Semantic Similarity Analysis between Future Topics and Their Neighbors in Topic Networks for Network-based Topic Evolution

Sukhwan Jung
Department of Computer Science
University of South Alabama
Mobile, Alabama
shjung@southalabama.edu

Aviv Segev
Department of Computer Science
University of South Alabama
Mobile, Alabama
segev@southalabama.edu

***Abstract*—Topic evolution automatically tracks a set of concepts within a given dataset over time, assisting researchers to overview various research domains. Network-based topic evolution is one of the recent approaches incorporating relational models instead of traditional text-based models for allowing the detection of topic correlation events. Topics are represented with co-occurrence relationships instead of word vectors, connected over time through their positions in a network instead of their semantic similarities. This paper shows that the topics and their network representations share meaningfully similar semantics. The existence of such contextual relationships allows topics to be labeled without having enough direct textual appearances in the document collection. Forty fields-of-study keywords with 64,215 to 5.8 million related articles were selected from the Microsoft Academic Graph dataset containing more than 200 million publications. The semantics of topics within forty topic networks were found using three sets of word embeddings trained from a collection of 10.5 million Medline abstracts from the year 2000 to 2016, and word embeddings of the topics are compared against their network-based representations, which are their neighborhoods in the previous timeslot. Cosine similarities between topics and their neighbors consistently resulted in moderate correlations from the year 2001 to 2015, showing higher values for topics that were already present in the topic networks compared to newly emerging topics. The result suggested that labeling topics based on the network structure are possible without semantic analysis, which is necessary for predicting topic evolutions such as the topic emergence when future documents are unavailable.**

***Index Terms*—topic evolution, network-based topic modeling, topic labeling, scientometrics**

I. Introduction

Scientific discoveries made around the globe expand the collective knowledge on the existing research topics and contribute to the creation of new topics. To ensure the validity and soundness of the outcomes, researchers contribute their work by gradually expanding on existing topics to base their findings. The current state of research fields provides insights into the outer boundary of knowledge, and their past evolution reflects the current demands and future impacts of new findings in the domain. Grasping the edge of the knowledge boundary is however becoming increasingly difficult for researchers in recent years as research fields are inundated with more publications each year. Increasing interdisciplinary collaborations between prominent research fields further aggravates the problem by enlarging the number of related fields as well. The number of relative publications often exceeds human capabilities with growing research communities in more interdisciplinary fields. The notion of topic modeling and topic evolution alleviates the workload by providing an overview of research publications, representing the shared theme among fellow researchers. Topic evolution and prediction are becoming increasingly important in research environments as it offers a semi-automated summarization of research fields, tracking shifts in interests and topical evolution over a given document collection without manually reading the overwhelming number of research articles [1]. Such knowledge would allow researchers to internalize current knowledge in the research domain and conduct a more targeted study on topics that are expected to be high in demand, which is important in both academic and industrial environments as such information is crucial in setting a future research goal. Researchers can make more valuable discoveries by focusing the research effort on more prospective goals, and industries can invest in potentially high-impact future technologies before competitors do. Knowledge about the projected *novel* topics would allow researchers to conduct preemptive research before the topics manifest. It will also inform the researchers which topics from other domains will be introduced to their fields for easier literature reviews.

Traditional topic evolution had limited success in automatic predictions [2]. Topic models specific to each timeslot can be compared over consecutive times using semantic similarity measures, connecting similar topics over time to analyze the semantic evolution within a given topic. Text-based topic models employed by the traditional models introduce an innate weakness to topic evolution, however; evolution within a topic and evolution involving multiple topics share the same semantic similarity measure as topic models do not distinguish between identity and semantics. Correlations between distinct topics over time are often not as well analyzed as such interactions are often interpreted as content transitions on

one topic instead. Recent research on topic evolution tried to overcome this limitation by proposing a network-based approach to topic evolution [3], where the topics are represented as the neighborhood set within the topic network built from the pre-defined bibliographical topic dataset. The topic network structures are independent of the domain contents they represent as the only information used is the relative relationships between different topics. Nodes represent the identity of topics separated from their semantics, bypassing the aforementioned limitation of text-based topic models in topic evolution to capture more evolutionary events such as *merge* or *split* between different topics over time. This approach showed detecting emerging topics using data available in previous timeslots, enabling prospective topic prediction over a wide range of research domains [4]. While the topics were predicted using their neighborhoods within the networks, the semantics of the individual topics were not analyzed in the research. Semantic similarities between topics are ignored, and only membership similarities are used to identify topical correlations including merging and splitting of topics. This paper is an attempt to verify the validity of the network-based topic evolution by evaluating the semantic similarities between the detected topics and the actual result. Similarities between topics' semantics and their network-based representations would indicate that the new approach can provide semantically consistent topic identification, which would allow topic naming at the application level as well.

The goal of the proposed method is to label new topics predicted to emerge from a temporal topic network, utilizing their correlation to the existing topics. Subgraphs in the given topic network defined as to-be-neighbors of new topics in the future are used to extract a common vector representing their future neighbors. The topic networks are first extracted from an open bibliographical dataset, with each network representing publications in a specific research journal with a focused set of research interests. The topic network is divided yearly to generate an evolving network, where each topic in a timeslot y is either *new*, appearing for the first time in y for the given topic network, or *old*. The *new* topics are then compared against the topic subgraphs found at the previous timeslot $y-1$ as the precursors to the newly emerging topics in y. The comparison is made by comparing semantic similarities between the two using word embeddings trained from the medical domain, and thirty topic networks were generated from publications related to medicine-related fields of studies from the Microsoft Academic Graph[1] dataset. Ten additional non-medical-related networks were also tested. The accuracy of predicted topic labels is analyzed with vector similarity comparison to objectively measure the semantic integrity of the predictions.

The word embedding of topics and their neighbors showed moderate to high correlations over all tested domains, indicating the topic labels can be derived from their neighborhood information. Using a default approach with a Fasttext embedding algorithm and skip-gram models, an average semantic similarity of over 0.48 was observed for all 40 tested networks, reaching over 0.52 on ten networks based on non-medical-related domains. Generic non-medical domains with remote connections to many medical fields showed higher similarities compared to the majority of highly medical fields because word embedding were learned from general medical publications and not being specialized to each field. The proposed method is shown to be generalizable to research domains not directly under the topical hierarchy of the training materials, and it is expected to perform similarly in other unexperimented research fields as long as the word embedding learned from the domain is provided.

