Identification and Prediction of Emerging Topics through Their Relationships to Existing Topics

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

Rituparna Datta
Department of Computer Science
University of South Alabama
Mobile, USA
rdatta@southalabama.edu

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

***Abstract*— Understanding the current research topics and their histories allow researchers to focus their capabilities on the current research trends. The field of topic evolution helps the understanding by automatically model and detect the set of shared research fields in the academic papers as topics. The authors propose a novel topic evolution method for identifying and predicting the emergence of new topics under the assumption that neighborhoods of new topics in the future have distinguishable structural features. Eight journals were selected from the Microsoft Academic Graph dataset, each representing topics networks with varying size, history, and research domains. Both retrospective classification and prospective prediction showed promising performance with classifications above 0.89 for six journals and coefficients of determination exceeding 0.95 for five journals. The result showed both the retrospective identification and the prospective prediction can be done, validating the assumption that topic evolution events can be predicted with a network-based approach.**

Keywords— Topic Evolution, Topic Prediction, Network-based Topic Modeling, Scientometric

I. INTRODUCTION

Scientific researches are conducted to contribute towards new discoveries at the boundary of the knowledge, expanding the boundary or filling the gaps. Knowing where the boundary is therefore an integral part of any research activities, which is done by understanding the recent research topics and their histories in the domain. Topics in the academic papers represent the set of shared themes, of research fields, among the fellow researchers. They can appear in various forms, including the philosophical category of the research, applications of the technology, and specific algorithms. Identifying such topics in the academic papers are therefore a crucial part of research activity. Author-specified keywords or automatically extracted topic models are used to filter out the related papers from the overflowing number of new articles and understand their relevance to the conducted research. Researchers understand the topics by reviewing a multitude of articles, internalizing the evolution occurring within the researchers' fields of interest which in turn allows them to ascertain the desirable paths the current and future research can take. A better understanding of such knowledge leads to better research aimed at the topics with high demands, hence holds academic values as well as industrial uses.

While understanding the evolution of research topics is one of the inherent tasks of researchers, automation of such a process is not easy with the complexities involved. Topics evolution is not only done through maturation over time with continuous research on the topic, but also affected by changes in the interests of authors, background topics, targeted applications, and external circumstances. Traditional topic evolution methods approach this problem by utilizing text-based topic models to understand the topic in a given document collection and track topical changes over time. Topic modeling methods extract statistical constructs based on word co-occurrences in the given document collection, where changes in topics can only be measured by the differences between the content of two topics; connections and correlations between different topics were not incorporated into the traditional topic modeling methods [1]. Topic evolution methods are therefore mostly limited to identifying content transition within a given topic, not how it is correlated to other topics. Merge and split events are less sought after, with limited success [2]. Neglecting such events, topic evolution based on traditional topic modeling methods is not suited to identify and predict new topics as new topics in most cases result from merging or splitting of existing topics.

Classifying topic evolutions is different from predicting such events that the latter requires prospective capability. Topic evolution prediction is therefore inherently limited with the nonexistence of textual data representing the future documents. One of the benefits of using a network-based approach is that the classification results can be extrapolated to allow prediction on when new topics are formed along with their ancestors. The main contribution of this paper is to propose a topic evolution method based on network-based topics, offering new functionalities by defining new topics with their neighboring topics. The goal of the method is to capture the emergence of new topics, which can be explained by their correlation to the existing topics. This can be formalized as classifying subgraphs in the given topic network as to-be-neighbors of new topics in the future based on their graphical properties. 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 timeslot *y* is either *new*, appearing for the first time in *y* for the given topic network, or *old*. Binary machine learning algorithms are

trained using the neighbors of each node in the previous years, classifying the neighbor subgraphs in the past having *new* or *old* topics as their future neighbors. A prospective approach is also tested to analyze the possibility of new topic prediction without the knowledge of neighboring nodes. Two community detection algorithms are run on differentially flattened topic networks, showing low mean squared error (mse) values from regression analysis between structural properties of communities and future new topics along with neighbors of future new topics associated with them. Both the identification and prediction of new topics were experimented on eight topic networks generated from publications of eight journals from the Microsoft Academic Graph[1] dataset, ranging from 194 years old New England Journal of Medicine (*NEJM*) to 14 years old Journal of Informetrics (*JoI*). The experiment results showed that it is feasible to classify the generation of new topics based on a given topic list using the structural properties of subgraphs and their members.

Section II reviews the related work on topic evolution, previous attempts on the prediction of new topics as well as background research for the proposed method. Section III and IV details the proposed method and experimentation, and the experiment results are shown in Section V.

II. Related Work

A. Identifying Evolution of Topics

Automatically identifying topical changes within the document set requires methods to extract machine-readable topics from the collection. Topic modeling provides a statistical approach to discovering *topics* within a given corpus, latent semantic structures in the form of word-popularity sets based on the statistical distribution and word co-occurrences. Latent Dirichlet Allocation(LDA) [3] finds latent topics within a document collection and is one of the most widely used topic modeling methods on which many other methods are based on [4], [5]. Word-topic links are iteratively assigned with word co-occurrences between documents; topics, defined as word distributions over a corpus dictionary, are then assigned to each document [6]. Topic evolution aims to identify the evolution of such topics in a sequentially ordered document collection. Document collection is first divided either uniformly or irregularly [7] into sequentially-ordered sub-collections on which topic models independent of the neighboring sub-collections are generated. Temporal topic models are then connected over time with similarity measures, and changes in the topics are sequentially analyzed to identify the evolution of topics. Dynamic topic models [8] is one of the early implementations of topic evolution, focusing on capturing the changes within a set of chained topics with fixed timeslots where Kalman filter and wavelet regression is used to approximate natural parameters of the topics found at different time slices. Evolutionary theme pattern mining is tried to capture not only the changes within each topic but also the sequential connections over multiple topics [9]. Kullback-Leibler divergence is used as a distance metric between topics, and the topics on different timeslots are designated as having an evolutionary transition when their distance stays below a threshold set specifically for different datasets. Collection of such evolutionary transitions result in detecting merge and split events over time as multiple connections are allowed between different topics. Similar approach is made by utilizing cross-citations between topic pairs' member documents as well [10].

