Journal of Informetrics 16 (2022) 101320

Contents lists available at ScienceDirect

Journal of Informetrics

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

DAC: Descendant-aware clustering algorithm for network-based topic emergence prediction

Sukhwan Jung*, Aviv Segev

Department of Computer Science, University of South Alabama, 150 Student Services Dr, Mobile, USA

ARTICLE INFO

Keywords:
Topic evolution
Topic prediction
Clustering
Topic emergence prediction
Scientometrics

ABSTRACT

Topic emergence detection aids in pinpointing prominent topics within a given domain, providing practical insights into all interested parties on where to focus the limited resources. This paper employs the network-based topic evolution approach to overcome limitations in text-based topic evolution, providing prospective topic emergence prediction capabilities by representing emergent topics by their ancestors. A descendant-aware clustering algorithm is proposed to generate non-exhaustive and overlapping clusters, utilizing the pace of collaborations and structural similarities between topics with iterative edge removal and addition processes. Over 100 datasets specific to a research topic were extracted from the Microsoft Academic Graph dataset for the experiments, where the proposed algorithm consistently outperformed existing clustering algorithms in generating clusters with a higher likelihood of being ancestors to an emergent topic up to three years in the future. Regression-based cluster filtering using five structural cluster features and topic cluster qualities showed that the prediction performance can be enhanced by automatically classifying undesirable clusters from previously known data. The results showed that the proposed algorithm can enhance topic emergence predictions on a wide range of research domains regardless of their maturities, popularities, and magnitudes without having access to the data in the predicted year, paving a road to prospective predictions on emergent topics.

Introduction

Scientific knowledge gradually expands with continuous research contributions; new discoveries are made to expand existing research fields and contribute toward new ones. Not all research is the same, however, as the participants and audiences have dynamically evolving interests. Topic evolution is a field of research dedicated to identifying how topics change over time, including survivability, maturity, and interactions between topics (Chen, Tsutsui, Ding & Ma, 2017). Tracking such changes provides insight into the current and future topical trends in a given research field, therefore, is a useful means to identify topics of high interest and popularity. Providing information on more prominent topics assists both research and industries, allowing preemptive resource investments on topics with growing needs (Carley, Newman, Porter & Garner, 2018).

Topic evolution captures changes in their semantics and relationships over time by analyzing various data sources such as unstructured documents and bibliographical datasets with metadata. Topic emergence in topic evolution is a topic evolution event where a previously unused topic is introduced to the field (Chen et al., 2017). New topics that experienced the emergence event are defined as emergent topics for the timeslot. Emergent topics can exhibit some characteristics of emerging technologies (Rotolo, Hicks & Martin, 2015) such as novel semantics, exponential growth rate, and high projected impact. They can also have connections to

* Corresponding author.
E-mail address: shjung@southalabama.edu (S. Jung).

https://doi.org/10.1016/j.joi.2022.101320
Received 18 February 2022; Received in revised form 13 July 2022; Accepted 29 July 2022
Available online 12 August 2022

other characteristics such as a high degree of uncertainty and ambiguity. Emergent topics can materialize on various levels, ranging from the addition of philosophical re-definitions of the research fields, theoretical improvements for research models, advancements in specific technologies, and new algorithms or applications. Such topics could be the outcome of topic evolution within existing topics; distinct evolutions within them culminated in a topic independent from the known evolution chains within the research domain. Topics from other research fields can migrate into a previously unaware research community as well (Osborne, Mannocci & Motta, 2017). A foreign concept with no semantical predecessors within the domain can be introduced from outside to provide novel approaches to the existing problems. In summary, emergent topics can manifest either by extensions or introductions; both types of emergent topics need to be considered for detecting or predicting emergent topics.

Traditional topic evolution methods mimic human knowledge processing by utilizing text-based topic models, extracting topics in a form of keyword vectors, then tracking their changes over time. Topic models are built with statistical word co-occurrences distributions without having content-independent identifiers. Similar topics are therefore only determined through the language model similarities; the most semantically similar topics in consecutive timeslots can be considered as a single surviving topic, with its semantic evolution determined by capturing the degree and direction of content transitions. The use of semantic similarities becomes problematic when multiple topics are linked, however, as identities of individual topics are dissolved into their shared semantics. Semantic evolution within a single topic and evolution caused by inter-topic relationships are shown with the same measure, resulting in a poor topical correlation detection (Gohr, Hinneburg, Schult & Spiliopoulou, 2009). Topical identities are tied to their semantics, therefore new topics are defined as topics with significantly distinct semantics compared to existing topics. This generates a limitation to topic evolution in terms of predictive capabilities as distinct semantics can only be retrospectively *detected* when semantic information is already known (Jung & Yoon, 2020). A semantically unique new topic is, by definition, unseen within the given dataset. Text-based topic models are therefore capable of detecting new topics as soon as the related textual data are given, but are inefficient at prospectively *predicting* new topics appearing in the future before related documents are available. Topic evolution methods on emergent topics are therefore mostly focused on detections, giving less focus to predictions. Network-based topic evolution is proposed to overcome such limitations by separating the topical identity and their semantics and provide a general foundation for a more advanced topic evolution research with topic emergence identification using underlying network patterns (Jung & Segev, 2021).

Emergent topics, be they semantically originated from existing topics within a given research domain, imported from outside, or novel to the whole research community, are almost always not *isolated;* they are used together with other existing topics when they appear. This can be represented as links within an evolving topic co-occurrence network. Detecting emergent topics in a certain timeslot, therefore, represents locating previously unseen topics in relation to their neighbors. Evolutionary event detection is done by retracing the evolution of topic networks based on the assumption that topic network experiences gradual evolution; many, if not most, neighbors of new topics are present in past timeslots as well, on which new topics are introduced as a common neighbor in the future. Emergent topics, or *descendants*, can therefore be represented by their *ancestors*, which are the presence of their (pre-existing) initial neighborhood topics in previous timeslots. Predicting emergent topics in the future is therefore transformed into a problem of identifying *ancestors* of such topics in the present and past. A previous network-based topic emergence identification method successfully detected emergent topics by training machine learning models with structural properties of their ancestors, distinguishing *ancestors* of future emergent topics from ancestors of existing topics (Jung, Datta & Segev, 2020). This approach was however limited to *predicting* emergent topics, as the method requires a set of ancestors, or *subgraphs*, to classify. Ancestors of future topics can only be known when the topic network data is available for the future timeslot, hindering prospective prediction into the future when no such data is given. Brute-forcing the possible combination is infeasible as possible combinations of subgraphs can reach up to 2^n, when large topic networks can easily have more than 10,000 nodes at a time.