Section 2 reviews the related works on network-based topic evolution as well as background research on word embeddings. Sections 3 and 4 detail the proposed method and experimentation including the 40 FoS-specific topic networks used. The semantics of extracted topics were found using word embeddings pre-trained on the Medline abstract text dataset. The experiment results show the correlation to the semantics of the actual topics in Section 5.

II. Related Work

Research topics represent the set of shared themes within a given collection of publications. They can appear in various forms, including the philosophical category of the research, theoretical development of research models, applications of the technology, and specific algorithms. Topic evolution is the field of research focusing on the temporal evolution of topics, identifying and predicting *content transition* within a single topic as well as *topical correlation* between different topics [1]. Six general event types are defined as topic evolution event, including *survive*, *dissolve*, *grow*, *shrink*, *split*, and *merge* [5].

Identifying emerging topics in academic papers is a crucial part of research activity. Traditional topic models represent topics as the latent semantic structures, discovered within a given document collection in the form of word-popularity sets [6]. Probabilistic models such as Latent Dirichlet Allocation and its variants are frequently used to represent the semantic structures over a series of timeslot-specific topic models. Links between the words and topics are iteratively assigned and updated with word co-occurrence frequencies between documents [7]. Topic evolution provides insight into how the topics evolve over time using text-based topic models, identifying topics in the given document collection, and tracking topical changes over sequentially ordered timeslots. Document collection is first divided either uniformly or irregularly [1] into sequentially-ordered sub-collections on which topic models independent of the neighboring sub-collections are generated. The topics are first extracted from each of the temporally ordered subcollections to generate a sequence of timeslot-specific topic models [8]. Temporal topic models are then connected over time with similarity measures, and

[1]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

changes in the topics are sequentially analyzed to identify the evolution of topics based on semantic similarities.

The limitation of traditional topic evolution comes from the topic modeling methods, where changes in topics can only be measured by the differences between the content of two topics [9]; connections and correlations between different topics are not incorporated into the traditional topic modeling methods [3], [10]. Other time-sensitive topic identification methods such as dynamic topic models [11] and evolutionary theme pattern mining [2] used semantic distances to approximate the same topics over time, identifying evolutionary transition events when changes within semantically similar topics are found. A network-based approach was proposed [3], [12] by defining topics as their neighborhood membership nodes in the topic networks, utilizing bibliographical meta-data recorded for research articles. Topic evolution in conjunction with bibliographical data has been successful in the past for better topic transition event identifications, augmenting topic models with citation contexts [13] and expanding document collection with referenced articles [14].

Several studies dedicate themselves to detecting emerging topics by utilizing alternative topic definitions where a topic is represented by a single word instead of a vector. First story detection is a sub-tasks of Topic Detection and Tracking, aiming to capture and catalog newly introduced topics in text stream data such as multilingual news articles [15]. Words with distinct neighboring word distributions are identified in real-time to capture novel topics as they are introduced to the text stream [16], [17]. Burst term detection instead monitors incoming data stream to capture rapid frequency growth within a given word in an attempt to capture a new topic in its infancy [18]. A combination of burst term detection and keyword co-word analysis was done to identify new topics in the format of a multidimensional research front exploration [19]. Other definitions of emerging topics, such as word frequency-based research front detection [20], new topic identification using query pattern mining [21], and integration of publication venues based on keyword similarities [22] to detect and predict topics newly introduced to the dataset in question.

Utilization of network-based topic representation allows extrapolations to future new topic predictions [23], [24], as the definition allowed the topics in a certain timeslot to be classified based only on the structural data available in the previous timeslots [25]. The network-based topic emergence prediction is a variation of the node prediction problem, where node newly introduced to a network is predicted in terms of their initial connections to the existing network using the network's structural properties [26]. Random node assignment with preferential link attachment is proposed to detect new publications within citation networks [27]. Topics, instead of being semantic constructs, are defined as their neighborhoods within a topic network to convert the topic evolution problem into a node prediction problem. This could provide a novel functionality to the field of topic evolution, predicting correlations between multiple topics over time. This paper provides labeling for the predicted topics by utilizing the word vectors of their neighbors.

The Word2Vec is a relatively recent natural language processing technique using a set of neural networks to learn word associations to produce a vectorized word set [28]. There are two basic approaches to the Word2Vec, the continuous bag of words (CBOW) and the skip gram. The CBOW architecture assumes the given document as a set of unordered bag of words, which is built by connecting a certain number of continuous words. The semantics of the words are predicted from the contexts of other members of the bag. The skip gram architecture on the other hand predicts the context of a given word by utilizing the word's distributed representation seen by surrounding words [29]. The Fasttext [30] is an advanced version of the Word2Vec algorithm, where the word vectors are augmented by additional information acquired from variable character n-grams. More recent approaches to language models include incorporation of global vectors such as the GloVe model [31]. Combining both the global and local language models, the GloVe performs as a log-bilinear regression model utilizing only nonzero portions of the global word co-occurrence matrix rather than having the entirety of sparse matrix. The BERT model [32] recognises the polysemous nature of words by allowing multiple instances of the same words in different sentences to be represented by distinctive embedding. This paper utilizes basic Word2Vec algorithms to identify embedded topics within academic publications; the authors aim to show that the semantic similarity between topics and their neighbors are significant enough to be seen with basic algorithms. The language models are then utilized to extract possible context for the new topics emerging from the topic networks.

