IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 2, 2020, accepted May 29, 2020, date of publication June 8, 2020, date of current version June 22, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3000948*

# TermBall: Tracking and Predicting Evolution Types of Research Topics by Using Knowledge Structures in Scholarly Big Data

**CHRISTINE BALILI[1], UICHIN LEE[1], (Member, IEEE), AVIV SEGEV[2], (Member, IEEE), JAEJEUNG KIM[1], AND MINSAM KO[3]**

[1]Graduate School of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
[2]Department of Computer Science, University of South Alabama, Mobile, AL 36688, USA
[3]Department of Human-Computer Interaction, Hanyang University, Ansan-si 15588, South Korea

Corresponding author: Minsam Ko (minsam@hanyang.ac.kr)

This work was supported in part by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea under Grant NRF-2019S1A5A2A03052192, in part by the Technology Advancement Research Program funded by Ministry of Land, Infrastructure and Transport of Korean Government under Grant 19CTAP-C152247-01, and in part by the KAIST-KU Joint Research Center, KAIST, South Korea.

**ABSTRACT** The exponential growth in the number of publications and the prevalence of interdisciplinary research in recent years call for new approaches for analyzing how topics in science are evolving at large. This paper proposes TermBall, a framework that tracks and predicts fine-grained topic evolution in terms of the evolution types: emergence, growth, shrinkage, survival, merging, splitting, and dissolution. TermBall builds the knowledge structure, which is a weighted dynamic network of co-occurring keywords in the literature, and then discovers key topic structures that consist of keywords and their relationships by performing community detection methods. Based on the topic structures, TermBall provides two applications: (1) Retrospective application to identify topic evolution in the past and (2) Predictive application to forecast upcoming topic evolution type based on the structural and temporal features of the topic structures. For the evaluation, we built the knowledge structure by applying TermBall to 19 million articles in PubMed that were published from 1980 to 2014. We conducted qualitative analysis on the derived topic evolution types and quantitative analysis on the prediction results. As a result, our qualitative analysis reveals that TermBall is able to find various topic evolution types from the knowledge structure and also can predict how topics will evolve after five years with an accuracy of 83%.

**INDEX TERMS** Knowledge structure, dynamic networks, topic evolution, evolution types.

## I. INTRODUCTION

Topics in science continually evolve. In response to the influx of novel discoveries and changing societal needs, new topics can surface and become part of the state-of-the-art, or existing ones can disappear after losing relevance. Topics can also persist for extended periods of time. Depending on the level of research interest they harbor, they can undergo periods of growth, shrinkage, or stagnation. When researchers from different domains collaborate, knowledge, and skills from both areas can become increasingly intertwined in the long term, thus leading to the convergence of the respective topics. It is also possible for a single coherent topic to fork into multiple separate subtopics as it enters a state of more specialization or when its research directions become polarized.

Given that research articles and related indexing data are exponentially growing, analyzing such scholarly big data brings novel opportunities for enabling data-driven analytics about knowledge on how the research topics evolve over time [1]. This data-driven approach can direct strategies at both individual and organizational levels [2], [3]. Keeping up with research trends allows scientists to recognize gaps in knowledge that they can address and emerging topics that they can pursue in future work. Meanwhile, funding agencies can leverage data-driven foresight on topic evolution to determine which areas present the most potential for

The associate editor coordinating the review of this manuscript and approving it for publication was Xiangtao Li.

growth. Data-driven insights can then be employed to allocate research grants strategically.

This work proposes *TermBall*, a novel framework for tracking and predicting fine-grained topic evolution based on topic keyword networks in terms of evolution types such as emergence, growth, shrinkage, survival, merging, splitting, and dissolution. TermBall utilizes the knowledge structure of a research domain as a dynamic network. The nodes in the network comprise of keywords that describe the key concepts covered in published research outputs. The undirected edges correspond to co-occurrences of keyword pairs in the literature, and the weight of an edge is equal to the frequency with which the corresponding keyword pair co-occurred within a given period. Thus, the linkages of these keywords depict the evolving associations of ideas and therefore expose shifting research foci and priorities over time.

Prior work examined various topic discovery [4]–[6] and topic evolution methods [7], [8]. However, most topic discovery approaches extract representative words based on statistical models, which make it hard to understand the relationship between topics and to track how topics evolve over time. Our framework provides ambient and useful information. In particular, our network-based representation (i.e., co-occurrence networks) sheds light on the strengths and patterns of conceptual relationships in a domain. This representation is informative because we can illustrate *conceptual relationships* beyond semantic and syntactic similarities. This can offer a bird's eye view of the collective knowledge in the scientific discipline from which the documents are generated.

Based on the knowledge structure, TermBall provides (1) the retrospective application to understand the past and (2) the predictive application to forecast upcoming evolution types. As the retrospective application, TermBall first extracts topic structures from the knowledge structure by adopting a community detection algorithm that outputs groups of nodes that are densely linked to each other but sparsely connected to the rest of the network [9]. Next, it identifies topic evolution types by analyzing changes in the topic structure over time. Furthermore, as the predictive application, TermBall provides predictions on how a given topic will undergo any events including growth, survival, shrinking, merging, splitting, and dissolution in the next snapshot, while the majority of the previous approaches to topic evolution have been retrospective only [10]–[14]. The proposed framework harnesses a feature set consisting of the structural and temporal attributes of persistent topics and trains a classifier to forecast their future states. This approach is inspired by previous studies of communities in social networks where attributes such as size and density have been shown to correlate with longevity and specific evolutionary patterns such as growth and survival.

We demonstrate the use of this proposed framework by examining topic evolution in the context of biomedical literature. Our dataset is composed of 19 million articles from PubMed, which is currently the largest repository of biomedical articles from MEDLINE,[1] life science journals, and online books. PubMed articles come with annotations from the MeSH[2] ontology, which is a controlled vocabulary of medical terms. We constructed a dynamic network based on the co-occurrences of these MeSH terms from 1980-2014 (35 years). We took snapshots of the network at five-year time intervals to encapsulate the states of topics at different periods in the recent history of biomedicine. Finally, our proposed framework achieved an accuracy of 83% in predicting which of the six possible evolutionary events a topic will experience within five years.

The contributions of this work can be summarized as follows. First, the proposed framework can be used for topic discovery through the detection of keyword communities in a dynamic network. Second, our approach can track and predict topic evolution corresponding to changes in the community structure of the keyword co-occurrence network. Although we only present the application of this framework to one large corpus, it can easily be applied to other collections of scientific texts that are indexed by a controlled vocabulary of keywords.

This work significantly extends our earlier work [15] as follows. First, we provided a comprehensive literature review on topic detection/evolution and community detection/evolution. Second, we significantly extended our framework by considering both disjoint and overlapping community detection algorithms. Furthermore, we incorporated both structural and temporal features in our prediction models. Third, we provided a comprehensive set of results on the development of the knowledge structure and predictive analyses of research topics in PubMed. Our model evaluation showed that the accuracy of our new model has significantly improved. The code and data for this study are available at this link.[3]

## II. BACKGROUND AND RELATED WORK

### A. TOPIC DISCOVERY

Topic discovery or modeling is the unsupervised task of identifying a set of latent topics pervading in a given collection of documents [16]. Majority of the work on this problem currently revolves around statistical topic models, which describe a probabilistic process by which documents are generated [17]. In the context of such models, each document is composed as follows. First, the proportion of topics is selected. Second, each word position in the document is assigned a topic. Finally, a word is picked from the vocabulary based on the topic that was chosen for its position. The topic proportion per document represents a multinomial distribution, and the topic itself also represents a multinomial distribution over the vocabulary words [5], [16].

Mathematically, the generative process is represented as a joint distribution of hidden random variables and

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

[2] https://www.nlm.nih.gov/pubs/factsheets/mesh.html

[3] https://github.com/christinebalili/ms_thesis_src.git

the observed word and document co-occurrence statistics. The latent variables in this case are the topics and per-document topic distribution. Probabilistic latent semantic analysis (PLSA) is a technique that is formulated in this manner, which employs an expectation-maximization algorithm to maximize the probability of the hidden variables given the observed ones. This value is also known as the posterior distribution of the model [18], [19]. Latent Dirichlet Allocation (LDA) extends PLSA by adding Dirichlet priors to its multinomial distributions [4], [7]. This algorithm infers topics more effectively, but the addition of hidden variables results in an intractable posterior distribution. Thus, the hidden variables are only approximated using sampling and variational techniques [20].

In both PLSA and LDA, the number of topics is set apriori. However, in most cases this information is not previously known, even to domain experts. Hierarchical Dirichlet Process (HDP) models overcome this limitation by automatically determining the number of topics based on the given dataset [6]. Following non-parametric Bayesian principles, the number of hidden parameters in HDP increase based on the data it encounters [21]. Similarly to LDA, these parameters are also estimated using Gibbs sampling and variational inference [22]. The Correlated Topic Model (CTM) represents another extension of LDA. However, instead of using a Dirichlet distribution to model the variability in per-document topic proportions, it employs a logistic normal distribution. This setting allows topics to have a covariance structure, and therefore enables the discovery of correlated topics [23].

The present work offers an alternative perspective on topic modelling by leveraging the knowledge structure of a domain. This differs from the previous models by representing topics as a community of terms from a keyword co-occurrence network. We explicitly use the linkages between words instead of deploying a bag-of-words treatment to infer the underlying topics in a corpus. Thus, the conceptual relationships beyond semantic and syntactic similarity can be accounted for in the analysis. Furthermore, this approach allows us to have a bird's eye view of the collective knowledge in the scientific discipline.

### B. TOPIC EVOLUTION

In the context of scientific literature, the methods for topic discovery and evolution mining employ additional metadata associated with the document collection. This additional information includes *citations* [24], [25], *authorship* [26]–[28], and *index terms* [29], [30]. These methods allow for persistent themes to be contextualized within the knowledge and social structures that are specific to a domain [2], [3].