A Descendant-Aware Clustering (DAC) algorithm is proposed to overcome this limitation and generate a manageable amount of topic subgraphs as *candidates* for network-based topic emergence prediction, without accessing data for the target timeslot. Instead of predicting emergent topics through their ancestors, the proposed algorithm aims to *generate clusters likely to be ancestors for emergent topics in the future*. In this problem, an emergent topic is predicted when a cluster positively matches one of the emergent topics' ancestor groups in the future. Clusters are detected under a few assumptions. Firstly, non-exhaustive and overlapping clusters are detected as not all topics contribute to the introduction of new topics. Structural properties found from trained machine learning models in previous studies (Jung et al., 2020) were incorporated into the clustering algorithm, detecting clusters with a high degree of activity over multiple years. Prospective topic emergence prediction was enabled by utilizing only ancestor information in previous years, and its performance has been validated against existing algorithms. DAC algorithm is designed to provide likely ancestors for emergent topics in the future, which can either be utilized as is or fed to the machine learning models as candidates for further processing. Data in the future timeslot is not utilized, therefore the clusters can be used to predict emergent topics before related papers are published. This paper aims to show the generalizability of the proposed method using various bibliographic datasets, each with a different research focus and interests. The algorithm first calculates the pace of collaboration (Salatino, Osborne & Motta, 2017) to measure structural similarities between topics. Edges between structurally dissimilar nodes were filtered out to signify adhesive connections between similar topics, leaving a number of connected components as cluster candidates. The components were then expanded and merged to generate a set of clusters. Once clusters were found, each cluster was regularized by edge filtration. Finally, heavily overlapping clusters are merged to find the final set of clusters. Multiple variations of DAC were implemented using different parameter values, such as *structural similarity threshold percentile* and *predicted similarity quartile*. Finally, a linear regression model was trained to learn cluster quality representing how similar they are to an emergent topic's ancestors in the future. Five structural properties were supplied as independent variables, and clusters' maximum similarity to ancestor groups was used as a dependent variable. Various filtering thresholds were used within the clustering algorithm therefore a range of threshold values were tested to

analyze their effect on the identified clusters. The proposed DAC algorithm outperformed existing algorithms over 100 topic networks from the Microsoft Academic Graph[1] dataset evolving over ten years. The DAC clusters showed higher ratios of positive matches to the answer set, outperforming not only exhaustive and overlapping clustering algorithms but also a clustering algorithm developed for the topic emergence detection task. The results indicate that the DAC algorithm is general enough to simulate ancestors of new topics over a wide range of research domains experiencing gradual topical evolution. Various research strategies were successfully captured without constructing domain-specific models, identifying *normal patterns of research* pertaining to the creation of new topics (Zhou, Huang, Zhang & Yu, 2019).

Section 2 reviews the related work on topic emergence prediction and overlapping clustering algorithms. Sections 3 and 4 detail the proposed Descendant-Aware Clustering algorithm and experimentation, and the experiment results are shown in Section 5.

Related work

Topic evolution and detection of emergent topics

Topic models are integral to topic evolution as automatically identifying evolutionary events requires means of extracting and comparing topics in a machine-readable format. Traditional topic models summarize a set of unstructured documents by extracting latent semantics in the form of word-popularity sets based on the word co-occurrence distributions. Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003) is one of the most widely used approaches to topic modeling, iteratively assigning inter-document word co-occurrence frequencies to discover topics. Topic models are represented as distinct word distributions, which are identified for each document (Steyvers & Griffiths, 2007). Topics in LDA-like models are compared by the similarities in word distributions. Word embedding (Bengio, Ducharme, Vincent & Jauvin, 2022) is another popular, and more recent, topic modeling approach, assigning numerical context to words instead of having topics as word distributions (Levy & Goldberg, 2022). Sharing the same numerical vector dimensions, topic similarities are measured in terms of vector similarities in a given multidimensional space.

Topic evolution aims to automatically track temporal changes in such topics. A collection of documents is assigned into sequentially ordered sub-collections based on their publication dates, generating a set of timeslots either with uniform or irregular lengths (Gohr et al., 2009). Topic models are generated at each timeslot to summarize its topical state at a given time (Ding, 2011). The topics in consecutive timeslots are then connected with similarity measures to form temporal topic chains, from which various evolutionary events can be detected. Dynamic topic models (Blei & Lafferty, 2006) utilized this process in the early days of topic evolution, where a dynamic topic is defined as consecutive topics chained by their word distribution similarities. Evolutionary events such as *enlarge* and *shrink* are found by analyzing the size changes over sequential neighbors. Automatic technology forecasting (Porter & Detampel, 1995) research captured chained technologies instead, employing various techniques such as extrapolation or fuzzy NLP to use the trend in a given topic to predict its future states (Battistella, 2014; Newman, Porter, Newman, Trumbach & Bolan, 2014).

More complex evolutionary events such as merge and split require recognizing interactions between multiple such topic chains, distinguishing evolution within a single topic versus evolution involving multiple topics as evolutionary theme pattern mining tried to capture (Mei & Zhai, 2005). A single topic at a specific timeslot can be connected to multiple temporal neighbors, allowing merge and split events to occur on top of each other. The use of a two-tiered topic model has been proposed for a better merge and split detection, where topic chains and topical evolutions are identified in each tier (Chen et al., 2017). Time-spanning global topics are retrieved from the whole corpus, representing a set of topics that are present over the whole document collection. Local topics, on the other hand, represent time-specific topics and are extracted from the yearly divided collections instead. The static global topics are matched to a series of dynamic local topics at each timeslot having cosine membership similarities above a given threshold. The number and sizes of matched local topics dictate the evolutionary event of the topic chain represented by the global topic; decreased and increased numbers of local topics connected to a global topic respectively represent the merging and splitting of the topic. The emergence events are detected when a previously unmatched global topic is matched to local topics in a given timeslot.