The Microsoft Academic Graph (MAG) bibliographic dataset generates a six-level hierarchical ontology to assign fields of study (FoS) to the recorded publications [33]. They are generated monthly using knowledge base type prediction with Wikipedia articles, employing graph link analysis and convolutional neural network methods. The hierarchical concepts are then tagged to the papers using a large-scale, multi-level text classification method on pre-trained word embedding vectors [34]. FoS are used as topics in this paper to build topic networks. The Microsoft Academic Graph (MAG) bibliographic dataset [33] was used to extract topic networks from various domains. The MAG dataset was deemed competitive with major bibliographic search engines such as Google Scholar or Scopus even with relatively recent creation [35], and the possibility of direct bulk download allows easier data manipulation without having to worry about the API call frequency limits often imposed by other online-based datasets. The MAG also has a built-in ontology called fields of study (FoS) representing each paper with different hierarchical concepts [34]. A six-level hierarchy of concept spans from high-level domains such as *chemistry* and *physics* to the specific keywords such as *topic based vector space model* and *p-glycoprotein interactions*. They are generated monthly using knowledge base type prediction with Wikipedia articles, employing graph link analysis and convolutional neu-

ral network methods. The hierarchical concepts are then tagged to the papers using a large-scale, multi-level text classification method on pre-trained word embedding vectors. The tagging is done weekly to keep up-to-date concept assignments. Identifying dataset-wide topics in a large-scale dataset is by itself a huge task; therefore the tagged FoS are defined as the topics for the document in this paper. While the author-assigned keywords in research publications also represent their topics, the MAG database does not have keywords as one of its relational database tables and therefore is not used in the experiments. Topics are represented by the FoS in this paper to build topic networks on which emerging topics are identified.

III. Comparison between word embeddings of future topics and their neighbors

A. Extracting Topic Neighbors in a Previous Timeslot

The proposed method utilizes a topic network, where emerging topics in a bibliographic dataset equate to new nodes in the topic network. Only the FoS identifiers were used to distinguish the topic nodes in the topic network; textual metadata such as the FoS labels and paper titles are not considered for analysis. The topic network $T_y = (V, R_y)$ represents co-occurrence frequencies R_y between topics V within the knowledge domain at given year y. Topic set V consists of the topic node u and the year the topic is first used in the dataset f_y, and R_y is the weighted edge set between nodes u and v, with w_y as co-occurrence frequencies in y.

$$T_y = (V, R_y),\ V = (u, f_y), \text{ and } R_y = (u, v, w_y) \quad (1)$$

Topics in year y are classified as *new* or *old* based on the structural features of their neighbors. Neighborhoods $neighbors(v, y)$ of each topic v in year y are first extracted to build a set of neighborhoods N_y from T_y. Each neighborhood is then categorized into two groups by the age of v, calculated by $f_y(v) - y$. This categorizes whether the topic v first appeared in the given year y, in which case $f_y(v) = y$. The state of v, $C(v)$ is calculated as the ceiling of topic age normalized by the oldest topic, where the new topics are denoted by $C(v) = 0$. Any preexisting topics have non-zero ages, and the normalized ceiling function result in $C(v) = 1$.

$$N_y = \{neighbors(v, y) | v \in V_y\}, \text{ and}$$

$$C(v) = \lceil (y - f_y(v))/(y - max(f_y(V))) \rceil \quad (2)$$

More prominent topics are likely to co-occur with more topics, and therefore the top 500 topics with the largest number of nodes in N_y are selected for each label $C(v) = 0$ and 1, resulting in a total topic count of 1,000 for each topic network. If the number of instances for one label is below 500, then the number of v for the other label is reduced further to have the same number of instances for both labels.

B. Retrieving Vectors for Topics and Their Neighbors

Word embeddings trained on the collection of research articles are used to extract the shared semantics within the given set of neighboring topics. The vocabulary size renders the one-to-one ratio between vocabulary and vector dimension impractical, and word embeddings with reduced dimensions are used for faster training and easier computation.

With $d = 200$ word embedding dimensions, a word embedding for the given vocabulary is a one-dimensional vector of length d, with each component representing the real-valued encoding for each dimension. A word embedding for the given topic v is calculated as the mean value for all words within the topic, and a word embedding for its neighbors is calculated as the mean value for all neighboring topics.

$$E_v = \frac{1}{|v|} \cdot \sum_{w}^{v} \langle w_1, w_2, \ldots, w_d \rangle,$$

$$NE_v = \overline{E_{x \in neighbors(v)}} \quad (3)$$

Other forms of vector averaging method were tested to analyze the effect of background semantics prevalent within the training document collection, and remove the effect of utilizing only a subset of the semantics when looking at one topic and its neighbors. The background semantic of the given topic set B^{freq} is first calculated as the average word embedding from all topics W. A binary background semantic B^{binary} is then calculated by selecting only distinct words from the set, which is denoted as $W_{\neq}$. The calculated background semantics are then subtracted from the word embeddings of topics and their neighbors to move the zero coordinate to the center of the given topic network.

$$W = \{w | w \in v,\ \forall v \in V\},\ B^{freq} = \frac{1}{|W|} \cdot \sum_{w \in W} \langle w_1, w_2, \ldots, w_d \rangle$$

$$W' = W_{\neq},\ B^{binary} = \frac{1}{|W'|} \cdot \sum_{w \in W'} \langle w_1, w_2, \ldots, w_d \rangle$$

$$E_v^{freq} = E_v - B^{freq},\ NE_v^{freq} = NE_v - B^{freq}$$

$$E_v^{binary} = E_v - B^{binary},\ NE_v^{binary} = NE_v - B^{binary} \quad (4)$$

The word embeddings for the topic v and its neighbors $neighbors(v)$ are compared by calculating the cosine similarity, or dot product of their unit vectors ($\widehat{E_v}$ and $\widehat{NE_v}$, respectively). This represents the similarity between the semantics of a topic and its neighbors in the previous timeslot, and a high outcome would suggest that the future topics' semantics can be predicted from current topic networks.

$$Sim(v) = \widehat{E_v} \cdot \widehat{NE_v} \quad (5)$$

TABLE I
THIRTY MEDICINE-RELATED FOS IN THE FEBRUARY 2020 MAG DATASET USED IN THE EXPERIMENT.

audiology	internal medicine	pathology
cardiology	medical emergency	pediatrics
dermatology	medical physics	physical medicine and rehabilitation
emergency medicine	medicinal chemistry	physical therapy
endocrinology	medicine	psychiatry
family medicine	nuclear medicine	radiology
gastroenterology	obstetrics	surgery
general surgery	oncology	traditional medicine
gynecology	ophthalmology	urology
intensive care medicine	optometry	veterinary medicine

TABLE II
TEN NON-MEDICINE-RELATED FOS IN THE FEBRUARY 2020 MAG DATASET USED IN THE EXPERIMENT.

atmospheric sciences	computer science	earth science
fishery	geography	history
law	marketing	mathematics
metallurgy		

IV. EXPERIMENTS

A. Extracting Topics and Their Common Neighbors

The MAG dataset snapshot in February 2020 is downloaded for preprocessing through the Microsoft Azure Databricks, containing 197,642,464 publications, 709,934 FoS, 48,829 journals, 1.5 billion citation links, and 1.3 billion paper-FoS links. Analyzing the whole graph would be too complex to compute, and therefore data subsets are extracted as the bibliographic records related to selected FoS, each representing subsets of topics focused on different research fields. FoS and paper-FoS links data files are converted into queryable database tables; publication metadata including publication venues and citation information is not required in this experiment and hence was not converted.