A citation network is a graph that connects papers based on their references. Citation networks can be directed, where an edge is drawn from the citing paper to the cited paper. They can also be undirected, such as in the case of co-citation networks, where an edge is set between two papers if they are both cited together in a publication. Link-LDA is a model that combines citation structures and textual information to discover latent topics [31]. In this framework, a citation relationship between a pair of documents is modeled as a Bernoulli random variable, which is parameterized by the topic proportions of the document pair. The topic mixture in a citing document is dependent on that of a cited document. Meanwhile, the citation-aware framework proposed by He *et al.* [32] employs a linear combination for the topic model, derived from both the cited and citing documents. In the same work, an alternative model was also proposed in which the topics in the citing set are dependent on the cited set.

There are also citation analyses that focus on network structures. In these models, the network is divided into clusters of papers that densely cite each other. The topic in a cluster is determined based on the word usage patterns within its corresponding papers [33]. Therefore, topic dynamics can be illustrated as a function of the structural attributes over time. For instance, Shibata *et al.* discovered incremental and branching topic evolution patterns in gallium nitride and complex networks research by measuring the within-cluster degree and participation coefficient for each paper in their respective citation network [34]. Intuitively, a timeline reflecting the emergence and continuity of topics can be generated by linking similar clusters in successive periods [25], [35].

Collaborations among researchers can also reveal the topic dynamics within a field. Co-authorship networks are undirected graphs that connect two researchers when they have written a paper together. Each node in such a network is associated with a set of documents that have been authored by a researcher. Under this setting, topic discovery can be performed by first identifying the clusters or communities based on the network structure and then performing statistical topic modeling on the documents in each cluster [36]. Another approach reverses this process by starting with the inference of topic proportions in each individual node and then constructing a topical community by grouping nodes based on topic mixture similarities [37]. In some cases, such as in topic-link LDA [38], the interplay between textual information and social contexts is modeled simultaneously using probabilistic methods that impose a conditional dependency between topics and an author's community membership. In addition, Kalyanam *et al.* explored this joint approach by introducing a common variable for both topic and community distributions per time step in a model based on non-negative matrix factorization [39].

Topic evolution is tightly coupled to the changing social interactions among researchers. The "burstiness" [40] of a topic within a certain research community can encourage authors from other communities to move towards it. This process of taking in new members tends to cause shifts in topic interests within the current community. This phenomenon is referred to as information diffusion [41]. The development of research foci in a community can thus result from people exploring new topics and topics being brought in by new people [27], [42].

The ebb and flow of research themes is perhaps most evident in the examination of the keywords that are used to index publications in digital libraries [43], [44]. In a co-word analysis [45], a keyword (or term) co-occurrence network is derived from a specialized corpus to depict the knowledge structure in a domain [30], [46], [47]. This network representation approach aims to capture the relationships between technical concepts in a research area at an aggregate level. In this context, a cohesive topic is represented by a cluster or community of words that are densely connected to one another. Over time, the manner in which concepts are linked can vary as new ideas are introduced and outdated ones disappear. Topological changes in the network can therefore signal topic evolution. As in citation networks [25], [35], topic persistence, splitting, merging, and dissolution can be monitored by thresholding the similarity in terms of membership between clusters in successive time slices [47], [48].

Each type of network constructed from each type of scholarly metadata (e.g. authorship and citations) is structured differently. Therefore, each network can offer different perspectives on topic evolution. The similarities of various scholarly networks have previously been investigated by Yan and Ding [49]. Their findings revealed a high degree of dissimilarity as measured by the cosine distance between co-authorship, citation/co-citation, and keyword networks. Co-authorship networks tend to reflect social events among researchers more than topic evolution, as they are constructed based on the relationships between people in the field. On the other hand, although citation networks depict information flow, the links between their components often lack context. In particular, the reason for the citation of a paper by another is not apparent [50]. Thus, interpreting the meanings of their underlying communities and the topics that they represent requires linking them to other available metadata such as keywords and co-authorship. In terms of structure, they are also sparser and are likely to exhibit high temporal variation [51].

Keyword co-occurrence networks are not riddled by the stated limitations of their counterparts. This type of network maps knowledge in a domain intuitively by linking ideas that are related to each other in published research works [52]. The context or topics of a community can be directly inferred by looking at the respective components. The framework proposed in this paper operates similarly to those in prior studies in terms of discovering topics and tracking their temporal evolution in scholarly publications. However, we extend our analysis to include a predictive component. Our work contributes to this area by identifying a set of features that can be used to foresee the future states of topics found in a dynamic keyword co-occurrence network.

### C. COMMUNITY DETECTION

Community detection involves identifying communities, i.e., partitions of nodes that are densely connected to one another compared to the rest of the network [53]. This process provides insights into a network's latent organization and the dynamics of certain processes that take place within it. There are a plethora of community detection algorithms, and they can be broadly categorized based on whether they identify *disjoint communities* or *overlapping communities* [54]–[57]. Each node in a graph can only represent part of one community in the former, while multiple memberships are permitted in the latter.

There are two main methods for disjoint community detection. One is the divisive algorithm, which is a top-down technique because at the start it considers the entire network as a single cluster and iteratively splits it by eliminating links joining nodes with low similarity and ends up with unique communities. It removes inter-cluster edges in a network based on low-similarity to separate communities from each other. Girvan-Newman [53] algorithm makes use of this method. Another method is the agglomerative algorithm, which is a bottom-up technique because at the start it considers each node as a separate cluster and iteratively merge them based on high similarity and ends up with the unique community. Newman [58] and Clauset [59] proposed a greedy algorithm that makes use of the modularity measure to define communities that have many edges within them and few between them. Such modularity has a high time complexity. Moreover, it tends to generate a large scale of super-communities.

In order to overcome these drawbacks, Blondal [60] proposed Louvain algorithm. It has a two-phase iterative process of also a heuristic greedy method for modularity optimization, and community aggregation. The initial phase, it starts with considering each node in the network is a community. Then it iteratively merges the nodes based on the gain of modularity until no gain is achieved. When no more improvement is possible, the second phase reconstructs the network in the way that communities identified in the first phase are replaced by supernodes. The time complexity of Louvain algorithm is $O(nlogn)$, while Newman's greedy algorithm is $O(n)^3$. Alternatively, spectral clustering techniques operate by vectorizing network nodes and then clustering them in a high dimensional space to reveal the community structure [61]. Walktrap approaches this task by using short random walks that end up traversing the same set of nodes, which could possibly constitute a community [62].

On the other hand, there have been many studies about overlapping community detection. This approach is motivated by the fact that entities represented by nodes can belong to multiple groups [57]. In the case of a social network, a person can be part of any number of communities dedicated to their work, a hobby, or a specific interest. One of the most well-known algorithms for this purpose is the clique percolation method, which relies on adjacent cliques to build communities from the bottom-up [63]. Probabilistic methods have also been proposed to model network organizations using a mixture of distributions and thus return fuzzy node memberships [64]. The other notable techniques involve partitioning edges instead of nodes [65], [66].

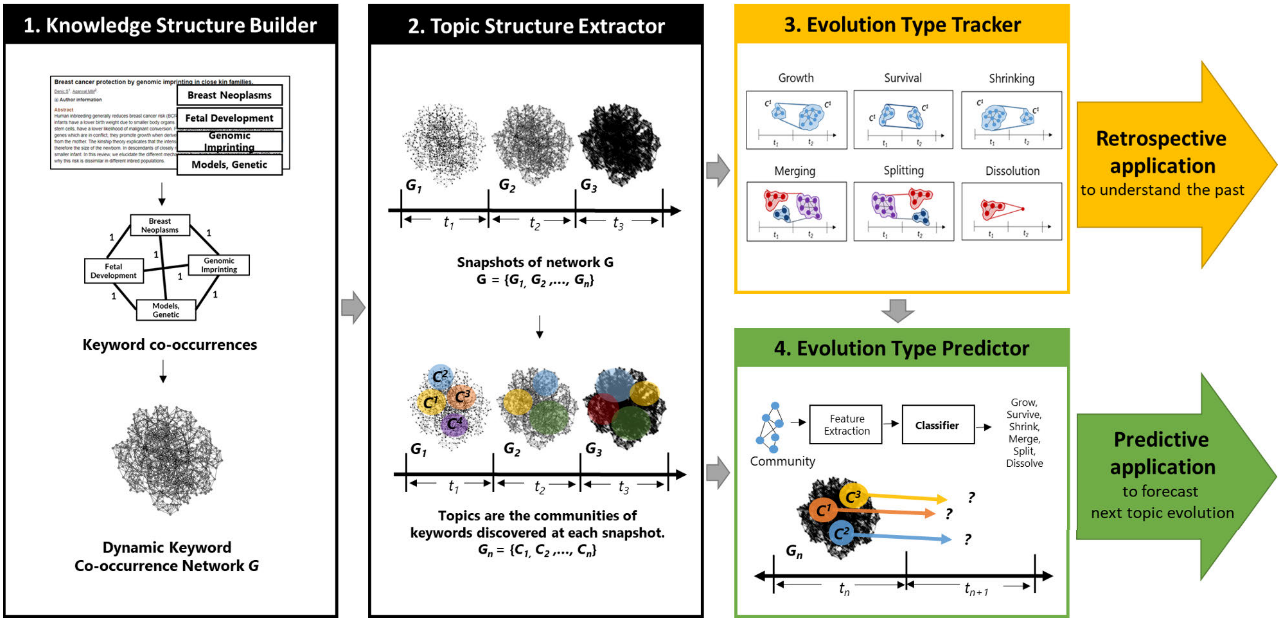

1. Knowledge Structure Builder
Breast cancer protection by genomic imprinting in close kin families.
Author information
Abstract
Breast Neoplasms
Fetal Development
Genomic Imprinting
Models, Genetic
Breast Neoplasms
Fetal Development
Genomic Imprinting
Models, Genetic
1
Keyword co-occurrences
Dynamic Keyword Co-occurrence Network G
2. Topic Structure Extractor
$G_1$ $G_2$ $G_3$
$t_1$ $t_2$ $t_3$
Snapshots of network G
G = {$G_1$, $G_2$ ,…, $G_n$}
$C^1$ $C^2$ $C^3$ $C^4$
Topics are the communities of keywords discovered at each snapshot.
$G_n$ = {$C_1$, $C_2$ ,…, $C_n$}
3. Evolution Type Tracker
Growth
Survival
Shrinking
Merging
Splitting
Dissolution
$C^1$
$t_1$ $t_2$
Retrospective application
to understand the past
4. Evolution Type Predictor
Community
Feature Extraction
Classifier
Grow, Survive, Shrink, Merge, Split, Dissolve
$C^1$ $C^2$ $C^3$
?
$G_n$
$t_n$ $t_{n+1}$
Predictive application
to forecast next topic evolution

**FIGURE 1.** **Overview of TermBall.**

## III. TermBall: A FRAMEWORK FOR TRACKING AND PREDICTING TOPIC EVOLUTION TYPES

The components of our proposed framework for tracking and predicting topic evolution types are illustrated in Figure 1. TermBall consists of four components: (1) Knowledge Structure Builder, (2) Topic Structure Extractor, (3) Evolution Type Tracker, and (4) Evolution Type Predictor.