Topic evolution in conjunction with bibliographical dataset analysis has been tried by numerous researchers to better identify the topic evolution events. The citation contexts are used in an iterative topic evolution learning framework to increase the performance of topic evolution with better topic models [11], where the document collection is expanded by the documents cited by its members. Inheritance topic model [12] is utilized to classify papers into autonomous parts with originalities and parts inherited from cited documents. Differentiating two parts allowed the method to overcome the topic dilution with cited papers, generating more new topics compared to LDA-based approaches. A more recent approach to topic evolution utilizes communities of keywords in a dynamic co-occurrence network [13]. Medical subject headings dataset from the PubMed[2] was used to build a filtered co-occurrence network of major subjects within the medicine domain divided into five-year snapshots. Word clusters were found and linked to generate the evolution of topics over time. Topic evolution based on two-tier topic models is tried for a better merge and split detection, where topic correlations in the same timeslot are used to identify topic evolution [2]. Timeslot-specific local topics are extracted from yearly divided sub-collection of documents, while time-spanning global topics are retrieved using the whole corpus. Global topics stays static, having connected to dynamic local topics at each timeslot with cosine similarities above a given threshold. Changes in the number of local topics connected to global topics are then used to define the topic evolution events; decreased and increased number of local topics connected to a global topic respectively represent merging and splitting of the topic.

B. Identifying and Predicting New Topics

Topic Detection and Tracking (TDT) [14] aims to capture the appearances of new topics in a continuously generated text data in real-time; a topic is defined as "*a seminal event or activity along with all directly related events and activities*" [14]. First story detection (FSD) is one of the parts of TDT research tasks, where the goal is to search and organize new topics from multilingual news articles, which is translated as identifying the first article introducing the new story [15]. Topic-conditioned FSD with a supervised learning algorithm first classified news articles into a set of pre-defined topic categories before identifying novelty within each topic [39]. FSD is also used in conjunction with document clustering to identify the earliest report to a certain event in news articles [16]. Identification of emerging topic trends has led to the division of research front and intellectual base, where the latter is an established foundation of domain knowledge on which the former is built. The underlying assumption is that the citation and co-citation between articles transfer the existing knowledge from the intellectual base to the research front. CiteSpace II [17] further utilized a keyword co-occurrence

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

[2] https://www.ncbi.nlm.nih.gov/pubmed/

relationship by employing a bipartite graph of keywords and articles. Research front terms are identified by the sharp frequency growth then used to identify research front articles, which in turn are absorbed into the intellectual base in the next time slice. Burst term detection, in conjunction with keyword co-word analysis, allows multi-dimensional exploration of the research front in question [18]. While these approaches allow the detection of merging and splitting of time-spanning topics and their transitional ratio at temporal level, the use of the text-based topic models inherently limits the predictive capabilities; the evolutionary events such as emergence, merge, or split can only be retrospectively analyzed once the topic is captured from the document set. The author previously introduced the use of author groups from a bibliographic dataset for determining topics connected over time by authors, showing that when topics defined by the authors are used instead of NLP-based topic models, topic evolution on the temporal network is possible; the topic evolution events defined by the network structures and therefore a predictive analysis is possible [1].

On top of the emergence events detected by the appearance of topic models dissimilar to the ones in the previous timeslots, there are a number of research dedicated to identifying new topics with a varying definition of the topic. One such field is the new topic identification, where the topic is defined as the entities the user is interested in during the search engine querying session; the query patterns and the intervals between queries are used to identify topics [19]. Neural network (NN) is introduced to reduce the errors in new topic estimations based on typos by utilizing the character n-gram method to bypass spelling errors in the queries [20]. There are also several researchers focusing on utilizing the queries' statistical characteristics such as search patterns, frequency of queries, and the relative position in the querying sessions [21]. Technology forecasting [22] is another field of research aiming to predict the characteristics of technology in the future; the technology, or topic, is defined as a representative keyword instead of a statistical model. Various techniques from simple extrapolation to organization management [23] and fuzzy NLP [24] are used to identify and predict changes in technology indicators[25]. Multiple applications of the predictive topic evolution have been proposed, including a semi-manual technology trend analysis which was done to identify the roots of new technologies with their projected impact on the research field [26].

The authors previously proposed a technology trend analysis approach with multiple data sources to show that while different data sources exhibit different forecast speeds, predicting the growth and shrinking in technology trends is possible extrapolating on a previously known technology growth curve [27]. A network-based approach was proposed by the authors in previous research to overcome the rigidity of trend-based forecasting where the prediction is dependent on the type and shape of the technology growth curve used. Node prediction based on preferential attachment link prediction is proposed to classify nodes in citation networks whether they have a connection to a new node in the future [28], labeling the new nodes by its neighboring nodes [29]. This showed that predicting nodes in bibliographic networks is possible based on the structural properties of the network. More complex contexts of the new nodes in knowledge networks were extracted by identifying the neighbors of the new node in the past timeslot to formulate the context of new node solely based on the metadata of its to-be-neighbors [30]. This paper is proposing a possibility of detecting and predicting the emergence of new topics from a topic network, where the pre-defined keywords are used to represent topics in a given document collection.

III. Network-based New Topic Identification and Prediction

A. Generating Topic Networks

Identifying emerging topics in a bibliographic dataset equates to identifying new nodes in a topic network. NLP-based topic modeling can be used in a retrospective analysis as the document set for the target topic is already present, but the goal of this paper is to present the prediction of new topics therefore textual metadata is not considered for analysis, and only the graphical structures are used.

The co-occurrence relationships R_y between topic set V with node u and the first year of usage $firstY$ on the bibliographic dataset are retrieved for each timeslot y, where R_y is the weighted edge set between nodes u, v with co-occurrence frequencies in y as weight w_y and G_y represent topic co-occurrences at year y within the target knowledge domain.

$$G_y = (V, R_y), \text{ and} \tag{1}$$

$$V = (u, firstY)\ R_y = (u, v, w_y) \tag{2}$$

Graphs in subsequent years show the evolution of topic co-occurrence patterns over time, therefore multiple consecutive graphs are merged to generate a topic network. Given the layer size l, the topic network for year y is defined as

$$MG_{y,l} = \{G_{y-l}, \ldots, G_y\} \tag{3}$$

which is then flattened to generate a single-layer topic network. Differential flattening [31] is used to incorporate the varying importance of different layers with layer coefficient α_i.

$$E_{y,l} = \{u, v, w\},\ w = [\alpha_0, \ldots, \alpha_l][w_{y-l}, \ldots, w_y]^T, \text{ and} \tag{4}$$

$$\alpha_i = 1 / (i + 1)^2 \tag{5}$$

Instead of utilizing optimization methods, the layer coefficients are calculated as inverse squared year distance (5) to represent exponentially diminishing importance of distant topic co-occurrences, with 1 added to deal with zero distance values. The result is a single-layer topic network at given year y containing l previous timeslots on a topic set $V_{y.l}$ containing the topics used in $E_{y,l}$.