Forty knowledge domains are extracted to test the proposed method, each represented by a single domain FoS. All FoS used with the domain FoS are extracted with their co-occurrence frequencies which dictate link frequencies in the topic network. A *FieldOfStudyId* for each domain FoS is first retrieved from the *FieldOfStudy* table, then all data rows in the *PaperFieldsOfStudy* table containing the matching *FieldsOfStudyId* are retrieved. Rows matching the filtered papers in *PaperFieldsOfStudy* and *FieldsOfStudy* tables are retrieved for FoS used in conjunction with the domain FoS and how they are assigned to different publications. With a series of SQL queries, *FirstUsed-Year* column is added to the *FieldsOfStudy* table to represent the first year f_y the each FoS is used in conjunction with the target FoS, and *FoSneighborCount* {*Node1, Node2, Year, Frequency*} table is created to represent undirected links within each dataset with node pair u, v, year y, and frequency w, where FoS are the nodes and the links represent the two FoS assigned co-occurring in the same publications. Frequency shows the co-occurrences between two FoS divided for each year to distinguish different FoS links and weights at different years.

Word embedding trained on medical publications were used for comparison as stated in the following subsection, therefore 30 Medicine-related domains are extracted for experimentation; spanning a wide range of medical domains including medical practice branches such as *endocrinology* or *family medicine*, specific treatments such as *surgery*, and academic fields related to medicine such as *medical physics* and *medicinal chemistry*. 10 non-medical domains are also extracted for comparison, selected among levels 0 and 1 of the MAG's FoS hierarchy level with minimal connection to the medicine-related domain topics; for example, *geography* was selected due to its connection to *family medicine* and *general surgery* through research on the relationship between patient addresses and received medical treatments. Domain FoS used to extract topic networks is shown in Table I and Table II.

To regularize the size of the experimented dataset, the topic network in any given year f_y is limited to a maximum of 1,000 nodes with the largest occurrences to the total maximum of 15,000 nodes per domain topic. To ensure label balance, this filtration is done by limiting the size of each of the *new* and *existing* node types to half of the total size limit; 500 per type. Most domain topics resulted in large enough networks that the total number of nodes reach more than 99% of the limitation except *atmospheric sciences* and *earth science*, where only 7,317 and 6,345 *new* topics were introduced over the course of 15 years. Age of the *existing* topics showed an average of 46.26 with a standard deviation of 7.79, with the exclusion of outlier *history* with an average age of 95.34. This is likely due to the fact that they are not as widely used as other topics that were already present within the domain. Topics recently introduced to the domain were not extracted with the minimum average age of *oncology* = 28.28. Structural differences between medical-related and non-related topic networks are kept to a minimum; neighborhoods of filtered topics showed statistically insignificant mean differences in the number of nodes ($p = 2.9842E - 24$), the number of edges ($p = 4.5082E - 9$), and the average degrees ($1.5397E - 264$) when two topic types are compared. A major distinction could be made between *existing* and *new* topics as in Table III, showing that *new* topics have not only small neighborhoods but also have much lower interactions within their neighborhoods. Fewer interacting entities coupled with a lower rate of interaction indicate that *new* topics will be harder to identify using information from their neighbors.

After the dataset preprocessing is done, the topic network T_y in Eq. 1 for each FoS is generated for $y = [2001, \ldots, 2015]$. The year y is ranged to retrieve the detection of newly used topics in the 21st century, disregarding the most recent five years to account for newly introduced words with less exposure in the Word2Vec training document collection. For each FoS, SQL queries are run on the *FoSneighborCount* table to extract topic co-occurrence with

TABLE III
AVERAGE NUMBER OF NODES AND EDGES WITHIN NEIGHBORS FOR 30 MEDICINE-RELATED AND 10 NON-MEDICINE-RELATED DOMAIN TOPICS.

		Medicine		Non-Medicine	
		existing	*new*	*existing*	*new*
#nodes	Mean	31.1125	10.3068	32.7383	9.4820
	Std	12.6196	1.7939	14.5300	2.1561
#edges	Mean	303.6757	37.4429	322.3877	28.4631
	Std	265.6215	14.3558	300.3694	13.5586
avg_degree	Mean	15.2853	6.5197	14.9419	5.1819
	Std	5.5851	1.3691	6.2119	1.2811

$FoSneighborCount.Year = y$ where the *Year* column in the *FoSneighborCount* table represents the year the topics co-occurred. The resulting edge data R_y are used to build neighborhood sets as in Eq. 2.

B. Comparing Embeddings of Topics and Their Neighbors

Word embeddings pre-trained on research publications are used in the experiments[2]. Three sets of word embeddings (*w2v_skip*, *w2v_cbow*, and *fast_skip*) are learned from the 10.5 million Medline abstracts from January 2000 to December 2016, producing 822,000 unique 200-dimensional vectors for 1.67 billion words in a corpus [36]. The word embedding sets are categorized by their prefixes and suffixes; the prefix (*w2v* and *fast*) defines the word embedding algorithms and the suffix (*skip* and *cbow*) defines the model used for training. *W2v* stands for Word2Vec, which is currently one of the main algorithms in natural language processing, embedding words into a dense N dimensional space using an unsupervised machine learning approach [37]. A continuously moving context window observes word co-occurrences within similar contexts, assigning them close to each other. Fasttext [30] (*fast*) is an advanced version of the Word2Vec algorithm, where the word vectors are augmented by additional information acquired from variable character n-grams. This allows the partial semantic assignments to prefixes, suffixes, stems, and other syntactic structures in words even when the word co-occurrences do not share similar contexts. *Skip* means the skip-gram model is used for learning while *cbow* stands for the use of the continuous bag-of-words model. The goal of this paper is to verify the semantic predictability of new topics based on the topical keywords used within the target domain; other bibliographic information such as co-authorship and citations [38] were not utilized to remove the possible effects of dense research communities.