There are several studies dedicated to identifying new topics with varying definitions of topics, from simple words, through end users' interests, to consolidated keywords from publication venues. First story detection (FSD) is one of the research tasks of Topic Detection and Tracking (Fiscus & Doddington, 2002), capturing emergent topics in continuously generated text data in real-time. The goal of FSD is to search and organize new topics from multilingual news articles or identify the first article introducing the new story (Allan, Carbonell, Doddington, Yamron & Yang, 1998; Zhang, Ghahramani & Yang, 2004). For term-based topic evolution visualizations, the topical similarity is calculated by Euclidean distances between topic centroids. Topic novelties in NSF project awards are captured when a topic's distance to other topics passes over the upper range and therefore cannot co-exist with other evolutionary events (Zhang, Zhang, Zhu & Lu, 2017). Burst term detection monitors a textual data stream to capture rapid frequency growth in an attempt to capture a new topic in its infancy (Kleinberg, 2003). A multi-dimensional exploration of the research front in question was also tested by combining burst detection with keyword co-word analysis (Li & Chu, 2017). Other approaches utilized various definitions of emergent topics, such as word frequency-based research front detection (Chen & CiteSpace, 2006), new topic identification using query pattern mining (Ozmutlu & Çavdur, 2005), and integration of publication venues based on keyword similarities (Furukawa, Mori, Arino, Hayashi & Shirakawa, 2015).

Emergent topics defined as *emerging technologies* in the bibliometric domain were captured through the use of multi-layer clustering, using the intersection between VOS citation clusters and co-citation clusters (Small, Boyack & Klavans, 2014). In a more recent study,

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

topics with sufficient emergence indicators such as *prevalence, persistence, growth*, and *community* (Garner, Carley, Porter & Newman, 2017) were enhanced by authors with emergent publication behaviors. A small number of highly emergent topics were identified from publication abstract and author-term usages, showing sharp term frequency growth after their first appearance in accordance with the innovative stage of technology trend curves such as the Hype cycle. Recombinative innovations were also identified using a hybrid approach, discerning normal search patterns within a curated domain-specific conceptual model. The Chinese research community on AI was analyzed with expert consultations and network centrality measures (Zhou et al., 2019).

Network-based approaches were proposed where topics are defined by node structures within a word network (Jung, Lai & Segev, 2016). This is based on the assumption that *inventions* and subsequent *innovations* (Schumpeter, 1939) have causal relationships to the components used for the invention and their combinations (Fleming, 2001). Node prediction based on preferential attachment link prediction is proposed to classify whether the nodes in citation networks have a connection to a new node in the future (Jung & Segev, 2013), labeling the new nodes by utilizing the metadata of their neighboring nodes (Jung & Segev, 2014). Similar approaches were made with multi-layer networks for enhancing the performance. Co-author and co-word bi-layer network was used to detect recombinations within the *information science* field, simulating networks with resource allocation link prediction algorithm (Zhang, Wu, Miao, Huang & Lu, 2021). A 5-layer network combined with a multi-layer clustering algorithm was used with bibliographic coupling, co-citation, and author-citation networks over the *human computer interaction* field (Jung & Yoon, 2020). A topic's ancestors (Jung et al., 2016) were tracked over time, utilizing machine learning models with their structural features for classifying evolution events observed for their successors. The network-based approach showed high generalizability with high accuracy (Jung & Segev, 2021); it, however, had limited prediction power with the required number of input subgraphs reaching infeasible numbers with larger graphs. Topics defined as keyword clusters (Balili, Lee, Segev, Kim & Ko, 2020) were tracked over time instead for predictive capabilities but had limited success in detecting emergent topics due to the necessity of managing the corpus with unknown keywords of new topics.

Clustering algorithms

Clustering algorithms divide a given network into smaller groups of similar nodes utilizing structural information (Lancichinetti, Fortunato & Radicchi, 2008). There are multiple tasks and applications of clustering algorithms; hence multiple definitions of clusters have been utilized. This is achieved by a malleable definition of similarity that holds clusters together; examples include structural properties such as modularity or density (Kaufman & Rousseeuw, 2009), geometric metrics such as centroids and vector distances (Kohonen, 1990; Likas, Vlassis & Verbeek, 2003), statistical characteristics such as distribution type (Rasmussen, 2000), and external meta-data such as geolocations (Henriques, Bacao & Lobo, 2012). Similarity governs the definition of clusters, and task-specific algorithms with more narrow clustering definitions generally were more successful than generic algorithms (Fung, 2022).

A recent surge in data sizes has changed the dynamic, however, as many popular algorithms have computational complexities unsuited for scaling data sizes. Simpler methods such as the *k*-means algorithm were preferred in large-scale applications as they had lower complexity and hence were more adaptable in the environment (Kulis & Jordan, 2012). With the number of clusters *k*, all data points join *k* clusters with randomly generated centers based on their Euclidian distances. The process is iterated after each cluster's means are updated with their centroids until no further updates are required. The modularity maximization algorithm is another simple algorithm capable of dealing with large-scale datasets where the number of clusters is unknown (Clauset, Newman & Moore, 2004). Starting from individual data points as clusters, adjacent clusters are merged to maximize each cluster's modularity. Label propagation employs a similar agglomerative strategy, with each point updating its label to match the majority of its neighbors (Cordasco & Gargano, 2010). One of the limitations of such simple methods is that they assume that all members of the data points belong to clusters with near-uniform sizes (Yamaguchi & Hayashi, 2017), which often is not the case in large networks with high densities. Without pre-existing knowledge of the number of clusters needed, existing algorithms often result in a single cluster containing the majority of the data points due to heavy connections between them (Jung & Segev, 2014).

