networks. *IEEE Transactions on Knowledge and Data Engineering*, *26*, 780–794. http://dx.doi.org/10.1109/TKDE.2013.56
He, Q., et al. (2009). *Detecting topic evolution in scientific literature: How can citations help? Proceedings of the 18th ACM conference on information and knowledge management (New York, NY, USA, 2009).* pp. 957–966.
Hoffman, M., et al. (2010). Online learning for latent dirichlet allocation. In J. D. Lafferty (Ed.), *Advances in Neural Information Processing Systems* (23) (pp. 856–864). Curran Associates, Inc.
Hong, L., et al. (2011). *A time-dependent topic model for multiple text streams. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD' 11 (San Diego, California, USA, 2011).* pp. 832.
Hopcroft, J., et al. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences*, *101*, 5249–5253. http://dx.doi.org/10.1073/pnas.0307750100
Huan, J., et al. (2003). *Efficient mining of frequent subgraphs in the presence of isomorphism. Proceedings of the third IEEE international conference on data mining (Washington, DC, USA, 2003), 549–.*
Hug, S. E., et al. (2017). Citation analysis with microsoft academic. *Scientometrics*, *111*, 371–378. http://dx.doi.org/10.1007/s11192-017-2247-8
Jiang, C., et al. (2010). *Frequent sub-graph mining on edge weighted graphs. Data warehousing and knowledge discovery*. pp. 77–88. Berlin, Heidelberg: Springer.
Jo, Y., et al. (2011). *The web of topics: Discovering the topology of topic evolution in a Corpus. Proceedings of the 20th international conference on world wide web (New York, NY, USA, 2011).* pp. 257–266.
Kay, L., et al. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology*, *65*, 2432–2443. http://dx.doi.org/10.1002/asi.23146
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*, 10–25. http://dx.doi.org/10.1002/asi.5090140103
Kim, J., & Lee, J.-G. (2015). Community detection in multi-layer graphs: A survey. *SIGMOD Record*, *44*, 37–48. http://dx.doi.org/10.1145/2854006.2854013
Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*, 373–397. http://dx.doi.org/10.1023/A:1024940629314
Levy, O. and Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. 9.
Li, M., & Chu, Y. (2017). Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *Journal of Information Science*, *43*, 725–741. http://dx.doi.org/10.1177/0165551516661914
Liu, Y.-B., et al. (2008). Clustering text data streams. *Journal of Computer Science and Technology*, *23*, 112–128. http://dx.doi.org/10.1007/s11390-008-9115-1
Louppe, G., et al. (2016). Ethnicity sensitive author disambiguation using semi-supervised learning. *Knowledge Engineering and Semantic Web*, 272–287.
Martin, S., et al. (2011). OpenOrd: An open-source toolbox for large graph layout. *Visualization and Data Analysis*, *2011*, 786806.
Mei, Q., & Zhai, C. (2005). *Discovering evolutionary theme patterns from text: An exploration of temporal text mining. Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining (New York, NY, USA, 2005).* pp. 198–207.
Menenberg, M., et al. (2016). *Topic modeling for management sciences: A network-based approach. 2016 IEEE International Conference on Big Data (Big Data) (Washington DC,USA, Dec. 2016).* pp. 3509–3518.
Mikolov, T. et al. Distributed Representations of Words and Phrases and their Compositionality. 9.
Mikolov, T., et al. (2013). *Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs]. (Sep. 2013).*
Mimno, D., et al. (2011). *Optimizing semantic coherence in topic models. Proceedings of the conference on empirical methods in natural language processing (Stroudsburg, PA, USA, 2011).* pp. 262–272.
Mucha, P. J., et al. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*, 876–878. http://dx.doi.org/10.1126/science.1184819
Newman, D., et al. (2010). *Automatic evaluation of topic coherence. Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics (Stroudsburg, PA, USA, 2010).* pp. 100–108.
Newman, N. C., et al. (2014). Comparing methods to extract technical content for technological intelligence. *Journal of Engineering and Technology Management*, *32*, 97–109. http://dx.doi.org/10.1016/j.jengtecman.2013.09.001
Papalexakis, E. E., et al. (2013). *Do more views of a graph help? Community detection and clustering in multi-graphs. Proceedings of the 16th International Conference on Information Fusion (Jul. 2013).* pp. 899–905.