The neighborhood set is used to identify word embedding for any given topic E_v and its neighborhood NE_v for all three sets of word embeddings, without applying origin transition in Eq. 4 as a default. Variations of origin transitions are also tested as well, adjusting the zero coordinate to the center of topic-specific words without duplicates (*adj*) and with duplicates (*adj_freq*). Vector subtractions are also tried in unit vectors, which are represented as *adj'* and *adj_freq'*. Cosine

[2]https://zenodo.org/record/802965

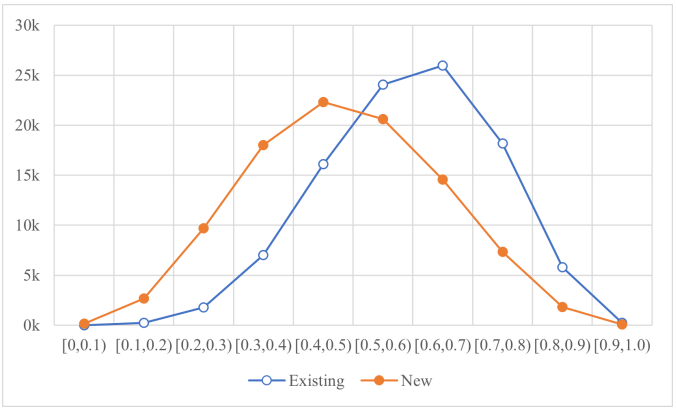

Fig. 1. Number of *existing* and *new* topics over all 40 topic networks, with the decile range as X axis and number of topics as Y axis.

similarities between topics and neighbors are measured for all combinations of topic networks and word embedding sets in order to analyze the effects of different learning approaches on the quality of topic labeling based on the network structures.

V. RESULTS

The word embeddings for the topic v and its neighbors $neighbors(v)$ are compared by calculating the cosine similarity between E_v and NE_v in Eq. 3, which is used as a *default* approach without applying origin transitions. The fasttext skip-gram model is used for the analysis. Over the span of forty domain topics, a total of 590,208 topics were measured for their similarity to their neighbors in the previous timeslot. *Existing* topics, in general, have extensive histories within the given topic network, having more than 40 years to assimilate with the neighboring topics. Such effect can be observed in Fig 1 where the *existing* topics draw positively skewed cosine similarity measures with skewness of 0.56. *New* topics showed lower similarities, which is to be expected as their neighbors previously had no common topic to bind them together; they show relatively normal distribution with a skewness of 0.28.

Analyzing topics with extreme similarity values showed that the descriptiveness of the topic word is related to the outcome. Topics such as *gpm* or *situs* shown in Table IV showcase the possibility of mixed semantics. *Gpm* in marketing can represent marketing terms such as Gross Processing Margin, Gross Profit Margin, or Graduated Payment Mortgage as well as non-marketing terms such as Gallons Per Mile or Graphical Path Method. *Situs* in the law field represents a legal term determining which law is applied to the properties but has a distinct usage in the biology field referring to the disposition of organs with left-right asymmetry. Homonyms have multiple semantic profiles, which were not distinguished in the experiment. There were other low-similarity topics without homonymy such as *dbz* (decibel relative to the reflectivity factor z), *ryr1* & *klrd1* (codenames for specific genes), and *gm6001* (a broad-spectrum matrix metalloproteinase inhibitor). While sharing the same initialized format, these topic words have specific semantic profiles and therefore are coupled

TABLE IV
TEN *Existing* AND *New* TOPICS WITH FIVE LOWEST AND HIGHEST COSINE SIMILARITIES TO THEIR NEIGHBORS.

	Fos	Year	Topic	sim
Existing	marketing	2012	gqm	0.0148
	geography	2015	dbz	0.0380
	atmospheric sciences	2012	dbz	0.0595
	law	2014	situs	0.0922
	traditional medicine	2013	trigeminal neuralgia	0.0932
	...	...	...	...
	law	2013	statute of the international court of justice	0.9508
	history	2012	anthropological linguistics	0.9512
	history	2014	psychology of religion	0.9539
	law	2012	political journalism	0.9548
	law	2014	international covenant on economic social and cultural rights	0.9562
New	obstetrics	2011	ryr1	0.0156
	physical medicine and rehabilitation	2013	gm6001	0.0177
	psychiatry	2011	klrd1	0.0282
	ophthalmology	2011	interleukin 28b	0.0284
	obstetrics	2014	heptafluorobutyric acid	0.0307
	...	...	...	...
	audiology	2011	rem sleep behavior disorder screening questionnaire	0.9522
	computer science	2011	web services cloud computing	0.9631
	medicine	2015	posterior cruciate ligament brace	0.9656
	physical therapy	2015	posterior cruciate ligament brace	0.9656
	surgery	2015	posterior cruciate ligament brace	0.9656

with more descriptive topics such as *trigeminal neuralgia* and *heptafluorobutyric acid*; their low performances indicate that they have a unique relationship within given domains different from their general use. For example, *klrd1* is a human gene related to the natural killer cells and their receptors and usually studied under its normal functions such as influenza susceptibility prediction [39], while the gene in the psychiatry field is studied for schizophrenia-relevant pathways [40] or gene expression in depressive disorders [41]. Topics with high cosine similarities indicate that each topic has a singular meaning across different domains. For *existing* topics, this equates to topics that are very well defined and used under a single semantic across different domains. *Statute of the international court of justice* and *international covenant on economic social and cultural rights* are such topics, as it is a well-defined topic with a specific global meaning with very little possibility of other uses. *New* topics with extremely high correlations to their neighbors are the result of unifying a single set of interdisciplinary topics under the same context, which is shown by topics such as *posterior cruciate ligament brace* which made an appearance in three different medical domains in the same year.