Overlapping clustering algorithms allows clusters to not hinge on neighboring clusters, reducing the occurrence of such problems. The Loop Edge Delete (*LED*) algorithm is an overlapping clustering designed to work on large-scale datasets with linear time complexity (Ma et al., 2016). Edge removal based on a structural similarity threshold is iteratively applied to divide clusters at each loop, starting from the whole network as a single cluster. Removed edges are then looped through to dictate cluster overlaps. The Advanced Clique Percolation Method (*ACPM*) classification algorithm was proposed to identify surging topic correlations (Salatino, Osborne & Motta, 2018). A novel topic at its embryonic stage is found in the semantically-enhanced topic evolutionary networks, which is represented by its core publications and related author information. Topic clusters with notable recent collaborations are then regarded as the ancestors of such novel topics (Salatino et al., 2017). The pseudo-clique definition of clusters in the original clique percolation method (Palla, Derényi, Farkas & Vicsek, 2005) resulted in cluster size disparity problems, and the advanced algorithm tried to overcome this with additional processes including intensity-based clique filtering and neighborhood extraction with local maxima. However, added processes made the algorithm more complex and not suitable for large networks.

The descendant-aware clustering algorithm

The descendant-aware clustering algorithm is focused on identifying possible common ancestors of new descendant topics in a domain of interest. Many topical qualities such as popularity, importance, and maturity affect the formation of such candidates; few topics will be ancestors of multiple new topics while some others remain un-contributing. Traditional exhaustive clustering

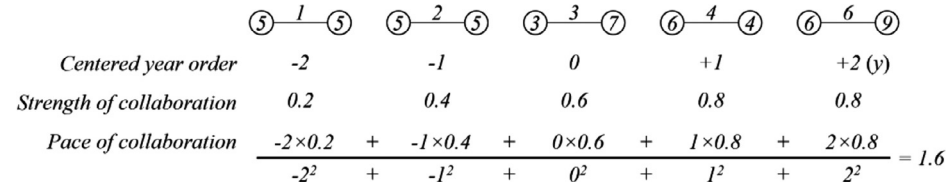

Fig. 1. Example of the pace of collaboration calculation with a five-year evolution window. Numbers represent topic frequencies and co-occurrence frequency each year.

algorithms assign all topics to a single cluster and hence are not suitable for detecting ancestor groups. A non-exhaustive overlapping clustering algorithm is proposed to overcome this issue, utilizing structural similarities in loops to allow algorithms to work with highly connected networks which are often the case for topics in bibliographical datasets.

The algorithm is divided into three steps. Bibliographic datasets with publication-topic links are first converted into a year-specific topic network, incorporating evolution in previous years. Edges between structurally dissimilar nodes are removed to signify adhesive connections between topics. Connected components in the filtered network are then expanded and merged to generate a set of clusters. Finally, each cluster is regularized by filtering edges and merging heavily overlapping clusters.

Generate topic networks from a bibliographic dataset

A bibliographic dataset is defined as a dataset containing publication records with related metadata; in this paper research topics are assigned to publication records in a form of metadata. The DAC algorithm is proposed with the assumption that interaction between topics alone is sufficient for performing the task at hand and access to their semantic information is unnecessary. The topics can either be manually assigned topics such as keywords or automatically detected topic models as long as a single channel is utilized throughout the process, having many-to-many connections to the publications. None of the metadata is used in clustering algorithms and hence is not required for other than identification purposes.

A given bibliographic dataset is first converted into a network format by defining topics as nodes $n \epsilon N$ and topic co-occurrences as edges $e=(u,v) \epsilon E$. Topic co-occurrences are observed when a pair is linked to a common publication, and edge weights W are measured as topic co-occurrence frequencies. Topics and co-occurrences that appeared in given year y are combined to form a topic network G_y with the non-zero occurrence and co-occurrences frequencies denoted as node and edge weights $W(n_y)$ and $W(e_y)$.

$$G_y = \left(N_y, E_y\right), and W\left(n_y\right) \in N^+, W\left(e_y\right) \in N^+ \tag{1}$$

Topics are dependent on predecessors and therefore multiple previous years should be considered when building a topic network to generate more accurate ancestors. Evolutionary window ω is defined to represent the number of previous years that are used to generate a year-specific topic network T_y. ω governs the length of past histories considered in building the topic network; in the basic scenario, $\omega = 1$ would indicate that only the current year y is considered without tapping into further historical behaviors. The pace of collaboration (Salatino et al., 2017) is calculated to incorporate topic co-occurrence patterns in previous networks as it is shown to signify the rapid growth rate often associated with emergent technologies. Node weights are first merged into edge weights by incorporating the strength of collaboration the edge represents, which is calculated as a harmonic mean between edge nodes normalized by the edge weight. The harmonic mean is used in an attempt to mitigate the effect of extremely frequent outlier topics with high co-occurrences with many others.

$$S_y(e) = \{Harmonic Mean(W(e)/W(v), W(e)/W(u)) | \forall u, v \epsilon N_y, Ee = (u,v) \epsilon E_y\} \tag{2}$$

Previous co-occurrence trends affect the formation of new topics. Topics with growing popularity are more likely to contribute towards the introduction of new ones with enough research impact, as the focused research interest in the topics will be extended to the new topics. Topics with declining popularity indicate otherwise, having not only a lower probability of participating in the generation of new topics but also having less research impact to convey. The strength of collaborations over the unilateral window with length ω in the past are then combined to capture such changes in the pace of popularity in a form of a regression slope for topic co-occurrences. A *centered year order* is used to weight the values at different years; starting from the most recent year y, diminishing weights are allocated for past years reaching negative values after the midpoint in the window ω. This is to grant more emphasis on recent behaviors while penalizing the distant past for a better regression outcome. Fig. 1 illustrates an example of how the pace of collaboration is calculated to be used as edge weight in the final topic network when $\omega = 5$. The *centered year order* starts from +2 for year y reaching down to −2 for y-4. For a list of evolution window years $l = (y - \omega, y]$, the pace of collaboration is denoted by

$$P_y(e) = (\sum\nolimits_l ((y - -\bar{y}) \times S_y(e)) / \sum\nolimits_l (y - -\bar{y})^2) \tag{3}$$

where $\bar{y}$ is the mean value of years used in topic network generation. This results in a topic network T_y with non-weighted nodes which are more commonly used for clustering, representing the first-degree linear regression results with normalized years. $P_y(e) < 0$ indicates a negative frequency slope over time, and all edges with negative collaboration paces are removed from the topic network.