Paul, S., & Chen, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, *10*, 3807–3870. http://dx.doi.org/10.1214/16-EJS1211
Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change*, *49*, 237–255. http://dx.doi.org/10.1016/0040-1625(95)00022-3
Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *The Journal of Artificial Intelligence Research*, *11*, 95–130. http://dx.doi.org/10.1613/jair.514
Rocklin, M., & Pinar, A. (2013). On clustering on graphs with multiple edge types. *Internet Mathematics*, *9*, 82–112. http://dx.doi.org/10.1080/15427951.2012.678191
Rosen-Zvi, M., et al. (2004). *The author-topic model for authors and documents. Proceedings of the 20th conference on uncertainty in artificial intelligence (Arlington, Virginia, united States, 2004).* pp. 487–494.
Rossi, R., et al. (2012). Transforming graph data for statistical relational learning. *The Journal of Artificial Intelligence Research (JAIR)*, *45* http://dx.doi.org/10.1613/jair.3659
Rudolph, M., & Blei, D. (2018). *Dynamic embeddings for language evolution. Proceedings of the 2018 world wide web conference on world wide web - WWW' 18 (Lyon, France, 2018).* pp. 1003–1011.
Schramm, T., & Steurer, D. (2017). *Fast and robust tensor decomposition with applications to dictionary learning. arXiv:1706.08672 [cs, stat]. (Jun. 2017).*
Segev, A. (2009). Adaptive ontology use for crisis knowledge representation. *International Journal of Information Systems for Crisis Response Management*, *1*, 16–30. http://dx.doi.org/10.4018/jiscrm.2009040102
Segev, A., et al. (2013). *Analysis of technology trends based on big data. 2013 IEEE international congress on big data (BigData congress) (Jun. 2013).* pp. 419–420.
Segev, A., et al. (2015). Analysis of technology trends based on diverse data sources. *IEEE Transactions on Services Computing*, *8*(2015), 903–915. http://dx.doi.org/10.1109/TSC.2014.2338855

Shiga, M., & Mamitsuka, H. (2012). A variational bayesian framework for clustering with multiple graphs. *IEEE Transactions on Knowledge and Data Engineering*, *24*, 577–590. http://dx.doi.org/10.1109/TKDE.2010.272
Sidiropoulos, N. D., et al. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, *65*(2017), 3551–3582. http://dx.doi.org/10.1109/TSP.2017.2690524
Silva, F. N., et al. (2016). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, *10*(2016), 487–502. http://dx.doi.org/10.1016/j.joi.2016.03.008
Steyvers, M., & Griffiths, T. (2007). *Probabilistic topic models. Handbook of latent semantic analysis. Lawrence Erlbaum Associates Publishers*. pp. 427–448.
Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(2012), 1820–1833. http://dx.doi.org/10.1002/asi.22695
Tang, W., et al. (2009). *Clustering with multiple graphs. 2009 ninth IEEE international conference on data mining (Dec. 2009)*. pp. 1016–1021.
Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *3*(2009), 11:1–11:29. DOI:https://doi.org/10.1145/1552303.1552304.
Wang, X., & McCallum, A. (2006). *Topics over time: A non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD' 06 (Philadelphia, PA, USA, 2006)*. pp. 424.
Wang, W., et al. (2017). Learning latent topics from the word Co-occurrence network. *Theoretical Computer Science*, (2017), 18–30.
White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(1998), 327–355. http://dx.doi.org/10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4
Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(2012), 1313–1326. http://dx.doi.org/10.1002/asi.22680
Zhang, Z.-Y., et al. (2013). Non-negative Tri-factor tensor decomposition with applications. *Knowledge and Information Systems*, *34*(2013), 243–265. http://dx.doi.org/10.1007/s10115-011-0460-y
Zhong, S. (2005). Efficient streaming text clustering. *Neural Networks*, *18*(2005), 790–798. http://dx.doi.org/10.1016/j.neunet.2005.06.008