Given a collection of scientific publications, the Knowledge Structure Builder in TermBall begins with the construction of a dynamic co-occurrence network using the keywords that appear in the documents and takes snapshots of this network along a specified time range. Then, TermBall extracts key sub-structures from the co-occurrence network by performing a community detection algorithm at each snapshot, in order to discover topics in the literature. The detected sub-structures are equivalent to the topics that pervade in the research publications for that time period. Next, to track the evolution types of topics, it matches the communities across snapshots and detects the changes in their compositions and sizes. Finally, a machine learning classifier, trained based on a set of structural and temporal features from previous topic structures, predicts whether a given topic will undergo any growth, survival, shrinking, merging, splitting, and dissolution events in the next snapshot.

### A. KNOWLEDGE STRUCTURE BUILDER

The collective knowledge of a research field is represented as a dynamic keyword co-occurrence network derived from a large collection of time-stamped documents. The dynamic network $G$ consists of a series of snapshots $\{G_1, G_2, \ldots, G_n\}$ taken at equally-spaced time intervals within the specified time span of the analysis. Each snapshot $G_t = (V_t, E_t)$ is an undirected weighted graph that contains the set of keywords $V_t$ and the set of co-occurrences $E_t$. The weight of an edge is the number of times a pair of keywords appear together in an article within the duration of the snapshot.

In this work, we considered two methods by which a snapshot can be obtained. The first approach involves a sliding window that is translated across the time intervals. Snapshots constructed in this manner only account for the co-occurrences within the most recent time interval. Using the sliding window, nodes and edges can disappear between snapshots. The second approach is an aggregate window. This method takes snapshots by incrementally collapsing all observed keyword co-occurrences from the beginning until the present time interval. The differences between these snapshot formulations are illustrated in Figure 3. Examining a dynamic network under these two viewpoints enables short-term and long-term regularities in the knowledge structure to be accounted for in the process of discovering topics.

### B. TOPIC STRUCTURE EXTRACTOR

In our framework, topic structure extraction is equivalent to community detection at each snapshot of the dynamic keyword co-occurrence network. This process yields a set of $n_t$ communities $\{C_t^1, C_t^2, \ldots, C_t^{n_t}\}$ for every snapshot $G_t$. A community $C_t^j$ is a subgraph $(V_t^j, E_t^j)$ that represents a research topic that exists within the time window of the snapshot. The communities in this representation are not explicitly defined, and so we employed detection algorithms that infer community structures in an unsupervised manner. Many community detection algorithms have been proposed, and they can be broadly categorized based on whether it identifies disjoint communities or overlapping communities

[54]–[57]. Each node in a graph can only represent part of one community in the former, while multiple memberships are permitted in the latter.

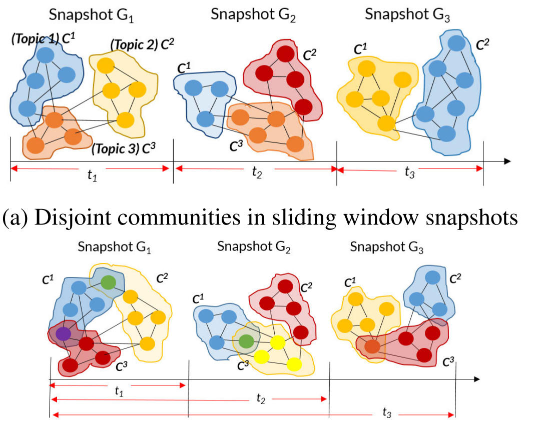
In this work, we considered and compared two representative community detection algorithms that they were selected from each of the approaches. First, for the disjoint community detection, we adopt Infomap [67] that identifies clusters by compressing random walks that encode information flow in the network. This choice was motivated by the size of the network at hand and the comparative performance of Infomap versus other state-of-the-art techniques in previous studies involving synthetic datasets [9], [54]. Second, we deploy the Order Statistics Local Optimization Method (OSLOM) for overlapping community detection. OSLOM optimizes a local objective function which represents the statistical significance of connections in a discovered cluster compared to a random subgraph [68]. This particular decision to use OSLOM is owing to its capability to detect clusters in weighted and dynamic graphs [68].

### 1) INFOMAP FOR DISJOINT COMMUNITY DETECTION

Infomap is a method based on information theory, wherein communities are discovered by learning an efficient encoding of a random walk in a network [67]. A random walk encapsulates how information flows in the graph. Ideally, an encoding that represents the network structure should be able to compress its most interesting topological features in the minimum number of bits possible. When a network has a modular organization, a random walk is statistically likely to traverse the same set of nodes in its densely-linked parts. These traversal patterns can be used to determine a minimum length encoding by only assigning unique codes to communities and allowing for node-specific codes to be re-used across communities. The algorithm specifies an exit code that marks when the walk moves from one community to another. The process of finding optimal community assignments is equivalent to minimizing the description length of a random walk.

### 2) OSLOM FOR OVERLAPPING COMMUNITY DETECTION

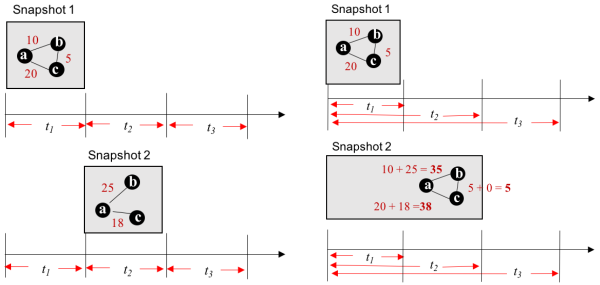
OSLOM allows detecting overlapped communities from the weighted networks [68]. OSLOM is a local optimization method applied to the statistical significance of individual sub-structures (i.e., communities) that can be measured by the probability of finding a similar structure (i.e., same size, degree sequence and internal connections) in a null model possessing no community structure. OSLOM begins to group neighbor nodes to obtain a collection of significant, possibly overlapping communities. Then, it tries to remove or add nodes to communities in order to increase their significance. In other words, an external node that has at least one edge connected to a community is added to it if the connection is determined to be statistically significant relative to the previous one. This process is repeated several times and stops when no new communities are found, and its stability is ensured by the repetition due to the stochastic nature.



(a) Disjoint communities in sliding window snapshots

(b) Overlapping communities in aggregate window snapshots

**FIGURE 2. Methods for topic structure extraction.**



(a) Sliding window

(b) Aggregate window

**FIGURE 3. Taking snapshots of the dynamic network.**

In the initial run of OSLOM, we employed the disjoint communities returned by Infomap in the first network snapshot $G_1$ as the seed subgraphs. For the succeeding runs, the communities determined by OSLOM in the previous snapshot $G_{t-1}$ were used as starting points for the communities to be identified in the current snapshot at $G_t$. To some extent, this setup imposes temporal smoothing in the process of community detection across snapshots. We employed the default settings in the R implementation of Infomap in the igraph library[4] and the code provided by the authors of OSLOM[5] in our experiments. In addition, we allowed for nodes to be singletons i.e. not part of any community in OSLOM.

Disjoint communities represent topics that are more disparate, whereas overlapping communities embody the existence of concepts or ideas that are shared by multiple topics in a research field. Infomap is employed to find topics in snapshots from the sliding window approach, while OSLOM is utilized for topic discovery in snapshots derived from the aggregate window scheme (as shown in Figure 2). Disjoint community detection is run on top of sliding window snapshots, as this approach allows the extent of a single cohesive topic to be measured for a particular point in time. In this

[4] http://igraph.org/r/

[5] http://www.oslom.org/index.html

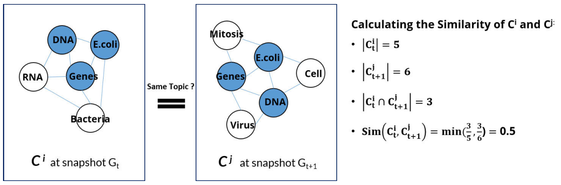

DNA
E.coli
RNA
Genes
Bacteria
$C^i$ at snapshot $G_t$
Same Topic ?
Mitosis
E.coli
Genes
Cell
DNA
Virus
$C^j$ at snapshot $G_{t+1}$
Calculating the Similarity of $C^i$ and $C^j$:
• $|C_t^i| = 5$
• $|C_{t+1}^j| = 6$
• $|C_t^i \cap C_{t+1}^j| = 3$
• $\mathrm{Sim}(C_t^i, C_{t+1}^j) = \min(\frac{3}{5}, \frac{3}{6}) = 0.5$

**FIGURE 4.** Topic continuity calculation.

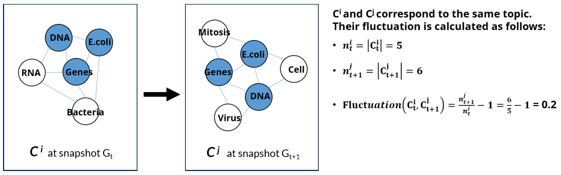

DNA
E.coli
RNA
Genes
Bacteria
$C^i$ at snapshot $G_t$
Mitosis
E.coli
Genes
Cell
DNA
Virus
$C^j$ at snapshot $G_{t+1}$
$C^i$ and $C^j$ correspond to the same topic.
Their fluctuation is calculated as follows:
• $n_t^i = |C_t^i| = 5$
• $n_{t+1}^j = |C_{t+1}^j| = 6$
• $Fluctuation(C_t^i, C_{t+1}^j) = \frac{n_{t+1}^j}{n_t^i} - 1 = \frac{6}{5} - 1 = 0.2$

**FIGURE 5.** Fluctuation calculation.

manner, we can rank topics in the domain according to the sizes of their communities. On the other hand, overlapping community detection is performed on aggregate window snapshots because this enables the identification of topics that exhibit increased concept sharing over time.