$$T_y = \left(N_y, E_y\right), \forall P_y(e) \epsilon \mathrm{R}^+ \tag{4}$$

Detect overlapping topic clusters

Once the topic network is built from a bibliographic dataset, overlapping clusters are found with a structural clustering approach. Structural similarities between all node pairs are first calculated, where the edges can be either unweighted or weighted. *Common neighbor size* (Ma et al., 2016) represents the ratio of friends shared by a node pair, which is calculated by normalizing the number of their common neighbors with the geometric mean of their neighborhoods. It does not require edge weight information and therefore is used as *the structural similarity value for unweighted networks*. With the neighborhoods of a node n in year y denoted by $\Gamma_y(n)$, the unweighted structural similarity of a node pair (u,v) is calculated as

$$\sigma_y(u,v) = |\Gamma_y(u) \cap \Gamma_y(v)| / |\Gamma_y(u)| \times |\Gamma_y(v)|^{1/2} \tag{5}$$

which can be replaced by $\sigma_y(e)$ for edge e connecting the pair.

Additional considerations are needed to deal with weighted networks as edge weights are often as important as the edge presence. This is achieved by multiplying *weighted vector similarity* and *normalized weight* to *common neighbor size*, reflecting the pace of collaboration between the paired nodes. *Weighted vector similarity* represents the similarity between the given nodes' collaboration patterns and their respective neighbors and is calculated as cosine similarity between edge weight vectors from each pair and their common neighbors. A pair of nodes can share no common neighbors, therefore each node in a pair is considered a neighbor to its counterparts. This ensures that there is at least one shared edge in any situation resulting in a non-zero outcome. *Normalized weight* represents the relative pace of collaboration between the given node pair, normalized by the maximum pace of collaboration in a given topic network. It is expected that larger values of both the *weighted vector similarity* and *normalized weight* would be observable for structurally similar nodes. Both variables range from 0 to 1 therefore the overall similarity will be lowered with significant diminishing values for dissimilar pairs. The *weighted structural similarity* of a node pair (u,v) in y is calculated as

$\sigma_y'(u, v) = \sigma_y(u, v) \times vecs(u, v) \times normw(u, v)$, $\mathrm{vecs(u, v)} = \mathrm{Z_y(u, v)} \bullet \mathrm{Z_y(v, u)} / ||\mathrm{Z_y(u, v)}|| \; ||\mathrm{Z_y(v, u)}||$,
$normw(u, v) = \mathrm{P_y}(u, v) / \max(\mathrm{P_y}(u, v))$ where

$$Z_y(a,b) = \left[P_y(a,x) \forall x \varepsilon \Gamma_y(a) \cap \Gamma_y(b)\right] \tag{6}$$

with each component shown per line.

Edge filtering is conducted as a next step to remove weak links from the network. This is especially necessary when working with dense networks, as an average of 88.54 edge-to-node ratio was observed over the 103 topic networks used in the paper. *Structural similarity threshold* α is used to filter out all edges with insignificant structural similarities ($\sigma_y(e) < \alpha$), which are removed from T_y along with any isolated nodes after edge removal. 0.381% of the topics were classified as emergent topics on average over the experimented datasets therefore a majority of the edges would be insignificant even with the high degree of edge-to-node ratio. Structurally dissimilar nodes are disconnected, transforming a dense topic network into a set of multiple connected components each sharing high structural similarities within them.

$$T_y = \left(N_y, E_y\right), \forall e_y = (u,v) E \sigma_y(u,v) \geq \alpha \tag{7}$$

More structurally homogeneous components remain as α increases, effectively rendering the connected components as clusters consisting of structurally similar nodes. A fixed α value would not perform well over multiple networks as different topic networks show various levels of structural similarities, therefore a *structural similarity threshold percentile* α' is introduced; α is calculated as the minimum threshold value that satisfies α'th percentile. Components with three or fewer nodes are considered too minor and were disregarded from further processing. Such an approach, however, is limited in that only direct similarities are measured. A node can be excluded when its connection to a cluster is diluted over multiple connections. Such nodes can be beneficial to clusters even when there are no links with similarity values $> \alpha$. Connected components are therefore expanded to capture non-direct structural similarities between them and their neighboring nodes.

Connected component expansion is done based on the PageRank (Page, Brin, Motwani & Winograd, 1999) value maximization, using a static value calculated across T_y. For each connected component $c \subset N_y$, an average PageRank score of its member nodes is compared against PageRank scores of boundary nodes $\Gamma_y(c)$ which are a non-overlapping neighborhood nodes for all members of a component c in the original topic network before edge filtration. A boundary node v is added to the connected component c if its score $PR(v)$ is greater than the component's average $|PR(n)|$; the PageRank from the original T_y is used to consider both the removed and remaining nodes.

$$\forall v \varepsilon \Gamma_y(c), c = c + v when PR(v) > |PR(n)| \forall n \varepsilon c \tag{8}$$

The expansion process is done from the perspective of clusters, without considering the structural similarity values of the removed edges. This is done to capture the importance of individual nodes as well as their similarities. The expansion process in Eq. (8) is repeated ε times to allow additional paths of maximum length ε to be added to any given component. The resulting connected components are defined as *topic clusters*, which are non-exhaustive and can overlap.

Cluster postprocessing and evaluation

Postprocessing is done once the clusters are identified to regularize the cluster size (Salatino et al., 2018). The highly dense nature of the topic networks results in very large clusters dwarfing smaller clusters and skewing structural properties. A maximum number of edges per cluster m and cluster merging threshold τ are used for postprocessing, each reflecting the ratio of less influential topics

Table 1
Five features used for training linear regression models.

Feature	Description
#nodes	Number of nodes in a component
#edges	Number of edges in a component
cohesion	Ratio of internal and external edges
deg	Mean degree in a component
pr	Mean PageRank in a component

within a cluster and how tolerant clusters are in sharing common topics. Modifications to these variables will result in clusters with different topical coverage; larger clusters will result in lower precision in exchange for higher recall.