$$T_{y,l} = (V_{y,l}, E_{y,l}) \tag{6}$$

B. Classifying Subgraphs by the Common Neighbors

The proposed method is run on the topic network $T_{y,l}$ in (6), where topics in year y and l previous timeslots 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 extracted to build a set of neighborhoods $N_{y,l}$ from $T_{y,l}$. Each

neighborhood is then categorized into two groups by the age of v calculated by $firstY(v) - y$ categorizing whether the topic v first appeared in the given year y ($firstY$ of v equals to y). The state of v $C(v)$ is calculated as the ceiling of normalized topic age, where the *new* topics are denoted by $C(v) = 0$. Any preexisting topics have non-zero ages, and the normalized ceiling function result in $C(v) = 1$.

$$N_{y,l} = \{neighbors(v, y) \mid v \in V_{y,l}\}, \text{ and} \quad (7)$$

$$C(v) = \lceil (firstY(v) - y) / max(firstY(v) - y) \rceil$$

More prominent topics are likely to cooccur with more topics, therefore top 100 topics with the largest number of nodes in $N_{y,l}$ are selected for each label $C(v) = 0$ and 1. In case the number of instances for one label is below 100, then the number of v for the other label is reduced further to have the same number of instances for both labels.

Evolution of existing topics such as merge and split are not targeted hence there is no need to train for the gradual evolutions within existing topics; only static features are used in the experiment. TABLE I. shows the list of 15 structural features of the neighbor subgraphs used to train binary classifiers. These features characterize the subgraph quality in several aspects and are grouped by the component they are used to measure, including six properties related to the whole subgraphs, four average values of node properties, two properties related to the number of edges, and three properties weighted by the topic co-occurrence frequencies.

TABLE I. STRUCTURAL FEATURES USED IN THE EXPERIMENT.

Features used	Description
Subgraph	
Node Count	Number of nodes
Cohesion	Number of internal/external edges
Density	Number of observed/possible edges
Transitivity	Number of observed/possible triangles
Normalized Triangles	Number of triangles/nodes
Mean Shortest Path	Mean of all node pairs' shortest paths
Nodes	
Mean PageRank	Mean PageRank for nodes in subgraph
Mean Degree Centrality	Mean degree centrality in subgraph
Mean Betweenness Centrality	Mean betweenness centrality in subgraph
Mean Node Age	Mean age of nodes in subgraph
Edges	
Edge Count	Number of edges in subgraphs
Mean Degree	Mean degrees in subgraph
Weighted	
Mean Degree Weighted	Mean degree with edge weights
Mean Edge Weighted	Mean edge weights
Mean Clustering Coefficient	Mean weighted clustering coefficient

The emergence of new topics is the only event being searched, therefore the binary classification on year y is trained by neighbor subgraphs in previous years and tested at y. Sets of open neighborhoods $Train_{y,l,t}$ and $Test_{y,l}$ are generated using t previous topic networks. The same set of neighbors n in $neighbors(v)$ are used to identify open neighborhood subgraphs of v in multiple previous timeslots, denoted by $T_{k,l}(n)$ where $y\text{-}t \leq k \leq y$.

$$sub(v,y,k) = \{(n, \{n_i,n_j\}) \mid n \in neighbors(v,y), \{n_i,n_j\} \in E_{k,l}\},$$

$$Train_{y,l,t} = \{sub(v,y,y\text{-}t) \cup \ldots \cup sub(v,y,y\text{-}1) \mid v \in V_{y,l}\}, \text{ and}$$

$$Test_{y,l} = \{sub(v,y,y) \mid n \in N_{y,l}\} \quad (8)$$

The length of flattened topic network l used in conjunction with variable training sizes t allows a different view of the topic interactions over the years; larger l results in more integrated topic interactions in each instance, while larger t results in more number of instances to better train the classifiers. Neighbor subgraphs in (8) represent interactions between direct predecessors of *new* topics and neighbors of preexisting *old* topics, which are assumed to have distinguishable structural features. The classification accuracies and area under the ROC curve (AUC) based on these features are compared to show the effect of different classification approaches on the performance of the proposed approach.

C. Predicting New Topics from Communities

Successful classification of subgraphs based on the state of their common neighbor validates that given the correct neighborhood for the topic, one can ascertain whether the topic would be new to the given domain. The limitation of this approach is that it cannot be prospective in practice. The neighborhoods are not necessarily connected nor have strong connections within them, therefore retrieving such subgraphs without the presence of the seed set V would become a problem of selecting a random number of random nodes. The possible combinations reach well over trillions in large networks, rendering the approach impractical.

Existing community detection algorithms are used as alternatives to the random subgraph sampling to enable prospective prediction. The unweighted variant of Clauset-Newman-Moore algorithm [32] maximizing the modularity of clusters (*Greedy)* and weighted Infomap algorithm [33] based on Map equation (*Infomap)* are implemented to generate communities in the topic networks $T_{y,l}$ which are built using the combinations of the same variables in (1). There is no one-to-one relationship between the communities and their common neighbor topic, therefore classification labels C cannot be identified. Regression analysis is done on the communities instead to prospectively predict the formation of new topics and neighbors of new topics in the future. The same set of features in TABLE I. are used for regression analysis as communities are similar to the neighborhoods that they both are subgraphs of the given graph and shares the same underlying structure. No filtering is done for the data instances, however, as there are no labels to balance and communities detected share similar sizes relative to the neighborhoods found in the previous section.

Four dependent variables are introduced for the regression analysis. $Vn_{y,l}$ is the set of new topics first appearing in year $y+1$ and $Nn_{y,l}$ is the combined set of all the neighbors of $Vn_{y,l}$; the former represents a set of new topics first observed in the following timeslot, and the latter represents the list of topics in the given timeslot which would cooccur with the new topics in the future. As the communities do not share the same structure patterns, all members of the new topics and neighbor nodes are considered as single sets instead.

$$Vn_{y,l} = \{v \mid v \in V_{y+1,l},\ C(v) = 0\}$$
$$Nn_{y,l} = \{n \in \cup neighbors(v) \mid v \in V_{y,l},\ C(v) = 0\} \quad (9)$$

Two dependent variables are generated for each community found; *NewTopicCount* measures the number of new topics linked to the given community members in the following timeslot, and *NewTopicFreq* measures the frequency sum of links to the new topics. *NeighborCount* and *NeighborRatio* on the other hand measure the number of to-be neighbors of new topics in $y+1$, each for the total number and ratio of community members having connections to the new topics in the following timeslot.