The experiments result showed that there are moderate to high correlations between the word embedding of topics and their neighbors in the previous timeslots over the range of tested topics. The correlations were consistently observed over all forty topic networks from the year 2001 to 2015, showing strength in capturing the semantics of topics that were already present in the network. Identifying semantics of emerging topics in their embryonic stage was done with limited success, validating the idea of network-based topic evolution and labeling. Table V shows the cosine similarities between them, dividing *Existing* ($C(v) = 1$) and *New* ($C(v) = 0$) topics. Five different cosine similarity measures are compared in the table, with *default* showing the result of Eq. 5. adj^{binary} and adj^{freq} represent the cosine similarities when the background semantics were subtracted from the embeddings to identify more topic-specific similarities. E_v and NE_v are replaced with $E_v^{binary}, NE_v^{binary}$ and E_v^{freq}, NE_v^{freq} respectively as shown in Eq. 4. Finally, adj^{binary}, and adj^{freq}, show modified versions of both measures when the subtractions were done in unit vectors instead. The result showed that *Existing* topics consistently showed high correlations to their predecessors compared to *new* topics, with the statistical differences signified with a p-value less than $1e^{-60}$ for all combinations of topic networks and embedding methods. This is to be expected as the existing topics have a chance to appear in the collected documents along with their previous neighbors to share their semantics. *fast_skip* showed the highest results with average value of 0.4204 compared to 0.3657 for *w2v_cbow* and 0.4140 for *w2v_skip*, consistently resulting in higher correlations with

TABLE V
TOPIC LABELING RESULT SUMMARY FOR ALL TOPIC NETWORKS BUILT FROM FORTY FOS MEASURED BY THE COSINE SIMILARITIES BETWEEN TOPICS AND THEIR NEIGHBORHOODS IN THE PREVIOUS TIMESLOTS.

Existing			
	fast_skip	*w2v_cbow*	*w2v_skip*
default	**0.5944**	0.4025	**0.5509**
adj^{binary}	0.4373	0.3486	0.4259
adj^{binary},	0.5302	**0.5537**	0.5419
adj^{freq}	0.3794	0.3113	0.3718
adj^{freq},	0.4610	0.4728	0.4685
New			
	fast_skip	*w2v_cbow*	*w2v_skip*
default	**0.4843**	0.2606	0.4301
adj^{binary}	0.2845	0.2051	0.2789
adj^{binary},	0.4258	**0.4854**	**0.4494**
adj^{freq}	0.2309	0.1772	0.2268
adj^{freq},	0.3700	0.4346	0.3895

TABLE VI
AVERAGE PEARSON CORRELATION COEFFICIENTS BETWEEN *default* AND OTHER SIMILARITY MEASURES OVER ALL TOPIC NETWORKS BUILT FROM FORTY FOS.

	Existing		
	fast_skip	*w2v_cbow*	*w2v_skip*
vs adj^{binary}	**0.8955**	**0.9739**	**0.9334**
vs adj^{binary},	0.8884	0.9107	0.9263
vs adj^{freq}	0.8244	0.9548	0.901
vs adj^{freq},	0.7969	0.8312	0.8665
	New		
	fast_skip	*w2v_cbow*	*w2v_skip*
vs adj^{binary}	**0.8848**	**0.9703**	**0.9208**
vs adj^{binary},	0.8479	0.8762	0.8874
vs adj^{freq}	0.8361	0.9490	0.8960
vs adj^{freq},	0.7597	0.7883	0.8195

default vector comparison. adj^{binary}, consistently provided relatively higher similarity values for new topics yet was unable to surpass the *default*. Both adj^{binary}, and adj^{freq}, resulted in higher similarities than unmodified counterparts, indicating the negative impact of scalar values when comparing topic semantics.

Correlations between the results of different vector comparisons were conducted to analyze the possibilities of topic-specific outliers causing such differences. Table VI shows the four alternative similarity measures compared against *default* resulted in high correlations over all forty FoS with adj^{freq}, having the lowest correlations over 0.75, indicating that there are no topic-specific variations in the result. Decreasing correlations from adj^{binary} to adj^{freq}, indicate domain differences are amplified when more calculations are done to measure the topic-neighbor semantic similarities. The information loss is compounded at each step, which led to less consistent outcomes.

A more detailed analysis of the cosine similarities showed that the non-medical related topic networks resulted in better topic semantic predictions compared to the medical-related counterparts when using *fast_skip* embeddings, using the *default* similarity measure. Fig 2 illustrates average cosine similarities between word embeddings of topics and their previous neighbors over all forty topic networks used in the experiments, displaying consistently higher values for topics that were already present within the networks compared to the newly introduced ones. This shows that nine out of ten non-medicine-related FoS reside within the top 20 except *earth science*, which is ranked 21st, with differences between topic networks significant with p-values on average $< 1e^{-100}$. This reflects the interdisciplinary nature of medicine-related research fields on which the word embeddings were learned, as the FoS-specific origin transition is not conducted with *default*. Background knowledge such as *computer science* or *mathematics* is closely related to the application and analysis of the medical procedures. Human-related aspects of medicine also necessitate knowledge on related fields such as *law* and *marketing* for appropriate use and dissemination of the research outcomes. For example, geography-related