Edge pruning is done under the assumption that the collaboration activities between topics affect their importance within the cluster. Edges in each cluster are first sorted by their pace of collaboration weight $P_y(e)$, and top m edges with the highest weights $P_y(e)$ are selected. The rest of the edges are discarded along with any nodes that become isolated as a result of edge deletion; only the core structure of each cluster is extracted, allowing a single to be disconnected. Cluster merging is done afterward to consolidate clusters with high overlap. Jaccard similarities between all cluster pairs are calculated using the membership node size and overlap, and any pair with similarity $> \tau$ are considered similar and merged into a single cluster instead. The merging process is repeated after initial cluster pairs are exhausted until no more merging can be done. Post-processed clusters are validated by matching resulting clusters against an answer set, actual ancestors of emergent topics in the same y. Jaccard similarity is again used to measure cluster similarities; the similarity between i th cluster $C_{y,i,}$ and *j-th* answer set $A_{y,j}$ in year y is calculated using their union and intersection sizes as shown in Eq. (9).

$$Jaccard(i,j) = |C_{y,i} \cup A_{y,j}| / |C_{y,i} \cup A_{y,j}| \tag{9}$$

Outcome values range from 0 to 1, with 0 showing zero similarity and 1 showing identical membership. A cluster is considered a positive match when its similarity to any answer set member is above a given *cluster similarity threshold* θ.

Evaluate topic cluster quality

Regression models are trained to evaluate *cluster qualities*, which are measured as the maximum membership similarities they have with the actual ancestors. Retrospective emergent topic detection is performed to test the possibility of multi-year prediction as the similarity to the actual ancestors in the future equates with the likelihood of a given cluster being the ancestor of an emergent topic in the future as well. Year distance parameter d represents the year distance between the cluster and answer sets and dictates how much future is being predicted; $d = 0$ indicates the clusters are analyzed in the same year, while $d > 0$ indicates the clusters are used to predict emergent topics in the future year $y + d$ using their ancestors in the time period indicated by y.

Multiple regression models were used in the previous research for a binary classification problem, predicting whether given ancestor topics have an *emergent topic* as their common future neighbor using their structural properties; the Linear Regression (LR) model showed high-accuracy results (Jung et al., 2020). While the tasks are not identical, this is in essence measuring how likely the found clusters are to have *emergent topics* as their common future neighbors. The authors believe similar models would perform well, and the LR model is selected for this experiment. Five cluster features shown in Table 1 were used as independent variables when training an LR model for each y, measuring the structural properties of detected clusters in y and future years. All features are standardized to have a mean value of 0 and a standard deviation of 1 to reduce the effect of having features with different value ranges.

Each cluster's maximum Jaccard similarity to an answer set is used as a dependent variable *cluster quality*. The year distance parameter d is used to capture the cluster quality for years $[y,\ldots,y+d]$ by comparing answer sets $[A_y,\ldots, A_{y+d}]$ *to* C_y as shown in Eq. (10). The variables in the last year $y + d$ are then used as the test set, while the data in years y to $y + d$-1 are used as the training set. With varying degrees of d, the independent variable *varX* and dependent variable *varY* are generated as

$varX_y(d) = C_{n,i}[\#nodes, \#edges, cohesion, deg, pr]$,
$varY_y(d) = \max(|C_{n,i} \bigcap A_{m,j}| \;/\; |C_{n,i} \bigcup A_{m,j}|)$,

$$\forall C_{n,i} \varepsilon C_n \forall A_{m,j} \varepsilon A_m where y \le n < y + d, m = y + d \tag{10}$$

Experiments

Preprocessing dataset

A heterogeneous bibliographic dataset called Microsoft Academic Graph (MAG) (Sinha et al., 2015; Wang et al., 2019) was extracted to generate a set of datasets each focused on a specific research topic. While it is scheduled to be retired at the end of 2021 and the dataset is not going to receive additional data input afterward, the records up until the late 2010s were comprehensive enough to be competitive with other major bibliographic search engines such as Google Scholar or Scopus (Hug, Ochsner & Brändle, 2017). The MAG dataset was used as it allowed bulk download of a data snapshot weekly, and a version of MAG in February 2020 is downloaded for preprocessing through Microsoft Azure Databricks.

Topics independent of research domains were used in this paper even though domain-specific terms such as author-assigned keywords are shown to produce better quality topics (Salatino, Osborne, Thanapalasingam & Motta, 2019). This is to minimize the effect of semantic variations between domains as well as term variations. The MAG provides a hierarchical ontology for document-assigned topics named fields-of-study (FoS) (Shen, Ma & Wang, 2018), each representing different research concepts found within the recorded research articles. A six-level FoS hierarchy is generated and updated monthly using Wikipedia articles, applying knowledge base type prediction methods along with graph link analysis and convolutional neural network techniques. Then the recorded documents are processed each week with large-scale, multi-level text classifications to update the FoS-document tagging relationships. The FoS represents dataset-wide representative terms and their assignments to each document and therefore was defined as the research topics in this paper. This removes the computational complexity of large-scale natural language processing over nearly two billion publications from the experiment, allowing larger scale experiments over wide range of domains.

The extracted dataset has a total of 197,642,464 publications, 709,934 research topics, more than 1.5 billion citation links, and 1.3 billion paper-topic links. It is extremely taxing for clustering algorithms to operate on a network of such size; therefore the dataset is divided into multiple datasets, each covering a specific research domain. Research topics in the second-highest level in the FoS ontology hierarchy with similar popularities were first extracted as core topics, each representing individual research domains. Then a dataset is generated by extracting all publications that share a link to the core topic, and all topics linked to the document collection. Topic co-occurrences frequencies are recorded yearly to produce an evolving topic network over given years. From 292 available topics, a total of 103 domain topics with medium popularity were selected to generate 103 domain-specific datasets. The resulting datasets had on average 48,467 topics and 3965,339 topic co-occurrences, with large standard deviations of 18,259.48 and 1217,796. Table 8 shows the full list of datasets with a number of topics and their co-occurrences.