IV. Experiments

A. Dataset Preprocessing

Multiple topic networks were generated from bibliographic records extracted from the Microsoft Academic Graph (MAG) [34], which is a heterogeneous bibliographic dataset [35]. MAG is selected as the source dataset for two reasons. Firstly, it was deemed competitive with major bibliographic search engines such as Google Scholar or Scopus even with relatively recent creation [36]. Secondly, MAG has a built-in ontology called fields-of-study (FoS) representing each paper with different hierarchical concepts [37]. A six-level hierarchy of concept is generated each month using knowledge base type prediction with Wikipedia articles, employing graph link analysis and convolutional neural network methods. They 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 are not used in the experiment.

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, more than 1.5 billion citation links, and 1.3 billion paper-FoS links. Analyzing the whole graph would be too complex to compute, therefore eight journals in TABLE II. are selected to represent subsets of topics shared by different research communities. *Nature*, *Science*, *NEJM*, *Cell*, and Physical Review (*Phys.Rev*) are selected as reputable venues sharing broad research interests, while Journal of High Energy Physics (*HEP*), Knowledge Based Systems (*KBS*), and *JoI* are selected as venues with more focused research topics.

TABLE II. Descriptions of Eight Journals in MAG Dataset in February 2020.

Journal	JournalId	Rank	Size	Topic	Year
Nature	137773608	1st	217,170	69,188	1869 ~
Science	3880285	2nd	161,856	61,385	1880 ~
NEJM	62468778	7th	83,581	33,778	1827 ~
Cell	110447773	14th	18,396	14,672	1973 ~
Phys.Rev	54862371	27th	40,592	8,703	1893 ~
HEP	187585107	178th	28,825	5,461	1997 ~
KBS	10169007	3,107th	4,316	6,508	1987 ~
JoI	205292342	9,612th	854	1,582	2007 ~

TABLE II. shows the eight journals with ranks measured by the possible importance along with the number of papers and related topics recorded in the MAG dataset, and the first year publication under the journal is recorded, where a wide range of size and starting years are included in the journal subsets. Eight journal-specific datasets are extracted into the SQL databases using high-performance computing service by Alabama Supercomputer Authority[3]. All data rows in the *Paper* table containing the matching *JournalId* are retrieved, then rows matching the filtered papers in *PaperFieldsOfStudy* and *FieldsOfStudy* tables are retrieved for FoS used in the journal and how they are assigned to individual publications. With a series of SQL queries, *FirstUsedYear* column is added to *FieldsOfStudy* tables to represent the first year *firstY* the given FoS is used within the journal, and *FOSneighborCount {Node1, Node2, Year, Frequency}* table is created to represent undirected links within each journal with node pair u, v, year y, and frequency w, where FoS are the nodes and the links represent the two FoS assigned cooccurring in the same publications. *Frequency* shows the co-occurrences between two FoS, which is divided for each *year* to distinguish different FoS links and weights at different years.

B. Generating Topic Networks

After the dataset preprocessing is done, the topic network $T_{y,l}$ in (1) for each journal j is generated with different combinations where j is the target journal, y is the target year, and l is the layer size.

$$y = [2000, \ldots, 2020],\ l = [1, 5, 10], \text{ and}$$
$$j = [\text{'Nature'}, \text{'Science'}, \text{'NEJM'}, \text{'Cell'}, \text{'Phys.Rev'}, \text{'HEP'}, \text{'KBS'}, \text{'JoI'}], \quad (10)$$

The target year y is selected to retrieve the detection of newly used topics in the 21st century, and the layer size l dictates the number of years to build the topic network for the analysis. For each journal j, SQL queries are run on the *FOSneighborCount* table to extract topic co-occurrence with $y - l < FOSneighborCount.Year \leq y$ where the *Year* column in the *FOSneighborCount* table represents the year the topics cooccur. The resulting edge data R_y is used to build a topic network for the given combination of variables using equations

[3] https://hpcdocs.asc.edu/

from (3) to (6). The natural log of the frequency is used as weight w_y in (4) to smooth the disparity between minimum and maximum frequencies, with 1 added as a constant to avoid getting 0 when frequencies equal 1.

$$w_i = \{ln(freq + 1)\} \quad (11)$$

A combination of 21 years and three layer sizes generates 63 topic networks iterations per a journal, resulting in a total of 504 networks for eight journals.

C. Classifying Subgraphs by the Common Neighbors

The total size of $N_{y,l}$ for all 8 journals reaches 945,424 sets, having 1,876 sets for each topic network on average. It is also highly skewed, with each iteration have on average 5.95 times more *old* topics than *new* topics; only 14.41% are for the new topics. Data downsampling is done to remove the total number of data and balance the number of labels for the classification. Isolated nodes *v* are ignored as they do not have any neighbors to analyze the structure.

1.11% of the journal, target year, and year length combinations result in a single label result where all instances belong to only one of the labels. *JoI* is responsible for all-new topics as the first publication entry in the dataset starts in the year 2007; all topic is new to the journal's first year. The behavior of topics in an initialization stage of the dataset is not a scope of this paper and hence removed from the experimentation. Training cannot be conducted with a single label dataset therefore iterations with all-old topics are removed from the experiment.

$$t = [1,3,5,7,9] \quad (12)$$

Five different training sizes *t* were tested to analyze the changes in the performances. Six binary classification machine learning algorithms were implemented in the experiment: *Decision Tree*, *Random Forest*, *K-Nearest Neighbors*, *Logistic Regression*, *Linear Discriminant Analysis*, and *Gaussian Naïve Bayes*. A fixed seed number is given to the random number generators in *Decision Tree* and *Random Forest* to retrieve a reproducible result. The number of neighbors to search in *K-Nearest Neighbors* is set to 5, or the size of *n* if $n < 5$. This is mainly because of *Phys.Rev* where there are very few topics assigned to the publications, reaching as low as 6 total topics in the year 2018. This is because sister journals of *Phys.Rev* such as *Physical Review A*, *Physical Review B*, *Physical Review C*, and *Physical Review D* were created in 2017 with more focused research fields, reducing the number of publications related to the *Physical Review* journal itself.