## C. EVOLUTION TYPE TRACKER

### 1) TOPIC CONTINUITY

We matched the identified communities across consecutive snapshots based on their similarities to track the continuity of a research topic. The similarity between a pair of communities is defined by Hopcroft *et al.* [69] as:

$$Sim(C_t^j, C_{t+1}^k) = min\left(\frac{|C_t^j \cap C_{t+1}^k|}{|C_t^j|}, \frac{|C_t^j \cap C_{t+1}^k|}{|C_{t+1}^k|}\right) \geq \theta \tag{1}$$

where $C_t^j$ is a community in snapshot $G_t$ and $C_{t+1}^k$ is a community in the subsequent snapshot $G_{t+1}$. These two communities represent the same topic if their similarity is greater than or equal to a certain threshold $\theta$.

In essence, the similarity function measures the amount of keywords that are common to the two communities from consecutive snapshots. An example is presented in Figure 4. If we set $\theta = 0.5$, then the two communities in the example are matched and correspond to the same topic because $Sim(C_t^i, C_{t+1}^j) \geq 0.5$. The threshold $\theta$ is a hyperparameter in the framework. If $\theta$ is set to a high value, then the number of persistent topics that can be discovered will be lower, as communities are more likely to dissolve. On the other hand, assigning a lower value would increase the number of persistent topics, with fewer similar keyword compositions over time. One possible approach to tuning $\theta$ is to consider the rate of change in the vocabulary of keywords that are used in a collection of scientific articles. In the case of PubMed literature, we set $\theta = 0.25$, as such a value enabled us to obtain samples for each type of topic evolution.

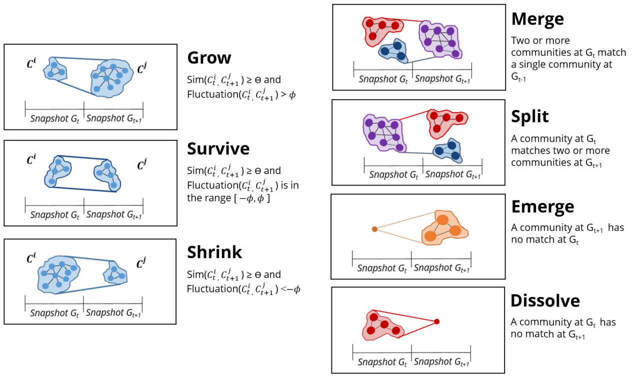

$C^i$
$C^j$
Snapshot $G_t$
Snapshot $G_{t+1}$
Grow
$\mathrm{Sim}(C_t^i, C_{t+1}^j) \geq \Theta$ and
$\mathrm{Fluctuation}(C_t^i, C_{t+1}^j) > \phi$
Survive
$\mathrm{Sim}(C_t^i, C_{t+1}^j) \geq \Theta$ and
$\mathrm{Fluctuation}(C_t^i, C_{t+1}^j)$ is in the range $[-\phi, \phi]$
Shrink
$\mathrm{Sim}(C_t^i, C_{t+1}^j) \geq \Theta$ and
$\mathrm{Fluctuation}(C_t^i, C_{t+1}^j) < -\phi$
Merge
Two or more communities at $G_t$ match a single community at $G_{t-1}$
Split
A community at $G_t$ matches two or more communities at $G_{t+1}$
Emerge
A community at $G_{t+1}$ has no match at $G_t$
Dissolve
A community at $G_t$ has no match at $G_{t+1}$

**FIGURE 6.** Topic evolution types.

### 2) TOPIC EVOLUTION TYPES

If a community has a match in the subsequent snapshot, then we identify it as a persistent topic. Otherwise, we say that its topic has dissolved. If a community at snapshot $G_{t+1}$ has no match in the preceding snapshot $G_t$, then we consider it as a new community, and more specifically an emergent topic. If a community at $G_t$ has two or more matches at $G_{t+1}$, then the topic that it represents has split. Meanwhile, if two communities at $G_t$ match the same community at $G_{t+1}$, then their corresponding topics have merged.

Persistent communities can be categorized further based on their fluctuation, which is defined by Ilhan and Oguducu [70] as:

$$Fluctuation(C_t^j, C_{t+1}^k) = \frac{n_{t+1}^j}{n_t^k} - 1 \tag{2}$$

where $C_t^j$ and $C_{t+1}^k$ are matching communities and $n_t^k$ and $n_{t+1}^j$ are the number of nodes (or keywords) in each respective community. In essence, the fluctuation measures the change in size of a topic. We present an example in Figure 5. If we set $\phi = 0.10$, then the topic represented by the two communities shown in the example has grown, because $Fluctuation(C_t^i, C_{t+1}^j) = 0.20 > 0.10$. Based on the threshold $\phi$, a persistent topic can either grow, survive, or shrink over time. We set $\phi = 0.10$, in order to consider a 10% increase in community size as substantial growth in our analysis of PubMed. The list of possible topic evolution is illustrated and defined in Figure 6.

## D. EVOLUTION TYPE PREDICTOR

The task of predicting the evolution of a research topic can be considered as a supervised classification problem. Communities at a snapshot $G_t$ are labeled based on the events that they undergo in the next snapshot $G_{t+1}$. Thus, topic evolution prediction becomes a six-label classification task. Our prediction model covers all previously defined events except for emergence. We first define a set of 42 handcrafted features to represent previously discovered communities. This feature set is then used to construct our prediction model.

**TABLE 1. Structural features of a topic.**

| Category | ID | Name | Description |
|---|---|---|---|
| Subgraph | 1 | Size | Number of nodes in the community |
| | 2 | Cohesion | Ratio of the number of internal edges to the number of external edges |
| | 3 | Average Clustering Coefficient | Average clustering coefficient of nodes based on subgraph of the community |
| | 4 | Density | Ratio of the number of actual edges observed compared to the number of possible edges |
| | 5 | Triangles | The number of triangles in the subgraph of the community |
| Nodes | 6-11 | PageRank | Mean, standard deviation, min, and the first three quartiles (25th, 50th, and 75th percentiles) of the global PageRank of nodes within the community |
| | 12-17 | Degree Centrality | Mean, standard deviation, min, and the first three quartiles (25th, 50th, and 75th percentiles) of the global degree centrality of nodes within the community |
| | 18-23 | Clustering Coefficient | The mean, standard deviation, min, and the first three quartiles (25th, 50th, and 75th percentiles) of the global clustering coefficient of nodes within the community |
| Edges | 24 | Internal Edges | Number of edges with both nodes inside the community |
| | 25 | Internal Non-edges | Number of non-existing edges between the nodes inside the community |
| | 26 | External Edges | Number of edges that are connected to a node outside the community |
| Weights | 27 | Total Internal Weight | Sum of the weights of internal edges |
| | 28-31 | Distribution of Internal Weights | The mean, standard deviation, and the first three quartiles (25th, 50th, and 75th percentiles) of the weights of internal edges |
| | 32 | Total External Weight | Sum of the weights of external edges |
| | 33 | Internal Weight Ratio | Ratio of the total internal weight to the total weight of all the edges with at least one node in the community |
| | 34 | External Weight Ratio | Ratio of the total external weight to the total weight of all the edges with at least one node in the community |
| | 35 | Overall Weight Ratio | Ratio of the total internal weight and the total external weight of the community |

**TABLE 2. Temporal features of a topic.**

| Category | ID | Name | Description |
|---|---|---|---|
| Subgraph | 36 | Size Change | Percentage of change in the number of nodes in the community |
| | 37 | Change in Density | Percentage of change in the density of the community from snapshot $G_t$ to $G_{t+1}$ |
| | 38 | Change in Average Clustering Coefficient | Percentage of change in the average clustering coefficient of the community from snapshot $G_t$ to $G_{t+1}$ |
| | 39 | Change in Triangles | Percentage of change in the number of triangles within the community |
| Edges | 40 | Change in Edges | Percentage of change in the number of edges in the community |
| Weights | 41 | Change in Internal Weights | Percentage of change in the total weight of internal edges |
| | 42 | Change in External Weights | Percentage of change in the total weight of external edges |

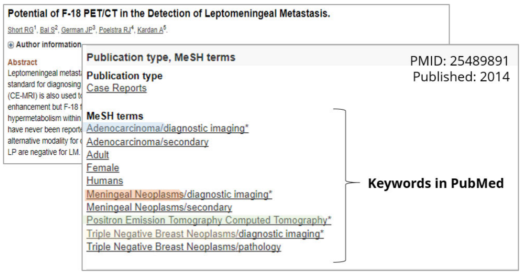
Structural and temporal properties were extracted from the communities to forecast their future states. The structural features of a topic, which are listed in Table 1, are static attributes that can be derived from its community representation within a single snapshot. These features aim to characterize the strengths of connections among community components in the current time snap. Hence, they are grouped according to the network components on which they are based: subgraphs, nodes, edges, and weights. Meanwhile, Table 2 lists the temporal features, which reflect changes in the properties of a community, using its former instance in a preceding snapshot as a point of reference. Similarly to the structural properties, we grouped these features according to the structural components from which they are taken.