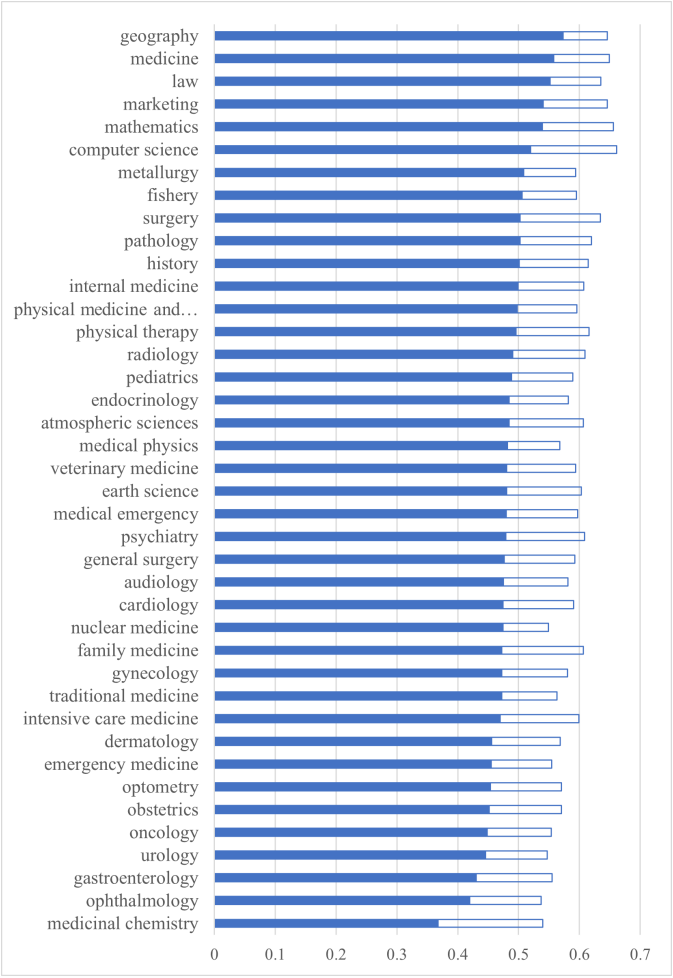

Fig. 2. The cosine similarities between topics and their neighborhoods in the previous timeslots for forty FoS, using *fast_skip* embedding and *default* measure. Average for *new* and *existing* topics are respectively shown as solids and hollows.

topics are researched with regard to many medical fields for medical applications (*geographic information system software* for *emergency medicine*), demographic analysis (*geographical diversity* for *family medicine* and *geographic population* for *general surgery*). Such prevalent connections to generic non-medical topics resulted in the non-medical topic networks showing higher performances over specialized medical fields such as *dermatology*, *ophthalmology*, or *traditional medicine*.

When word embedding vectors were adjusted to the fields, the medicine-related fields resulted in higher similarity measures as shown in Fig 3. The figure illustrates average cosine similarities between word embeddings of topics and their previous neighbors, comparing the average between medical and non-medical related fields. The figure shows that more alterations are done to the similarity calculation, the more *new* topics get classified better than the *existing* ones. While the overall performance is diminished compared to the *default*, all of the four similarity measure variations captured semantics of future topics better in medical fields. This is because the

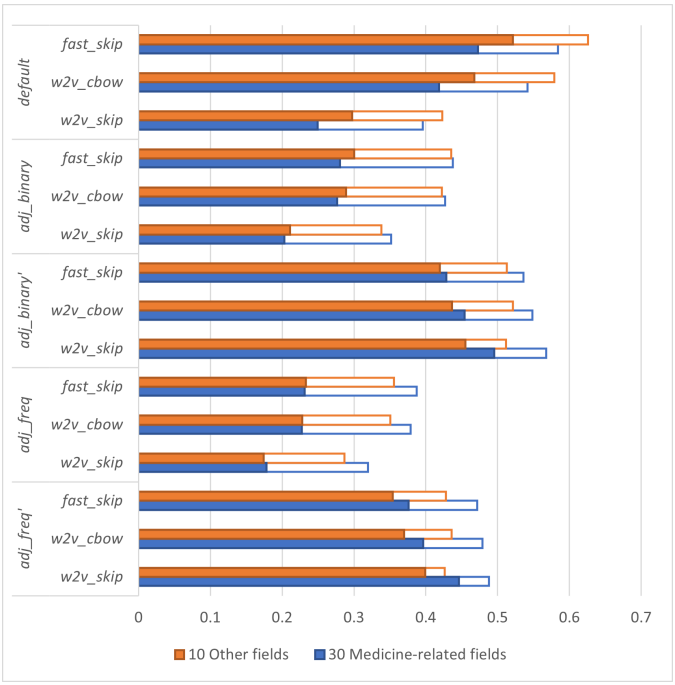

Fig. 3. The cosine similarities between topics and their previous neighborhoods for medicine-related and other fields, using three embedding sets and five measures. Average for *new* and *existing* topics are respectively shown as solids and hollows.

words were trained under the medical background, and words associated with the other topics have distinctive differences from others, resulting in more drastic changes with origin transition.

VI. Conclusion

A network-based topic evolution was proposed to bypass the innate limitations of text-based topic evolution and provide better event detection and prediction. The method introduced neighborhood information in topic networks to represent topics intersecting over time, successfully capturing complex events such as *merge* and *split*. This paper showed that text-based semantics can be attached to the network-based topic models after evolutionary events were captured. Semantic similarities between the detected topics and the actual result showed medium correlations over forty different domains, validating the assumption that the network-based topic evolution method captures semantically consistent topics without accessing textual data.

The results showed that topics share moderate to high degrees of semantic similarities to their neighbors in topic networks, even when the topics are newly introduced to the network without any previous interactions with existing nodes. The Fasttext embedding resulted in the best result combined with the skip-gram model, showing semantic similarity over 0.48 on average. The outcomes were consistent over forty different topic networks over various research domains, including fields that were not directly related to the document collection on which word embeddings were learned. Consistent performances shown across the 40 tested datasets indicate that the proposed method is not domain-specific and would be generalizable to other research fields not tested within the experiments as long as the word embedding is learned from a domain with acceptably wide research interests. The Medline papers used for the training cover heavily interdisciplinary medical fields, thus satisfying the requirements in the results. Word embedding trained on a domain-specific document set would result in better performances as well; the FoS *medicine* was shown as one of the top-performing topic networks. The general nature of training material resulted in several low-performing topics over various domains which can be attributed to either homonymy or polysemy. Acronyms were especially susceptible to homonymy issues as they can represent different words. Polysemy on the other hand was noticeable in matters such as genes or enzymes when specialized medical domains utilize the matter in unconventional applications. Word embedding trained with general medical publications are distant to specific words in these cases.