D. Predicting New Topics from Communities

Linear, quadratic, and cubic regression analysis is done for each of the four dependent variables using all 15 features. The feature sets in non-linear regression are first transformed to polynomial formats containing all the features including derivatives up to the given degree, then linear regression is run on the polynomial features. Training sets and Test sets are divided by the years the same as in the binary classification with different lengths of the training set (7). Regression analysis results are measured with coefficients of determination (R^2) and Normalized Root Mean Square Error (NRMSE). R^2 score was implemented to use the scikit-learn python library's linear regression score function[4], where negative values can be generated for computational purposes; R^2 value of 0 indicates that the model can explain none of the data, therefore any negative values were replaced to 0 for the analysis. NRMSE values follow the same unit as the source data and therefore are normalized by mean ($NRMSE_{mean}$), differences between minimum and maximum ($NRMSE_{minmax}$), and interquartile range ($NRMSE_{interquartile}$) to enable comparison between different variables.

V. Results

A. Retrospective Classification of Neighbor Subgraphs

Majority of features for *new C(v) = 0* and *old C(v) = 1* topics had consistently different values over different combinations of journals, layer sizes, years, and training sizes in (10) and (12). The only exception is *Phys.Rev* having 7 features showing opposite patterns. This is partially explained by the publishing of its sister journals in 2017 indicating the topical evolution was halted in recent years, resulting in statistically insignificant p-values of 0.19 on average for 15 features. Another possible explanation is that the differences in the knowledge domains the journals belong to; *Phys.Rev* is the only medical journal tested. This could indicate that the structural differences between old and new topics are specific to a given domain.

On 7 journals excluding *Phys.Rev* with 20 years and 3 different layer size, *new* topics always had lower values for *Node Count*, *Cohesion*, *Mean Shortest Path*, *Mean Betweenness Centrality*, *Edge Count*, and *Mean Degree* while having higher values for *Density*, *Transitivity*, *Mean PageRank*, *Mean Degree Centrality*, *Mean Edge Weighted*, and *Mean Clustering Coefficient* throughout all the iterations. The average differences between labels were statistically significant for all 15 features. T-test with non-equal variance is done on all 15 features, resulting in p-value ranging from 0.004 for *Mean Node Age* to $4.4e^{-99}$ for *Mean Betweenness Centrality* with 0.0005 as an average p-value for 15 features, 79.44% of all t-test results showing $p < 1.0e^{-10}$. It is however worth noting that differences in *Normalized Triangles*, *Mean Degree Weighted*, *Mean Node Age*, and *Cohesion* were respectively insignificant for 4, 8, 7, and 5 out of 24 Iterations with 8 journals and 3 layer sizes respectively with the majority of features for *Phys.Rev* having p-value > 0.05, indicating that simple value comparison would not perform well for identifying subgraphs having new topics as their common neighbors.

A comparison between different ML algorithms in TABLE III. shows that all six algorithms perform with high accuracy and AUC, indicating that it is feasible to distinguish whether the common neighbor of a given subgraph will be a new topic to the domain. Both the accuracy and AUC increase when *Phys.Rev* is excluded from the calculation reaching up to 0.9684 and 0.9053, indicating that the journal has specific quality hindering the classification performance. *Random Forest* and *Logistic Regression* show the highest values while *Decision Tree* showing the lowest. This is because the datasets

[4] https://scikit-learn.org/stable/index.html

TABLE III. AVERAGE BINARY CLASSIFICATION ACCURACY AND AUC SCORE FOR SIX ML ALGORITHMS.

ML Algorithm	Acc	AUC	Acc[a]	AUC[a]
Decision Tree	*0.8110*	*0.8084*	*0.8491*	*0.8496*
Gaussian Naïve Bayes	0.8420	0.8822	0.8964	0.9356
K-Nearest Neighbors	0.8635	0.8788	0.8962	0.9271
Linear Discriminant Analysis	0.8528	0.8825	0.8856	0.9315
Logistic Regression	**0.8701**	0.9085	**0.9053**	**0.9684**
Random Forest	0.8451	**0.9140**	0.8788	0.9613

[a] Excluding *Phys.Rev.*

used in the experiment are relatively large and complex; feature importance fixed in a single decision tree lessens the chances of better training. There were no anomaly patterns detected with six ML algorithms with different combinations of variables, hence *Logistic Regression* is selected to be analyzed further.

The experiment result showed that there are statistically significant structural differences between subgraphs that would result in new topics and those that would not. Analyzing results for the selected algorithm showed the classification accuracy exceeds 0.895 in 6 out of 8 journals as shown in TABLE IV. where the average value for five training year lengths is shown for each of the iterations over the 20 years. Different l values represent the number of years used to build the topic network in (5), resulting in a slight increase in both the accuracy and AUC in the majority of journals. *Phys.Rev* is the exception in this case, where the larger l result in a further drop in accuracy reaching 0.58 near the 0.5 threshold where it's becoming a random coin flip. This is due to the sudden drop in topic network activities; 548,823 topic co-occurrences were recorded prior to 1970 having 7,128 links each year, while only 3,823 links were recorded from 1970 to 2020. This is explained by the creation of sister journals; 1970 is the year when *Physical Review A, B, C,* and *D* were established. The publication activities were spread across sister journals, effectively rendering structure-based classification void for *Phys.Rev* in recent years.

TABLE IV. BINARY CLASSIFICATION RESULTS BY JOURNALS AND LAYER SIZES IN ACCURACIES (ACC) AND AUC SCORE.

	l = 1		*l = 5*		*l = 10*	
Journal	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***
Nature	**0.9414**	0.9930	**0.9429**	0.9943	**0.9496**	**0.9956**
Science	0.9051	**0.9934**	0.9231	**0.9957**	0.9314	**0.9956**
NEJM	0.9067	0.9886	0.9181	0.9892	0.9182	0.9910
Cell	0.9307	0.9868	0.9289	0.9906	0.9345	0.9906
Phys.Rev	*0.6501*	*0.4569*	*0.6121*	*0.4665*	*0.5754*	*0.4884*
HEP	0.9107	0.9819	0.9144	0.9787	0.9062	0.9730
KBS	0.8245	0.9023	0.8788	0.9403	0.8948	0.9581
JoI	0.6944	0.7391	0.8667	0.9064	0.9295	0.9751

JoI exhibits relatively low accuracy compared to the other journals but experiences a significant increase with l=5. This is likely due to the recency of the journal starting in the year 2007; the structural features have not been stabilized and integrating multiple years was needed to build topic networks with more pronounced differences. The recency of *JoI* can also be observed with a different number of years t in the training set. Fig. 1 shows the accuracy of *Logistic Regression* mostly increases with other journals while *JoI* experiences lower accuracy with larger t. This is because the journal only had 13 years in the MAG dataset, and longer training years result in initial years being included in the training set. The majority of the topics in initial years are by definition new to the topic introduced to the journal, not necessarily representing adaptation or creation of new topics based on past research. Inclusion of years with such erratic patterns would negatively affect the prediction accuracy as seen in Fig. 1.