Based on the derived features, we employed a supervised classification model to predict what type of evolution will occur for a topic in the next snapshot of the keyword co-occurrence network. Depending on the nature of the corpus, the representation of each possible event can be highly imbalanced. Hence, our framework employs the Synthetic Minority Oversampling TEchnique (SMOTE), which undersamples the majority class and oversamples the minority class to arrive at a relatively balanced dataset prior to model training [71]. To predict topic evolution, we trained a set of classifiers including support vector machines (SVM), k-nearest neighbor (KNN), logistic regression, AdaBoost, and random forest.

## IV. EVALUATION

In this section, we discuss the insights provided by our proposed framework in the context of PubMed literature. In the first part, we provide a global view of the overall development of knowledge in the biomedical domain as evident in PubMed. In the second part, we delve into a retrospective analysis of topics. We identify the major topics that have been prevalent in the history of PubMed. We also illustrate the different types of topic evolution undergone by some of the discovered persistent topics. Finally, we present a predictive analysis of topic evolution based on the model that resulted from our proposed framework. We utilize this model to predict how the current major topics will change after five years.

For the purpose of demonstrating how this framework performs the tasks of finding topics and analyzing their evolution, we apply it to a large corpus of scientific articles in the biomedical domain. The analysis presented in this paper is based on 18.54 million papers that were published from



Potential of F-18 PET/CT in the Detection of Leptomeningeal Metastasis.

Short RG1, Bal S2, German JP3, Poelstra RJ4, Kardan A5.

⊕ Author information

Abstract

Leptomeningeal metast
standard for diagnosing
(CE-MRI) is also used t
enhancement but F-18 f
hypermetabolism within
have never been report
alternative modality for
LP are negative for LM.

Publication type, MeSH terms

PMID: 25489891
Published: 2014

Publication type

Case Reports

MeSH terms

Adenocarcinoma/diagnostic imaging*
Adenocarcinoma/secondary
Adult
Female
Humans
Meningeal Neoplasms/diagnostic imaging*
Meningeal Neoplasms/secondary
Positron Emission Tomography Computed Tomography*
Triple Negative Breast Neoplasms/diagnostic imaging*
Triple Negative Breast Neoplasms/pathology

**FIGURE 7.** **A sample article in PubMed and its MeSH terms.**

1980 to 2014 and are currently contained in PubMed.[6] Each article in PubMed is annotated using a specialized set of keywords from a biomedical ontology called Medical Subject Headings (MeSH).[7] A sample article and its corresponding MeSH terms are shown in Figure 7. The highlighted terms are the major subjects or nodes in the keyword co-occurrence network. This corpus is distributed using XML files, where articles and their metadata are accessible in a structured format. We only employed the MeSH terms that were tagged as the primary subjects of a publication for constructing the keyword co-occurrence network. Also, we limited the edges in our network representation to only be between MeSH terms that co-occurred more than five times within the interval of the snapshot. These filtering processes ensure that we constrain our analysis to the most relevant conceptual relationships in the domain. Finally, community detection at each snapshot of the dynamic co-occurrence network enabled us to identify the dominant research topics in the PubMed corpus at different points in history. The dominant topics found in the disjoint and overlapping setup show a high degree of similarity even when the duration of the time windows in their snapshots are different.

### A. TOPIC EVOLUTION TYPES IN PubMed

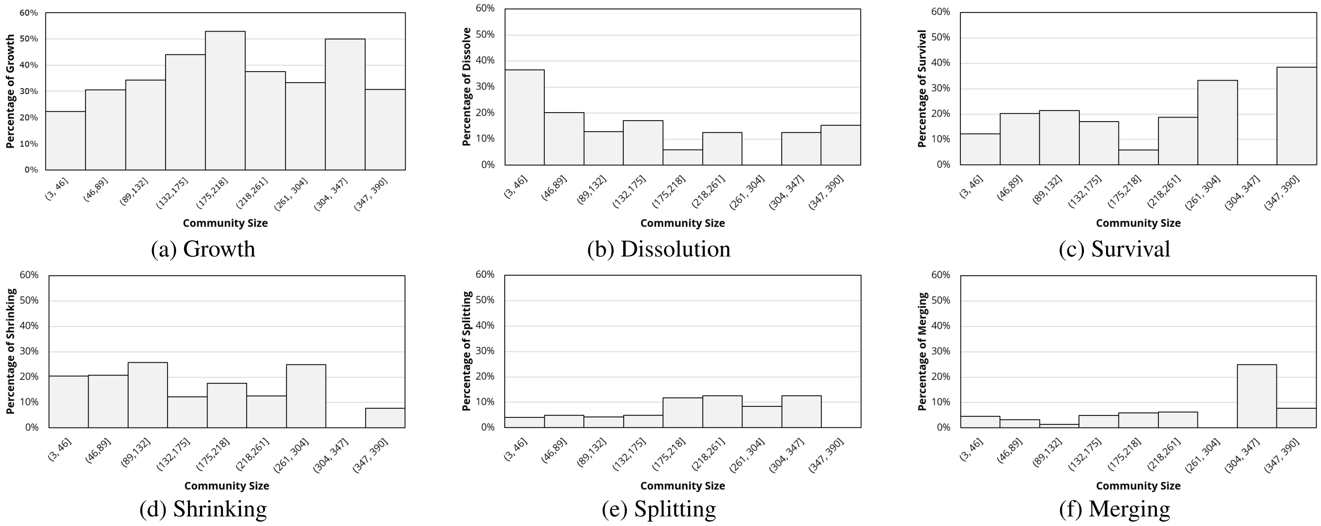
TermBall identified existing topic evolution types in PubMeb by tracking each topic structures. The distribution of evolution types across community size groups is displayed in Figure 8. In Figure 8, we can see that the likelihood of growth increases from smaller to larger communities, while the likelihood of dissolution decreases. This pattern suggests that the evolving knowledge structure in PubMed also exhibits the preferential attachment (''rich-get-richer'') phenomenon [72], [73]. The merging and splitting of topics are less frequent events, because persistent communities tend to remain cohesive over their lifetime. Our detailed qualitative analysis results on the topic evolution types in PubMeb are followed.

#### 1) GROWTH, SURVIVAL, AND SHRINKING

The continuity of a research topic can be traced by matching communities found between successive timesteps. We were able to identify 41 major topics that persisted throughout the duration of our analysis (1980-2014). We consider a community to be a major topic if it has been composed of at least 100 MeSH terms at some point in the snapshots. Of these major topics, 80% exhibit continual growth, but at varying rates over time. We could observe that the keyword communities of *Cancer Research*, *Mental Disorders*, and *Phytotherapy* have considerably expanded over the years. The other persistent communities, such as *Neuroscience* and *Anti-bacterial Agents*, appear to maintain a relatively steady growth compared to the others. Community growth can be interpreted as the elaboration and specialization of a research area. In terms of publications, this implies that new associations are formed and new ideas are being embraced in the field.

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

[7] https://www.nlm.nih.gov/pubs/factsheets/mesh.html



(a) Growth

(b) Dissolution

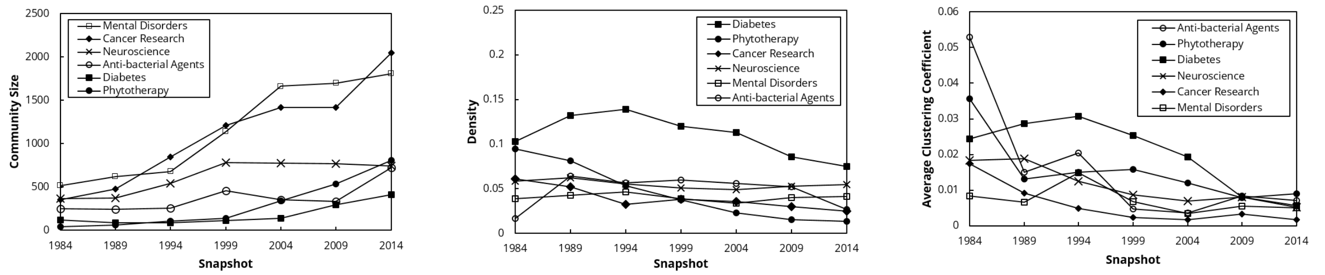
(c) Survival

(d) Shrinking

(e) Splitting

(f) Merging

**FIGURE 8.** **Histogram of evolution events across community sizes (equal bin size of 43 was selected for exact mapping).**



(a) Community size (b) Density (c) Average coefficient

**FIGURE 9. Network statistics of selected major topics.**

In Figures 9a, we present a temporal examination of the size of the subgraphs corresponding to six of the largest persistent communities in the keyword co-occurrence network. These topics dominate the research landscape in the period 2010-2014. The subgraph densities of these communities are also shown in Figure 9b. As opposed to the behavior of the network at large, the subgraphs of *Diabetes*, *Phytotherapy*, and *Cancer Research* actually exhibit a decreasing density over time. On the other hand, the *Neuroscience* and *Mental Disorders* communities appear to maintain a relatively stable density. The *Anti-bacterial Agents* community noticeably plateaued before decreasing in the final snapshot.

In addition, we present the average clustering coefficients of the aforementioned communities in Figure 9c. Concerning their respective clustering coefficients, the subgraphs of interest exhibit a decreasing trend. We can interpret these structural changes in light of the sustained growth of these communities. As more MeSH terms are acquired by these communities, their number of possible internal connections increases exponentially, thus resulting in the decline of the aforementioned structural properties. We can further relate these observations to that of Palla *et al.* [10], wherein large groups in time-evolving networks were found to prevail longer if they are capable of dynamically altering their membership. The dynamic reconfiguration of community members enables a cluster to handle the influx of new nodes. If this property holds, then large communities become more adaptable and resilient to fluctuations in the network. In the case of PubMed's evolving knowledge structure, a decrease in density and the clustering coefficient allows these major topics to accommodate more keywords and continue to expand in the future.