Existing topics showed higher similarities to their past neighbors compared to the *new* topics over all combinations of topic networks, word embeddings, and similarity variations. The difference between the two topic types can be attributed to the fact that the new topics had less time to appear in conjunction with the other topics in the research articles, having less chance to be tuned along with the others. While moderate semantic similarities indicated semantic similarity over topic network alone was not accurate enough to detect topic evolution on its own, consistent differences between topic types suggest that the idea of network-based topic labeling is valid. The utilization of non-direct neighborhoods, connection weights, and semantic distances could improve the accuracy of neighbor-based topic labeling. Reducing overlapping acronyms and differentiating domain-specific topics, word embedding filtering, or domain-specific training would alleviate the homonymy and polysemy issues and will be investigated in future works.

Acknowledgment

Effort sponsored by the U.S. Government under Other Transaction number W9124P-19-9-0001 between AMTC and the Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

References

[1] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou, "Topic evolution in a stream of documents," in *Proceedings of the 2009 SIAM international conference on data mining*. SIAM, 2009, pp. 859–870.

[2] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 198–207.

[3] S. Jung and W. C. Yoon, "An alternative topic model based on common interest authors for topic evolution analysis," *Journal of Informetrics*, vol. 14, no. 3, p. 101040, 2020.

[4] S. Jung and A. Segev, "Analyzing the generalizability of the network-based topic emergence identification method," *Semantic Web*, vol. 13, no. 3, pp. 423–439, Apr. 2022.

[5] N. Ilhan and Ş. G. Öğüdücü, "Feature identification for predicting community evolution in dynamic social networks," *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 202–218, 2016.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[7] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of latent semantic analysis*. Psychology Press, 2007, pp. 439–460.

[8] Y. Shao, X. Li, Y. Chen, L. Yu, and B. Cui, "Sys-tm: A fast and general topic modeling system," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2790–2802, 2019.

[9] B. Chen, S. Tsutsui, Y. Ding, and F. Ma, "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval," *Journal of Informetrics*, vol. 11, no. 4, pp. 1175–1189, 2017.

[10] C. Balili, A. Segev, and U. Lee, "Tracking and predicting the evolution of research topics in scientific literature," in *2017 IEEE international conference on big data (big data)*. IEEE, 2017, pp. 1694–1697.

[11] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[12] A. A. Salatino, F. Osborne, and E. Motta, "How are topics born? understanding the research dynamics preceding the emergence of new areas," *PeerJ Computer Science*, vol. 3, p. e119, 2017.

[13] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: how can citations help?" in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 957–966.

[14] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 233–240.

[15] J. G. Fiscus and G. R. Doddington, "Topic detection and tracking evaluation overview," in *Topic detection and tracking*. Springer, 2002, pp. 17–31.

[16] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[17] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," *Advances in neural information processing systems*, vol. 17, 2004.

[18] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data mining and knowledge discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[19] M. Li and Y. Chu, "Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis," *Journal of Information Science*, vol. 43, no. 6, pp. 725–741, 2017.

[20] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[21] H. C. Ozmutlu and F. Çavdur, "Application of automatic topic identification on excite web search engine data logs," *Information Processing & Management*, vol. 41, no. 5, pp. 1243–1262, 2005.

[22] T. Furukawa, K. Mori, K. Arino, K. Hayashi, and N. Shirakawa, "Identifying the evolutionary process of emerging technologies: A chronological network analysis of world wide web conference sessions," *Technological Forecasting and Social Change*, vol. 91, pp. 280–294, 2015.

[23] A. Salatino, *Early Detection of Research Trends*. Open University (United Kingdom), 2019.

[24] S. Jung, R. Datta, and A. Segev, "Identification and prediction of emerging topics through their relationships to existing topics," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 5078–5087.

[25] A. A. Salatino, F. Osborne, and E. Motta, "Augur: forecasting the emergence of new research topics," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 2018, pp. 303–312.

[26] S. Jung, T. M. Lai, and A. Segev, "Analyzing future nodes in a knowledge network," in *2016 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2016, pp. 357–360.

[27] S. Jung and A. Segev, "Analyzing future communities in growing citation networks," *Knowledge-Based Systems*, vol. 69, pp. 34–44, 2014.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[30] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, and R. Rogahn, "A review of microsoft academic services for science of science studies," *Frontiers in Big Data*, vol. 2, p. 45, 2019.

[34] Z. Shen, H. Ma, and K. Wang, "A web-scale system for scientific knowledge exploration," *arXiv preprint arXiv:1805.12216*, 2018.

[35] S. E. Hug, M. Ochsner, and M. P. Brändle, "Citation analysis with microsoft academic," *Scientometrics*, vol. 111, no. 1, pp. 371–378, 2017.

[36] V. Major, A. Surkis, and Y. Aphinyanaphongs, "Utility of general and specific word embeddings for classifying translational stages of research." in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 1405.

[37] L. Wu, I. E. Yen, K. Xu, F. Xu, A. Balakrishnan, P.-Y. Chen, P. Ravikumar, and M. J. Witbrock, "Word mover's embedding: From word2vec to document embedding," *arXiv preprint arXiv:1811.01713*, 2018.

[38] Y. Xiao, A. Krishnan, and H. Sundaram, "Discovering strategic behaviors for collaborative content-production in social networks," in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2078–2088. [Online]. Available: https://doi.org/10.1145/3366423.3380274

[39] E. Bongen, F. Vallania, P. J. Utz, and P. Khatri, "Klrd1-expressing natural killer cells predict influenza susceptibility," *Genome medicine*, vol. 10, no. 1, pp. 1–12, 2018.

[40] N. J. van Beveren, L. C. Krab, S. Swagemakers, G. Buitendijk, E. Boot, P. van der Spek, Y. Elgersma, and T. A. van Amelsvoort, "Functional gene-expression analysis shows involvement of schizophrenia-relevant pathways in patients with 22q11 deletion syndrome," *PLoS One*, vol. 7, no. 3, p. e33473, 2012.

[41] R. Jansen, B. Penninx, V. Madar, K. Xia, Y. Milaneschi, J. Hottenga, A. Hammerschlag, A. Beekman, N. Van Der Wee, J. Smit *et al.*, "Gene expression in major depressive disorder," *Molecular psychiatry*, vol. 21, no. 3, pp. 339–347, 2016.