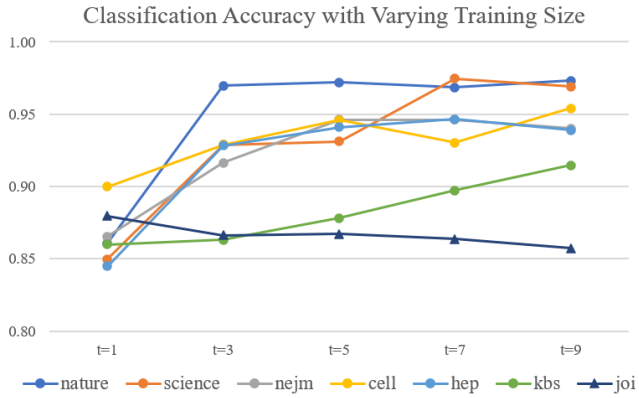

Fig. 1. Binary Classification Accuracy of *Logistic Regression* with Different Training Years t.

B. *Prospective Regression of Communities*

The classification of neighbor subgraphs validated that topic's novelty is statistically correlated to the structure of its neighborhood in previous years. Multiple degrees of regression analysis are done to test if this finding can be applied in a prospective approach for predictions. *Phys.Rev* removed from the analysis as many of the property values result in NaN in later years, which was also the cause of lower classification accuracy shown in the previous section. $T_{2007,l}$ for *JoI* is also excluded from the analysis; 2007 is the first year *JoI* publications appear in the MAG dataset, therefore topics are all treated as new. This is not the normal behavior of topic networks falling outside the scope of this research.

Each of the four dependent variables was first tested if they have significant correlations with 15 properties. Spearman correlation is used as the properties are not normally distributed, to capture non-linear correlations. The average value of correlation coefficients for four dependent variables in Fig. 2 shows that 10 out of 15 properties have moderate correlations with coefficient > 0.4 on average with all dependent variables. Three variables *Node Count*, *Mean Shortest Path*, and *Edge Count* are positively correlated to all four dependent variables with coefficient > 0.5, and five variables *Density*, *Transitivity*, *Mean PageRank*, *Mean Degree*

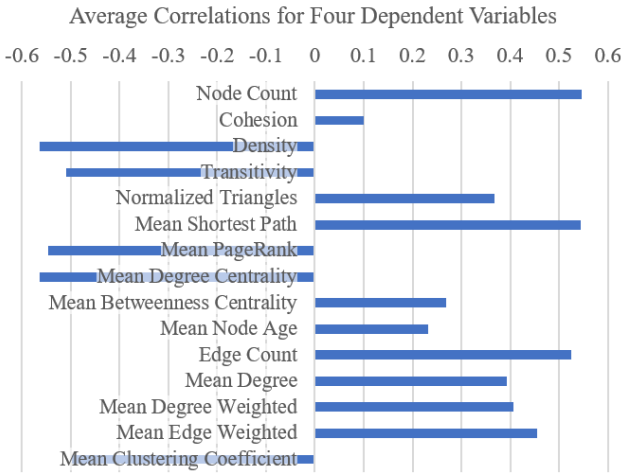

Fig. 2. Average Spearman correlation between 15 independent and 4 dependent variables.

Centrality, and *Mean Clustering Coefficient* are negatively correlated with coefficient < -0.5. Only three variables showed an absolute coefficient value less than 0.3, with *cohesion* showing little to no monotonic relationships to four dependent variables. This suggests the external links from the community are not a good indicator of the community's relationship with new topics in the future, while the structural properties of the community itself, weighted or otherwise, are connected to such information to some degree.

Comparison between two clustering algorithms and seven journals showed that 87.62% of the correlation coefficients found from 15 features and 4 dependent variables show weak correlations exceeding |0.3|, while 55.71% showed moderate correlations with values larger than |0.5|. A comparison between the correlation coefficient of dependent variables revealed *NeighborRatio* had a significantly lower value of 0.2497 compared to the coefficient of *NewTopicCount, NewTopicFreq,* and *NeighborCount* respectively with 0.4658, 0.4715, and 0.4832. The absolute number of neighbors of future new topics has a more significant relationship with the graph structures compared to their ratio to the subgraph itself. This indicates that the correlation outcomes are less affected by the community size, which frequently dictates many of the structural properties.

Linear, quadratic, and cubic regressions are done for each of the four dependent variables over 7 journals. R^2 scores of the regression results over different variable combinations are averaged to first determine the best regression degree. R^2 values in TABLE V. show that the value experiences sharp drops with quadratic and cubic regressions, indicating that the best fit for the regression lines is linear for all variables but *NeighborRatio* which showed $R^2 < 0.05$ even with linear regression. The R^2 value for three variables reached near 0.8, explaining more than 80% of the data variabilities. A comparison between the two algorithms shows that the weighted algorithm resulted in strictly better results in the three variables, validating the assumption that topic co-occurrence frequencies would positively be related to the prediction performance.

TABLE V. R^2 VALUES AVERAGED BY REGRESSION DEGREES.

Communities from *Greedy* Algorithm			
Dependent Variable	***Linear***	***Quadratic***	***Cubic***
NewTopicCount	0.7902	0.0828	0.0056
NewTopicFreq	0.8139	0.1234	0.0071
NeighborCount	0.8572	0.1000	0.0090
NeighborRatio	0.0252	0.0022	0.0020
Communities from *Infomap* Algorithm			
Dependent Variable	***Linear***	***Quadratic***	***Cubic***
NewTopicCount	0.8103	0.2419	0.0021
NewTopicFreq	0.8366	0.2993	0.0032
NeighborCount	0.8913	0.3007	0.0052
NeighborRatio	0.0320	0.0024	0.0000

R^2 scores plateaued when the number of training years *t* become 5 for three dependent variables except for *NeighborRatio*, which was removed from the analysis for its performances. Comparisons between different *t* values were unnecessary as the variance between different *t* falls under $5.61e^{-5}$ to $1.43e^{-5}$ when the *t=1* was excluded; the middle-value *t=5* was selected for further.