DAC predicts the emergence of novel topics by measuring a likelihood of a current topic set being direct ancestors of a future topic. The topic novelty is defined within the confinement of a single domain; topics from other domains are considered new when they are introduced to the research domain for the first time. Topics in multiple bibliographical datasets are therefore clustered independently of their states and relationships in other datasets. A bibliographic dataset is used to generate a topic network where emergent topics equate to nodes that are newly introduced in the network at a given timeslot. Validation of the experiment result requires a set of outcomes predetermined to be correct, therefore a set of emergent topics is extracted from evolving topic networks to validate DAC's performances. Utilizing all known emergent topics is impractical and could be tainted by several low-quality topics; hence a series of filtration is done to select topics with top quality.

The bibliographic dataset with topic-publication relationships is used to generate a yearly set of emergent topics $t \in T_y$ for each given year y. The emergent topic is defined as the topics with at least min_*freq* number of appearances before y; min_*freq* = *0* would result in the simple definition of emergent topics as any topic that was never used before. While being the simplest way to define emergent topics, this is not always a safe approach to follow. The MAG allowed retrospective term assignments assigning topics to a distant source paper, which went dormant for a long time before gathering community attention. Misused labels could occur as well when the labels for a given topic are mentioned in unrelated publications under different semantics. In order to smooth the outlier cases, min_*freq* = 5 is used instead to filter out possible outlier topic usages in the initial stages of its lifespan, such as irregular uses, retrospective topic assignments, label misuses, or minor topics without sustained uses. Innovation of topics, which are separate from the actual technical invention of said topic and often materialize at different times, are captured by filtering out initial outlier occurrences (Fleming, 2001).

Ancestors of an emergent topic $A_{y,t}$ are defined as the set of topics with non-zero co-occurrences to t in y. For example, $A_{2005,social\_engagement}$ in a *human-computer interaction* domain would include *multimedia* and *user interface* as a basis along with more directly related topics such as *robot, virtual machine, psychology, artificial intelligence,* and *everyday life*. This includes a large portion of links with less importance; hence the ancestors are filtered by their contributions to the emergent topics. This is measured in terms of co-occurrence intensity inspired by the intensity of collaboration (Salatino et al., 2018), resulting in higher values for ancestors with rare connections to the topic in question. To reduce the yearly variations from the equation, five years after y are searched to analyze the initial topic trend instead of just looking at topics in y. Co-occurrence intensity between t and $A_{y,t}$ is calculated for each ancestor member *anc* as below, using topic occurrence frequency and topic co-occurrence frequencies.

$$CI_{t,a} = \left(\sum_{z} \left(|t_z| - |anc_{z,t}| \right)^2 \right)^{1/2}, y \le z < y + 5 \tag{11}$$

CI is the Euclidian distance between two vectors, each representing emergent topic frequency and co-occurrence frequency with the given ancestor topic. The smaller the distances, the more commonly two topics appeared in publications together over an initial stage of the emergent topic. To maintain the regular quality of ancestor topics, $A_{y,t}$ is filtered with CI values to select the top *25 anc* with the smallest distances to t. The golden set is extracted for each of the 103 domain-specific datasets generated from the previous section for $y = [2001,\ldots,2010]$. The number of emergent topics steadily increased from 14,414 in 2001 to 24,854 in 2010, averaging at 139.94 to 241.30 for 103 datasets. Most of the emergent topics had more than 25 neighbors and thus filtered out based on the CI value, resulting in average neighbor sizes of 24.80 per emergent topic. After reducing duplicate selections, the number of topics selected as neighbors of the golden set takes more than a quarter of available topics in the years; around 26% of the existing topics were regarded as significantly contributing towards at least one emergent topic. Detailed information on the golden set for two sample

Table 2
Parameters used in the experiment and their values.

Parameter	Value	Description
y	[2001,…,2010]	Experimented year
ω	5	Evolutionary window
d	[0,1,2,3]	Year distance parameter
α'	[0.90, 0.95]	Structural similarity threshold percentile
ε	3	Connected component expansion iteration counter
m	15	Maximum number of edges per cluster
τ	0.70	Cluster merging threshold
θ	[0.01, 0.02, …, 0.75]	Matching cluster similarity threshold
θ'	[0, 1, 2, 3]	Predicted similarity quartile

years y = [*2005, 2010*] is shown in Table 8. The golden answer sets along with the list of graphs generated from the previous section are uploaded as python object binary files to the Zenodo repository[2] for open access.

Evaluating DAC's performance

Table 2 shows the list of variables used in the experiment. The experiments were focused on the 21st century, applying clustering algorithms on ten datasets with y = [2001, …, 2010]. Years from 1997 to 2000 were used to accommodate a five-year evolution window ($\omega = 5$), while years from 2011 to 2013 were used to evaluate the performance when predicting emergent topics for up to 3 years (d = [0,1,2,3]). Edges with structural similarities $\sigma_y(u,v)$ below α are filtered out to retain the top 10% and 5% of most similar topic pairs (α' = [0.90, 0.95]), which form a set of connected components which are then expanded by adding neighboring nodes with enhanced PageRank values up to distances of three (ε = 3). Clustering postprocessing is followed by selecting the top fifteen edges with the highest pace of collaborations per cluster, then merging clusters with Jaccard similarity above 0.7.

LR models are trained to predict the performance of detected clusters through the metrics of how close the clusters are to becoming ancestors of emergent topics in the present and future. The authors assume that the predicted similarity to an answer group is positively correlated to the cluster having a positive match to one. This is tested by two levels of thresholds. First, the experiments were conducted with incremental *cluster similarity threshold* θ until no positive matches were detected (2*dp*); the maximum threshold value with positive matches was found to be $\theta = 0.75$. Then three sets of *predicted similarity quartile* θ' were used to first filter out the clusters with low predicted Jaccard similarities to ancestor groups. Instead of using a set number, θ' is dynamically assigned by using 1st, 2nd, and 3rd quartile values for any given cluster outcomes with the unfiltered outcomes denoted as $\theta' = 0$. Both weighted $\sigma_y'(e)$ and unweighted $\sigma_y(e)$ structural similarities were calculated, resulting in a total of up to 48,000 results gathered for each of the 103 datasets with 75 θs and three predicted similarity quartiles.