There are also persistent topics that exhibit continuous shrinkage over time. The topic of *Occupational Diseases* exemplifies this case. This keyword community began with 118 nodes in the first snapshot of the network (1980-1984), but has since reduced to include only 73 MeSH terms in the most recent snapshot (2010-2014). This phenomenon can signify a decrease of interest in this topic, and hence the pool of keywords has shrunk over time. This finding could also be explained by the increased co-occurrence of keywords originally belonging to this topic with keywords from another topic, resulting in a change in membership from the former to the latter.

**TABLE 3. Splitting of *Brain-related Diseases*.**

| 2005-2009 | 2010-2014 |
|---|---|
| Brain Neoplasms<br>Glioma<br>Neurosurgical Procedures<br>Glioblastoma<br>Meningeal Neoplasms<br>Meningioma<br>Radiosurgery<br>Antineoplastic Agents<br>Alkylating<br>Hydrocephalus<br>Craniotomy | Brain Neoplasms<br>Glioma<br>Neurosurgical Procedures<br>Glioblastoma<br>Radiosurgery<br>Meningioma<br>Meningeal Neoplasms<br>Antineoplastic Agents, Alkylating<br>Craniotomy<br>Skull Base Neoplasms |
| | Hydrocephalus<br>Neuroendoscopy<br>Cerebrospinal Fluid<br>Cerebral Ventricles<br>Intracranial Hypertension<br>Ventriculostomy<br>Intracranial Pressure<br>Creation of ventriculo-peritoneal shunt<br>Third ventricle structure<br>Cerebrospinal fluid shunts procedure |

### 2) MERGING AND SPLITTING

With the prevalence of interdisciplinary practice in science, it would be interesting to ask how collaborations among different fields of expertise have shaped the temporal knowledge landscape within PubMed through the merging and splitting of keyword communities. In Table 3 and Table 4, we present notable examples of these events in our analysis. The first case illustrates the splitting of the cluster of *Brain-related Diseases*. This topic existed as one cohesive community since the first snapshot but has recently divided into two communities corresponding to *Brain Neoplasms* and *Hydrocephalus* in 2010-2014. The splitting of a community can represent the decomposition of a topic into more specialized subareas or the divergence of research questions towards different trajectories. The other instances of splitting that we detected include the forking of the *Chromosomes* topic into the more specific foci of *Biological and Molecular Evolution* and *Chromosome Aberrations, Translocation, and Deletions* in 1995-1999. The two segments that originated from this split have been consistently detected as separate communities until the present.

The palpable effect of interdisciplinary research on the overall knowledge structure is perhaps most evident in topic merging events. The increasing number of publications

**TABLE 4.** Merging of *Coronary Heart Disease* and *Arterial Diseases*.

| 1995-1999 | 2000-2004 |
|---|---|
| Heart<br>Electrocardiography<br>Coronary heart disease<br>Myocardial Infarction<br>Hemodynamics<br>Heart failure<br>Cardiac Surgery procedures<br>Heart Diseases<br>Thoracic Surgery Specialty<br>Heart rate | Disease<br>Postoperative Complications<br>Heart<br>Myocardial Infarction<br>Myocardium<br>Coronary heart disease<br>Cardiovascular system<br>Heart failure<br>Cardiovascular Diseases<br>Electrocardiography |
| Arteriosclerosis<br>Thrombosis<br>Angiogram<br>Vascular Diseases<br>Aorta<br>Leg<br>Vascular Surgical Procedures<br>Ischemia<br>Arteries<br>Stent, device | |

**TABLE 5.** Development of *Medical Image Processing*.

| Years | Topics |
|---|---|
| 2000 -2004 | X-Ray Computed Tomography, Biological Models, algorithm, Image Processing, Gene Expression Profiling, Imaging, Three-Dimensional, Computer software, Models, Statistical, Computer Simulation, Oligonucleotide Array, Sequence Analysis |
| 2005 -2009 | Magnetic Resonance Imaging, algorithm, Biological Models, X-Ray Computed Tomography, Image Interpretation, Image Enhancement, Imaging, Three-Dimensional, Theoretical Model, **Pattern Recognition**, Computer software |
| 2010 -2014 | Magnetic Resonance Imaging, X-Ray Computed Tomography, algorithm, Positron-Emission Tomography, Image Interpretation, Imaging, Three-Dimensional, Image Processing, **Pattern Recognition**, Image Enhancement, Contrast Media |

resulting from collaborative projects induces the flow of ideas from one discipline to another. In our model, this knowledge transfer can be observed as an increase in edges between communities as their keywords co-occur more often in the literature. The merging of *Coronary Heart Disease* with *Arterial Diseases* is an example of the convergence of closely related communities. These topics began as independent modules in our analysis but have been combined in a singular cluster since 2004.

### 3) EMERGING RESEARCH TOPICS

In the context of our proposed framework, we define emerging research topics as communities that were not present in past instances of the network but have gone on to become major topics from their initial point of discovery until the present time. This definition naturally excludes communities that were found in the first network snapshot. Some of the fairly recent emerging topics we have discovered in our analysis are presented in Table 5 and Table 6. The formation of the *Medical Image Processing* community can be associated with the continued integration of computer vision research into medical applications. This process can also be attributed to the increase in computing power over recent years, allowing for more sophisticated and resource-intensive algorithms to be employed for such purposes. Meanwhile, we can explain the emergence of *Disease Outbreaks* as a possible response of the biomedical community to the severe acute respiratory syndrome (SARS) outbreak that became a major global health threat in 2002-2004 [74].

In addition to detecting emerging themes in the domain at large, we also found that shifts in interests within a persistent topic can be observed by examining changes in the centralities of its member nodes. For instance, in Table 7, the interest in *Drug Resistance* in research on *Anti-bacterial Agents* became apparent when this node garnered a higher PageRank within the community after 1995. On the other hand, the rise of *Cardiovascular Diseases* in the ranks of keywords within the topic of *Diabetes* (Table 8) can be credited to the growing scientific output investigating the

**TABLE 6.** Development of *Disease Outbreaks*.

| Years | Topics |
|---|---|
| 2005 -2009 | Disease Outbreaks, Influenza, Antibodies, Viral, Vaccination, Viral Vaccines, Population Surveillance, Viral Envelope Proteins, Influenza virus vaccine, Enzyme-Linked Immunosorbent Assay, Vaccines, DNA |
| 2009 -2014 | Influenza, Disease Outbreaks. **Influenza A Virus**, **H1N1 Subtype**, Vaccination, Antibodies, Viral, Influenza virus vaccine, Swine Diseases, Viral Vaccines, Immunologic Adjuvants, Influenza A virus, Orthomyxoviridae Infections |

**TABLE 7.** Development of *Anti-Bacterial Agents*.

| Years | Topics |
|---|---|
| 1980 -1984 | Anti-Bacterial Agents, Bacterial Infections, Bacteria, Neonatal disorder, Cephalosporins, Disease Outbreaks, Staphylococcal Infections, Cross Infection, Urinary Tract Infection, Sepsis |
| 1995 -1999 | Anti-Bacterial Agents, Staphylococcal Infections, Penicillins, Bacteria, Antibiotics, Antitubercular, Anti-Infective Agents, Genus Staphylococcus, Disease Outbreaks, **Drug Resistance**, **Microbial**, Streptomycin |
| 2010 -2014 | Anti-Bacterial Agents, Bacterial Proteins, Escherichia coli, Escherichia coli Proteins, Staphylococcal Infections, Cross Infection, Gene Expression Regulation, Bacterial Microbial Biofilms, **Drug Resistance, Bacterial** Anti-Infective Agents |

**TABLE 8.** Development of *Diabetes*.

| Years | Topics |
|---|---|
| 2000 -2004 | Insulin, Obesity, Glucose, Diabetes Mellitus, Diabetes Mellitus, Non-Insulin-Dependent, Hypoglycemic Agents, Diabetes Mellitus, Insulin-Dependent, Blood Glucose, Adipose tissue, Islets of Langerhans |
| 2005 -2009 | Obesity, Diabetes Mellitus, Non-Insulin-Dependent, **Cardiovascular Diseases**, Insulin, Diet, Glucose, Hypoglycemic Agents, Diabetes Mellitus, Insulin-Dependent, Feeding Behaviors |
| 2010 -2014 | Obesity, Diabetes Mellitus, Non-Insulin-Dependent **Cardiovascular Diseases**, Diet, Hypoglycemic Agents, Insulin, Feeding Behaviors, Dietary Supplements, Glucose, Diabetes Mellitus, Insulin-Dependent |

co-morbidities of these lifestyle diseases. For the recently formed *Medical Image Processing* community in Table 5, the increase in the centrality of *Pattern Recognition* reflects how machine learning has established itself at the core of biomedical applications in recent years. Finally, the remarkable surge of *Influenza A Virus, H1N1 Subtype* within the *Disease Outbreaks* community in 2010-2014 (Table 6) can be considered as the aftermath of the AH1N1 outbreak in 2009 [75]. These examples illustrate how our framework could be employed to screen for "hot" keywords within more general topics discovered in a large corpus.

### B. PREDICTION PERFORMANCE ON TOPIC EVOLUTION TYPES

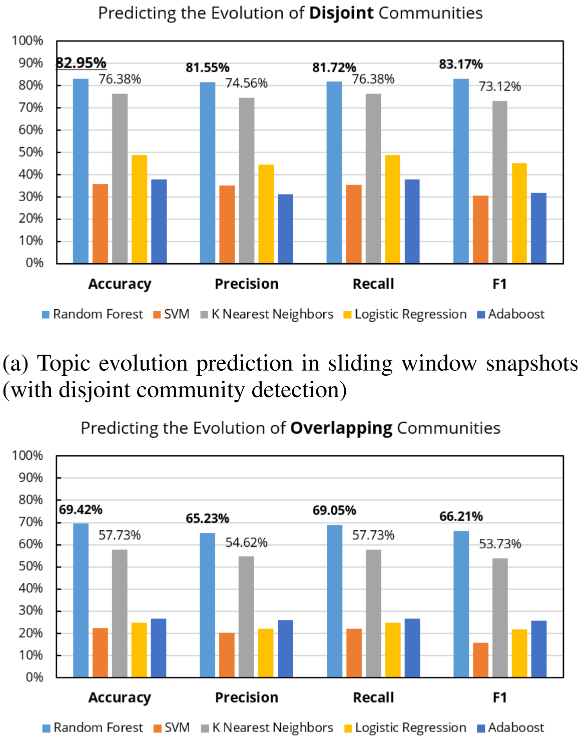
We trained several classifiers including SVM, KNN, logistic regression, AdaBoost, and random forest to predict how topics will evolve after five years in both the sliding and aggregate window setups.