Linear regression results on communities found with *Infomap* algorithms for seven journals over the 20-year timeslots with *t=5* is shown in TABLE VI. where values over 0.9 are marked with **bold** and values below 0.8 are marked with *italic*. *Nature, Science, NEJM, Cell,* and *HEP* show higher values compared to *KBS* and *JoI*. The distinctively lower score for *JoI* can again be attributed to the journal being new, not having a stabilized pattern to be analyzed with only 13 years of history. The effect of the journal immaturity can also be seen with *KBS* with the second lowest value, as it had 33 years of history in the MAG dataset. *HEP* dataset had only 23 years of publications recorded but showing high R^2 scores, but the effect could be seen with $R^2 < 0.5$ when the *greedy* algorithm was used. Comparing three dependent variables showed that *NeighborCount* returns the highest score followed by *NewTopicFreq* in 6 out of 7 journals. This validates the assumption that identification and prediction of new topics can be done through their neighbors, and the frequencies of topic co-occurrences lead to better predictions.

TABLE VI. R^2 SCORE OF LINEAR REGRESSION RESULTS FOR INFOMAP COMMUNITIES FROM DIFFERENT JOURNALS TRAINED BY 5 PREVIOUS YEARS.

Journals	**NewTopicCount**	**NewTopicFreq**	**NeighborCount**
Nature	0.8499	0.8764	**0.9576**
Science	**0.9044**	**0.9153**	**0.9742**
NEJM	**0.9415**	**0.9438**	**0.9613**
Cell	**0.9159**	**0.9394**	**0.9767**
HEP	**0.9667**	**0.9579**	**0.9834**
KBS	*0.7280*	*0.7638*	*0.7878*
JoI	*0.4922*	*0.5024*	*0.6900*

NRMSE with different normalization methods in TABLE VII. show that the mean square errors for three variables are not normally distributed using linear regression on infomap communities with $t=5$. $NRMSE_{minmax}$ showing values more than 20 times larger than $NRMSE_{mean}$ suggests that there are a small number of large errors present. The largest value in $NRMSE_{interquartile}$ indicates that such large error values are sparse while there are numerous low-error values in the 1st quartile, meaning that the NRMSE follows power laws for each of the three variables. The least amount of errors are observed in *NeighborCount* indicating that it is the most accurate predictor. The *NewTopicFreq* variable shows the largest NRMSE value in contrary to its high R^2 scores, which can be attributed to the nature of the variable; the variable counts the number of topic co-occurrences, which can reach up to thousands depending on the number of publications made each year unlike the other two. The large variance of the possible values results in a relatively lower $NRMSE_{minmax}$ as well.

TABLE VII. NRMSE SCORE FOR THREE DEPENDENT VARIABLES WITH DIFFERENT NORMALIZATIONS.

NRMSE	NewTopicCount	NewTopicFreq	NeighborCount
$NRMSE_{minmax}$	0.0654	0.0447	0.0435
$NRMSE_{mean}$	1.3909	1.8034	0.8763
$NRMSE_{interquartile}$	2.3124	8.1458	2.3644

Out of three normalization methods, $NRMSE_{minmax}$ is selected to compare the NRMSE score of three dependent variables for different journals in Fig. 3. Five journals which showed high R^2 scores also showed low NRMSE values for all of three dependent variables, having less than 3% of the maximum observed value of errors on average. *KBS* showed slightly higher normalized errors around 5%, while communities in *JoI* resulted in significantly higher NRMSE value around 20% and more. The values suggest that the regression analysis can be done to predict how new topics will be related to the given community in various topic networks, consistently resulting in up to 0.95 R^2 score and normalized RMSE ranging around 3%. The topic networks however need to be mature enough to have stabilized topic co-occurrence patterns to achieve such high performance.

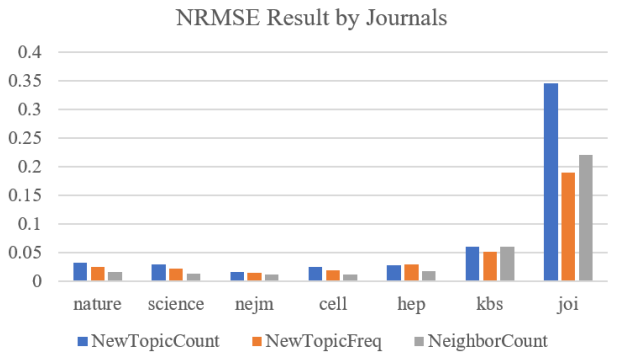

Fig. 3. $NRMSE_{minmax}$ Score of Linear Regression results for Infomap Communities from Different Journals Trained with 5 Previous Years.

VI. CONCLUSION

Topic models derived from processing unstructured documents can capture the number of topics shared throughout given document collection and can be used to detect and track changes in such topics over time. The text-based approaches however have an innate limitation of requiring the textual data for modeling topics, inhibiting the effective prediction of topic evolutions where such data is nonexistent. The paper proposed an alternative to such an approach by utilizing the network structure to model topics, where topics can be defined by its relative neighbors alone. Retrospective classification is first done to validate the assumption that the new topics can be distinguished by the structural properties of their neighborhoods in the past. Prospective prediction is done next to apply the findings in the classification for the prediction of new topics.

Binary classification result showed that the detection of new topics can be done solely based on the structural features of their neighbors with high classification accuracy up to 0.9, and linear regression on communities showed that the prospective prediction of new topics and their to-be-neighbors can be done on communities found from existing community detection algorithms with R^2 score value up to 0.97. The analysis of the retrospective classification and prospective prediction showed that the performance is closely related to the maturity of the topic networks as well as the range of topics; the dataset with more stabilized topic co-occurrence patterns results in better prediction performances. The case of *Phys.Rev* and *JoI* showed that the small, or new, topic networks result in poor performances as the network-based topic model requires certain sizes and histories to work properly.

Future works incorporate experiments on topic networks based on research domains rather than individual journals for better coverage of the focused research interest. An algorithm will then be proposed to predict individual new topics based on network-based topic models. Instead of *NeighborCount*, the authors plan to analyze each member of communities to detect individual community members being to-be neighbors of new topics in the future. This would reveal a more focused context of the new topics derivable from the given communities, which in turn would enable distinguish individual new topics connected to the communities. The network-based topic model calls for the community detection algorithm conscious of the properties used for the new topic prediction, which will be tackled based on the existing algorithms. A new graph generation algorithm would be proposed afterward with the gained knowledge on topic emergence as well as other evolutionary events, simulating the topic evolution in the given domain.