As illustrated in Figure 2, we used two modes of topic discovery for labeling: (1) disjoint community detection (Infomap) with slide windowing, and (2) overlapping community detection (OSLOM) with aggregated windowing. There were *323 disjoint communities* discovered by Infomap in the sliding window snapshots, while there were *1,271 overlapping communities* found by OSLOM in the aggregate window snapshots. These previously discovered topics, whose evolutions are already known, were employed in training and testing the prediction model.

In Figure 10, we present the cross-validated evaluation for the accuracy, precision, recall, and F1-score in the two cases. In both setups, the random forest model consistently outperformed all the other classifiers. The best performance was achieved when the topic evolution prediction was modeled using disjoint communities in sliding window snapshots. The overall accuracy in this setting was at 82.95% for predicting which of the six possible events will occur for a topic within five years.

We present the normalized confusion matrix for the random forest model in Figure 11. We can infer that most of the mistakes incurred by the model are concentrated on the growth and dissolution events. On the other hand, the model is able to predict merge and split events with considerably high accuracies. This difference in performance can be attributed to the limited examples available for these two event types. It is also worth noting that even though our model can predict topic merging, it is unable to determine which two communities will specifically converge together.

Furthermore, we assessed the importance of the handcrafted structural and temporal features in community evolution predictions from the resulting random forest classifier. As shown in Table 9, the distribution of node clustering coefficients, cohesion, and density change rank as the highest among all the derived features. We note that the majority of the most informative features are structural.



(a) Topic evolution prediction in sliding window snapshots (with disjoint community detection)

(b) Topic evolution prediction in aggregate window snapshots (with overlapping community detection)

**FIGURE 10. Topic evolution prediction results.**

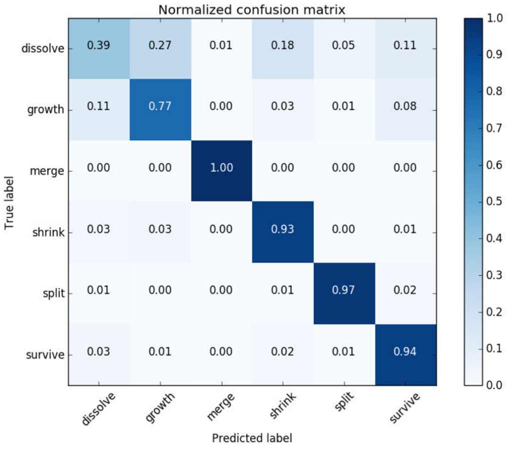


**FIGURE 11. Confusion matrix for topic evolution prediction.**

Finally, we apply our trained model to predict the evolutions of topics discovered in the most recent snapshot of the PubMed knowledge structure. Among the largest keyword communities, continued growth is foreseen for *Mental Disorders*, *Anti-Bacterial Agents*, and *Diabetes* by 2019. Meanwhile, our model predicts the future splitting of *Artificial Nanoparticles*, *Neuroscience*, and *Environmental Monitoring*. Furthermore, *Cancer Research* and *Phytotherapy* are expected to dissolve. However, this prediction does not necessarily mean that these topics will disappear entirely. Due to the relative sizes of these communities compared to the

**TABLE 9.** Ranking of features.

| Rank | Feature Name | Feature Type |
|---|---|---|
| 1 | 50th Percentile of Clustering Coefficient | Structural |
| 2 | Standard Deviation of Clustering Coefficient | Structural |
| 3 | 25th Percentile of Clustering Coefficient | Structural |
| 4 | Cohesion | Structural |
| 5 | Density Change | Temporal |
| 6 | Subgraph Clustering Coefficient | Structural |
| 7 | Mean of Clustering Coefficient | Structural |
| 8 | Maximum of Clustering Coefficient | Structural |
| 9 | 75th Percentile of Clustering Coefficient | Structural |
| 10 | 25th Percentile of Internal Weights | Structural |

others, it is possible that they will only divide into smaller and more specialized communities in the coming years.

## V. DISCUSSION

Tracking the overall behavior of the dynamic network in terms of its global structural properties enabled us to quantitatively describe the evolution of knowledge at large over extended periods of time. These insights serve as a ''bird's eye view'' of the domain which can provide researchers and other decision-makers with the ability to understand the general trajectory of a highly diverse and dynamic research area. The results of this analysis extend the view of stakeholders from their respective specialized research areas to the field in its entirety. Furthermore, because we have defined a set of measures that can describe the state of knowledge at a given point in time, it is possible to extrapolate these values in the future through a time series analysis. Therefore, we can forecast how a field will evolve in an aggregate manner based on the previous network size, density, and diameter.

The retrospective analysis of topic evolution in our framework is based on the changes in size and composition of keyword communities. This allows stakeholders to take a high-level view of the organization of the domain knowledge at a given time and the influences of historical events on the structure of the dynamic network. Our experiments on biomedical literature demonstrated that the emergence of topics, as evidenced by the formation of new keyword communities, can be explained by notable events such as global outbreaks of diseases and the application of new computational techniques to biomedical challenges. In our analysis, we identified topics based on the nodes in their communities that bear the highest PageRank. These nodes are considered as the core concepts representing a topic. Our investigation revealed that shifts in research interests or foci within topics become apparent when these core concepts are examined over time. As with the formation of new communities, the increase of keywords in the centrality of a topic is also impacted by events and trends in research practice.

Aside from a historical account of topic evolution, our framework offers a predictive analysis of research trends in science. We cast topic evolution prediction as a supervised classification problem. Our experiments have demonstrated that we can predict the future state of prevalent research topics with high accuracy. We were also able to assess which of the handcrafted structural and temporal features were informative in the prediction task. The foresight provided by our proposed framework can be employed by researchers and funding agencies to determine which particular topics are predisposed to growth and which are becoming less popular, so that they can allocate their resources accordingly. In effect, they can sift through the massive volume of literature and determine the most interesting topics.

One possible limitation that arises from our network-based representation is the reliance of our technique on the existence of a controlled vocabulary of keywords such as the MeSH ontology in PubMed. In the absence of such a specialized ontology, natural language processing techniques to extract entities of interest from the raw text are warranted to obtain appropriate nodes that will comprise the dynamic network. However, because most scholarly work is now distributed online, the process of recommending keywords from domain-specific vocabularies for indexing articles is increasingly becoming the norm. Hence, the proposed framework can be still applied in other research areas.

In addition, our approach has a limitation that cannot capture new emerging topic, if all the related terms to the topic are not included MeSH terms. However, we think that the effects of this limitation are not serious in practical situations. First, MeSH terms well represent a study because the researchers manually select the terms as the most representative ones of their study, and MeSH terms are often used as ground-truth data in topic discovery research. Second, the limitation mostly occurs in the first stage. At first, some keywords in the new emerging topic would not be included in MeSH terms. However, if the topic keeps growing, and many researchers begin to cover them, the keywords will be likely included in MeSH terms, and our framework can work with the topic. Furthermore, if the smaller window size is used, the limitation of our framework can be mitigated by tracking topic evolution at a micro-level.

As future work, we aim to explore the effect of coupling the dynamic keyword co-occurrence network with citation and co-authorship networks in the task of tracking and predicting topic evolution. These networks, constructed from different types of scholarly metadata, introduce a variety of perspectives on the interaction of ideas and the flow of information in science. Hence, they may be able to complement each other and together can capture the evolution of a research field in a more holistic and comprehensive manner. We also aim to employ the framework in the construction of a knowledge exploration platform that can visualize knowledge structures and enable users to zoom in and out to their areas of interest. Finally, we intend to leverage this paradigm in the task of recommending hypotheses to researchers.

## VI. CONCLUSIONS

We have proposed TermBall, a holistic framework that models knowledge structure of research topics and tracks/predicts the evolution of research topics. TermBall represents research topics as communities of keywords in a dynamic co-occurrence network. Based on a large number of scientific publications, our framework has provided a retrospective and predictive analysis of how these topics will develop over time. The progress of research in an area is reflected by its persistence and growth as a community of keywords. Furthermore, we have demonstrated that the evolution of the research landscape is apparent not only in terms of community events but also in the changes in centrality among member nodes or keywords. Our framework draws the big picture from millions of published studies in a domain and thus provides insights into shifting research trends in science. We believe that this approach can be further applied to other scholarly archives with existing ontologies of technical terms.

## REFERENCES

[1] S. Khan, X. Liu, K. A. Shakil, and M. Alam, "A survey on scholarly data: From big data perspective," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 923–944, Jul. 2017.

[2] J. A. Evans and J. G. Foster, "Metaknowledge," *Science*, vol. 331, no. 6018, pp. 721–725, Feb. 2011.

[3] G. Goth, "The science of better science," *Commun. ACM*, vol. 55, no. 2, pp. 13–15, Feb. 2012.

[4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5228–5235, Apr. 2004.

[5] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook Latent Semantic Anal.*, vol. 427, no. 7, pp. 424–440, 2007.

[6] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," *Bayesian Nonparametrics*, vol. 1, pp. 158–207, 2010.

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

[8] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 113–120.

[9] Z. Yang and R. Algesheimer, "A comparative analysis of community detection algorithms on artificial networks," *Sci. Rep.*, vol. 6, Aug. 2016, Art. no. 30750.

[10] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, Apr. 2007.

[11] M. Goldberg, M. Magdon-Ismail, S. Nambirajan, and J. Thompson, "Tracking and predicting evolution of social communities," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 780–783.