REFERENCES

[1] 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, Aug. 2020, doi: 10.1016/j.joi.2020.101040.

[2] 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, Nov. 2017, doi: 10.1016/j.joi.2017.10.003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] Z. Guo, Z. M. Zhang, S. Zhu, Y. Chi, and Y. Gong, "A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 780–794, Apr. 2014, doi: 10.1109/TKDE.2013.56.

[5] L. Kay, N. Newman, J. Youtie, A. L. Porter, and I. Rafols, "Patent overlay mapping: Visualizing technological distance," *J Assn Inf Sci Tec*, vol. 65, no. 12, pp. 2432–2443, Dec. 2014, doi: 10.1002/asi.23146.

[6] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of latent semantic analysis*, Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2007, pp. 427–448.

[7] 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*, 0 vols., Society for Industrial and Applied Mathematics, 2009, pp. 859–870.

[8] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, pp. 113–120, doi: 10.1145/1143844.1143859.

[9] 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*, New York, NY, USA, 2005, pp. 198–207, doi: 10.1145/1081870.1081895.

[10] Y. Jo, J. E. Hopcroft, and C. Lagoze, "The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus," in *Proceedings of the 20th International Conference on World Wide Web*, New York, NY, USA, 2011, pp. 257–266, doi: 10.1145/1963405.1963444.

[11] 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*, New York, NY, USA, 2009, pp. 957–966, doi: 10.1145/1645953.1646076.

[12] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised Prediction of Citation Influences," in *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA, 2007, pp. 233–240, doi: 10.1145/1273496.1273526.

[13] 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)*, Dec. 2017, pp. 1694–1697, doi: 10.1109/BigData.2017.8258108.

[14] J. G. Fiscus and G. R. Doddington, "Topic Detection and Tracking Evaluation Overview," in *Topic Detection and Tracking*, Springer, Boston, MA, 2002, pp. 17–31.

[15] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic Detection and Tracking Pilot Study Final Report," Jan. 1998, doi: 10.1184/R1/6626252.v1.

[16] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," in *Proceedings of the 17th International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Dec. 2004, pp. 1617–1624.

[17] 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, doi: 10.1002/asi.20317.

[18] 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, Dec. 2017, doi: 10.1177/0165551516661914.

[19] 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, Sep. 2005, doi: 10.1016/j.ipm.2004.04.018.

[20] B. C. Gencosman, H. C. Ozmutlu, and S. Ozmutlu, "Character n-gram application for automatic new topic identification," *Information Processing & Management*, vol. 50, no. 6, pp. 821–856, Nov. 2014, doi: 10.1016/j.ipm.2014.06.005.

[21] S. Ozmutlu, "Automatic new topic identification using multiple linear regression," *Information Processing & Management*, vol. 42, no. 4, pp. 934–950, Jul. 2006, doi: 10.1016/j.ipm.2005.10.002.

[22] A. L. Porter and M. J. Detampel, "Technology opportunities analysis," *Technological Forecasting and Social Change*, vol. 49, no. 3, pp. 237–255, Jul. 1995, doi: 10.1016/0040-1625(95)00022-3.

[23] C. Battistella, "The organisation of Corporate Foresight: A multiple case study in the telecommunication industry," *Technological Forecasting and Social Change*, vol. 87, pp. 60–79, Sep. 2014, doi: 10.1016/j.techfore.2013.10.022.

[24] N. C. Newman, A. L. Porter, D. Newman, C. C. Trumbach, and S. D. Bolan, "Comparing methods to extract technical content for technological intelligence," *Journal of Engineering and Technology Management*, vol. 32, pp. 97–109, Apr. 2014, doi: 10.1016/j.jengtecman.2013.09.001.

[25] A. Bongers and J. L. Torres, "Measuring technological trends: A comparison between U.S. and U.S.S.R./Russian jet fighter aircraft," *Technological Forecasting and Social Change*, vol. 87, pp. 125–134, Sep. 2014, doi: 10.1016/j.techfore.2013.12.007.

[26] A. Segev, C. Jung, and S. Jung, "Analysis of Technology Trends Based on Big Data," in *2013 IEEE International Congress on Big Data (BigData Congress)*, Jun. 2013, pp. 419–420, doi: 10.1109/BigData.Congress.2013.65.

[27] A. Segev, S. Jung, and S. Choi, "Analysis of Technology Trends Based on Diverse Data Sources," *IEEE Transactions on Services Computing*, vol. 2015 Vol.8, no. 06, pp. 903–915, Dec. 2015, doi: 10.1109/TSC.2014.2338855.

[28] S. Jung and A. Segev, "Analyzing future communities in growing citation networks," in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2013) International Workshop on Mining Unstructured Big Data using Natural Language Processing*, New York, NY, USA, Oct. 2013, pp. 15–22, doi: 10.1145/2513549.2513553.

[29] S. Jung and A. Segev, "Analyzing future communities in growing citation networks," *Knowledge-Based Systems*, vol. 69, pp. 34–44, Oct. 2014, doi: 10.1016/j.knosys.2014.04.036.

[30] 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)*, Jun. 2016, pp. 357–360, doi: 10.1109/BigDataCongress.2016.57.

[31] J. Kim, J.-G. Lee, and S. Lim, "Differential Flattening: A Novel Framework for Community Detection in Multi-Layer Graphs," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, p. 27:1–27:23, Oct. 2016, doi: 10.1145/2898362.

[32] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.

[33] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *PNAS*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008, doi: 10.1073/pnas.0706851105.

[34] K. Wang *et al.*, "A Review of Microsoft Academic Services for Science of Science Studies," *Front. Big Data*, vol. 2, 2019, doi: 10.3389/fdata.2019.00045.

[35] A. Sinha *et al.*, "An Overview of Microsoft Academic Service (MAS) and Applications," in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, May 2015, pp. 243–246, doi: 10.1145/2740908.2742839.

[36] S. E. Hug, M. Ochsner, and M. P. Brändle, "Citation Analysis with Microsoft Academic," *Scientometrics*, vol. 111, no. 1, pp. 371–378, Apr. 2017, doi: 10.1007/s11192-017-2247-8.

[37] Z. Shen, H. Ma, and K. Wang, "A Web-scale system for scientific knowledge exploration," *arXiv:1805.12216 [cs]*, May 2018, Available: http://arxiv.org/abs/1805.12216.