[12] S. Saganowski, B. Gliwa, P. Bródka, A. Zygmunt, P. Kazienko, and J. Koźlak, "Predicting community evolution in social networks," *Entropy*, vol. 17, no. 5, pp. 3053–3096, 2015.

[13] M. Takaffoli, R. Rabbany, and O. R. Zaiane, "Community evolution prediction in dynamic social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 9–16.

[14] S. R. Kairam, D. J. Wang, and J. Leskovec, "The life and death of online groups: Predicting group growth and longevity," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (Wsdm)*, 2012, pp. 673–682.

[15] C. Balili, A. Segev, and U. Lee, "Tracking and predicting the evolution of research topics in scientific literature," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1694–1697.

[16] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[17] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 1, pp. 161–208, 2015.

[18] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell.* San Mateo, CA, USA: Morgan Kaufmann, 1999, pp. 289–296.

[19] L. Hong, "A tutorial on probabilistic latent semantic analysis," 2012, *arXiv:1212.3900*. [Online]. Available: http://arxiv.org/abs/1212.3900

[20] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1353–1360.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1385–1392.

[22] C. Wang, J. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 752–760.

[23] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 17–35, Jun. 2007.

[24] K. W. Boyack, R. Klavans, and K. Börner, "Mapping the backbone of science," *Scientometrics*, vol. 64, no. 3, pp. 351–374, Aug. 2005.

[25] H. Small, K. W. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Res. Policy*, vol. 43, no. 8, pp. 1450–1467, Oct. 2014.

[26] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5200–5205, Apr. 2004.

[27] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 44–54.

[28] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *J. Informetrics*, vol. 5, no. 1, pp. 187–203, Jan. 2011.

[29] A. H. Renear and C. L. Palmer, "Strategic reading, ontologies, and the future of scientific publishing," *Science*, vol. 325, no. 5942, pp. 828–832, Aug. 2009.

[30] S. Chae, A. Segev, and U. Lee, "Cannibalism in medical topic networks," *Knowl.-Based Syst.*, vol. 108, pp. 168–178, Sep. 2016.

[31] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 542–550.

[32] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: How can citations help?" in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 957–966.

[33] Y. Jo, C. Lagoze, and C. L. Giles, "Detecting research topics via the correlation between graphs and texts," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 370–379.

[34] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, no. 11, pp. 758–775, Nov. 2008.

[35] S.-H. Chen, M.-H. Huang, and D.-Z. Chen, "Identifying and visualizing technology evolution: A case study of smart grid technology," *Technol. Forecasting Social Change*, vol. 79, no. 6, pp. 1099–1110, Jul. 2012.

[36] Y. Ding, "Community detection: Topological vs. topical," *J. Informetrics*, vol. 5, no. 4, pp. 498–514, Oct. 2011.

[37] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, 2008, pp. 101–110.

[38] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: Joint models of topic and author community," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 665–672.

[39] J. Kalyanam, A. Mantrach, D. Saez-Trumper, H. Vahabi, and G. Lanckriet, "Leveraging social context for modeling topic evolution," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 517–526.

[40] J. M. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining Knowl. Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[41] R. Lambiotte and P. Panzarasa, "Communities, knowledge creation, and information diffusion," *J. Informetrics*, vol. 3, no. 3, pp. 180–190, Jul. 2009.

[42] D. Zhou, X. Ji, H. Zha, and C. L. Giles, "Topic evolution and social interactions: How authors effect research," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2006, pp. 248–257.

[43] A. T. McCray and K. Lee, "Taxonomic change as a reflection of progress in a scientific discipline," in *Evolution of Semantic Systems*. Cham, Switzerland: Springer, 2013, pp. 189–208.

[44] R. L. Ohniwa, A. Hibino, and K. Takeyasu, "Trends in research foci in life science fields over the last 30 years monitored by emerging topics," *Scientometrics*, vol. 85, no. 1, pp. 111–127, Oct. 2010.

[45] X. Polanco, "Co-word analysis revisited: Modelling co-word clusters in terms of graph theory," in *Proc. 10th Int. Conf. Scientometrics Informetrics*, vol. 2. Stockholm, Sweden: Karolinska Univ. Press, Jul. 2005, pp. 662–663.

[46] H.-N. Su and P.-C. Lee, "Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in technology foresight," *Scientometrics*, vol. 85, no. 1, pp. 65–79, Oct. 2010.

[47] M. Herrera, D. C. Roberts, and N. Gulbahce, "Mapping the evolution of scientific fields," *PLoS One*, vol. 5, no. 5, May 2010, Art. no. e10355.

[48] X. Wang, Q. Cheng, and W. Lu, "Analyzing evolution of research topics with NEViewer: A new method based on dynamic co-word networks," *Scientometrics*, vol. 101, no. 2, pp. 1253–1271, Nov. 2014.

[49] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 7, pp. 1313–1326, 2012.

[50] N. J. van Eck and L. Waltman, "Visualizing bibliometric networks," in *Measuring Scholarly Impact*. Cham, Switzerland: Springer, 2014, pp. 285–320.

[51] L. Bornmann, R. Haunschild, and S. E. Hug, "Visualizing the context of citations referencing papers published by eugene garfield: A new type of keyword co-occurrence analysis," *Scientometrics*, vol. 114, pp. 427–437, Dec. 2017.

[52] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLoS One*, vol. 12, no. 3, Mar. 2017, Art. no. e0172778.

[53] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[54] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 5, Nov. 2009, Art. no. 056117.

[55] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.

[56] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, K. Padmanabhan, and N. Samatova, "Community detection in large-scale networks: A survey and empirical evaluation," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 6, no. 6, pp. 426–439, Nov. 2014.

[57] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, p. 43, 2013.

[58] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.

[59] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.

[60] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

[61] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[62] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.

[63] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: Locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, Apr. 2006.

[64] S. Gregory, "Fuzzy overlapping communities in networks," *J. Stat. Mech., Theory Exp.*, vol. 2011, no. 02, Feb. 2011, Art. no. P02017.

[65] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, Aug. 2010.

[66] T. S. Evans and R. Lambiotte, "Line graphs of weighted networks for overlapping communities," *Eur. Phys. J. B*, vol. 77, no. 2, pp. 265–272, Sep. 2010.

[67] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008.

[68] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS One*, vol. 6, no. 4, Apr. 2011, Art. no. e18961.

[69] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5249–5253, Apr. 2004.

[70] N. Ilhan and Ş. G. Ögüdücü, "Feature identification for predicting community evolution in dynamic social networks," *Eng. Appl. Artif. Intell.*, vol. 55, pp. 202–218, Oct. 2016.

[71] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[72] C. C. Aggarwal, "An introduction to social network data analytics," in *Social Network Data Analytics*. Springer, 2011, pp. 1–15.

[73] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 2, Mar. 2007.

[74] World Health Organization. (2003). *sARS: Chronology of a Serial Killer*. Accessed: Nov. 19, 2017. [Online]. Available: http://www.who.int/csr/don/2003_07_04/en/

[75] World Health Organization. (2009). *What is the Pandemic (h1n1) 2009 Virus?* Accessed: Nov. 19, 2017. [Online]. Available: http://www.who.int/csr/disease/swineflu/frequently_asked_questions/about_disease/en/

**CHRISTINE BALILI** received the B.S. degree *(magna cum laude)* in computer science from the University of the Philippines Diliman, in 2015 and the M.S. degree in knowledge service engineering from KAIST, South Korea, in 2018. She had worked as a Management Associate with Globe Telecom. She also co-founded Roadmob and served as a volunteer for some developer communities in Manila. As an undergraduate at the UP Diliman Computer Science, she served as the Chairperson of UP ACM, from 2014 to 2015 and a Google Student Ambassador, from 2013 to 2014. Her research interests include product development, service design, and machine learning.

**UICHIN LEE** (Member, IEEE) received the B.S. degree in computer engineering from Chonbuk National University, in 2001, the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2003, and the Ph.D. degree in computer science from UCLA, in 2008. He continued his studies at UCLA as a Postdoctoral Research Scientist from 2008 to 2009 and then worked for Alcatel-Lucent Bell Laboratories as a member of technical staff till 2010. He is an Associate Professor with the Department of Knowledge Service Engineering, KAIST. His research interests include social computing systems and mobile/pervasive computing.

**AVIV SEGEV** (Member, IEEE) received the Ph.D. degree in technology and information systems from Tel-Aviv University, in 2004. He is an Associate Professor with the Department of Computer Science, School of Computing, University of South Alabama. His research interests include the DNA of knowledge, an underlying structure common to all knowledge, through analysis of knowledge models in natural sciences, knowledge processing in natural and artificial neural networks, and knowledge mapping between different knowledge domains.

**JAEJEUNG KIM** received a B.S. degree in computer science and technology from Tsinghua University, in 2007 and the M.S. degree in culture technology and the Ph.D. degree in knowledge service engineering from KAIST, in 2009 and 2019, respectively.

From 2009 to 2013 he was an HCI Researcher with the KAIST Institute for IT Convergence (KIITC), designing and evaluating novel interaction techniques for smart IoT devices. He holds over 50 patents from both South Korea and USA. His research interests include understanding human behavioral patterns through the IoT data and designing/implementing behavior change support systems for mental/physical healthcare.

**MINSAM KO** received the B.S. degree in computer science and electronic engineering from Handong Global University, South Korea, in 2009, and the M.S. and Ph.D. degrees in knowledge service engineering from KAIST, South Korea, in 2011 and 2016, respectively.

He was a Machine Learning Researcher with the Artificial Intelligence Research Institute (AIRI), South Korea, from 2016 to 2017. He was a Data Scientist with the Artificial Intelligence Team, Samsung Electronics Mobile Division, from 2017 to 2018. Since 2018, he has been an Assistant Professor with the Department of Human–Computer Interaction, Hanyang University, South Korea. His research interests include interaction between human and artificial intelligence, from exploring opportunities and challenges for novel interaction patterns according to advances in artificial intelligence to designing and implementing new AI-derived UX applications.