DAC's performance is compared with other existing methods, including *ACPM*, Clauset-Newman-Moore greedy modularity maximization (*Greedy*) (Clauset et al., 2004), and *LED* using the structural similarities as edge weights. All algorithms shared the same network generation and postprocessing steps mentioned in sections 3.1 and 3.3 to share identical topic networks with a single weight parameter and regulate the cluster numbers and sizes. The authors used a high-performance computing service by Alabama Supercomputer Authority[3] to process the algorithms. The Networkx library's *greedy modularity communities* function is used to run the *Greedy* algorithm,[4] while *ACPM* and *LED* were implemented following respective publications (Ma et al., 2016; Salatino et al., 2018). Some modifications were made to the *LED* algorithm due to the lack of explanation. For example, expansion with isolated nodes is not implemented as the structural similarity used for attaching nodes to clusters is edge-specific values; the similarities are iteratively re-calculated for each cluster during each loop, having no static global values to be compared with. It is inconclusive which values are to be used when outside nodes are concerned and which edges should be restored. They are not likely to cause a significant difference in this experiment setting where only core structures of clusters are used, and hence were deemed unnecessary; the modified *LED* algorithm is defined as *LED-m* in the result section.

Results

Medium popularity topics are selected to reduce the effect of computational complexities caused by network sizes; FoS with the highest level in the MAG's FoS ontology were not used during experiments as their size caused some of the existing algorithms to slow down considerably. Table 3 shows the resources required to run clustering algorithms on a dataset extracted for the *Computer Science* domain with y = [*2005, 2010*], which processed 180,967 topics and 129,915,515 topic co-occurrences in total. A modified version of *LED* (*LED-m*) is showing very little resource usage because isolated nodes were not considered, which is computationally expensive with large graphs. On the other hand, *ACPM* used far more memory resources and took more time, returning an out-of-memory error

[2] https://zenodo.org/record/5746108
[3] https://hpcdocs.asc.edu/
[4] https://networkx.org/

Table 3
Computational time and memory usage for five clustering algorithms on a *Computer Science* dataset.

Algorithm	Year	Time (H:M:S)	Memory (Gb)
LED-m	2005	**0:06:12**	1.31
LED-m	2010	**0:08:07**	2.09
DAC, $d = 0$	2005	0:20:07	1.81
DAC, $d = 0$	2010	0:33:41	2.84
DAC, $d > 0$	2005	2:52:32	1.81
DAC, $d > 0$	2010	4:31:40	2.84
Greedy	2005	6:29:14	**0.97**
Greedy	2010	10:35:02	**1.45**
ACPM	2005	*22:30:27*	*491.72*
ACPM	2010	–*	–*

a.*The run was crashed with a memory out error exceeding 500Gb.

Table 4
Linear regression performance metrics predicting how similar clusters are to the actual ancestors averaged over ten years on all 103 domain-specific datasets.

	α'	d	R^2	MAE	MSE		α'	d	R^2	MAE	MSE
$\sigma_y(e)$	90	**0**	0.6631	0.0314	0.0017	$\sigma_y'(e)$	90	**0**	0.6879	0.0311	0.0017
		1	0.5711	0.0355	0.0022			**1**	0.5885	0.0355	0.0021
		2	0.5844	0.0346	0.0020			**2**	0.6046	0.0345	0.0020
		3	0.5907	0.0345	0.0020			**3**	0.6142	0.0343	0.0020
	95	**0**	0.5986	0.0338	0.0020		95	**0**	0.6084	0.0330	0.0019
		1	0.5115	0.0374	0.0024			**1**	0.5145	0.0366	0.0023
		2	0.5211	0.0367	0.0023			**2**	0.5289	0.0358	0.0022
		3	0.5291	0.0365	0.0023			**3**	0.5351	0.0357	0.0022

with maximum requestable memory of 500Gb from the Alabama supercomputing center. The proposed method is faster and memory efficient compared to the *Greedy* algorithm as well.

LR models are trained for each of 103 datasets over ten years, having either unweighted or weighted edges filtered by two *structural similarity threshold* percentiles. Five independent variables are used as independent variables, while a maximum Jaccard similarity to a set of ancestors is used as a dependent variable. The models' performances are first measured by their regression accuracies predicting *cluster quality* (with the value range of 0 to 1) in the given year and following future years. Table 4 shows the trained LR models' performances in coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). Negative R^2 values can be generated for computational purposes; negative R^2 values indicate that none of the data can be explained by the given model, therefore are considered zero in the table. The predicted values are compared against the answer sets up to three consecutive future years from 2001 to 2010. The results show that the regression model in DAC is capable of predicting *cluster qualities* with less than 3.5% differences over all combinations tested during the experiment, showing on average 0.0348 differences in MAE. R^2 averaged at 0.5782; around 58% of the future cluster qualities can be explained by the regression models.

While it can be regarded as adequate explainability, R^2 values are noticeably lower than the results of a previous topic emergence prediction research with $R^2 > 0.9$ (Jung et al., 2020). This can be attributed to the differences between *detection* and *prediction*; The DAC clusters are detected without having access to the timeslot where the actual ancestors are located as opposed to the previous research. Extracting the exact member of the subgraphs before they form is a challenging task, therefore the DAC algorithm instead aims to detect the influential core members of the ancestor groups. Such an approach is to overcome the innate limitation of prospective prediction at the expense of lower ancestor membership explainability; some ancestor members are left unexplainable by the DAC clusters on purpose to better capture the core members. The matching cluster similarity threshold θ is not used during the regression training to boost the results as well. Lower R^2 values in higher structural similarity thresholds support this explanation as higher α' results in more tightly connected clusters. On average 6.23% and 7.71% lower explainabilities were observed for $\sigma_y(e)$ and $\sigma_y'(e)$ as there are fewer cluster members to be matched. Another likely explanation is that the variance in 103 datasets with different research foci and publication behaviors caused randomness in the outcome as opposed to the previous research using datasets that had fewer variabilities. Another possibility is that there are different stages of evolution in research domains, where the proposed method is no longer applicable due to the drastic changes in topic co-occurrence patterns. A *mineralogy* domain is an example showing extremely negative R^2 values averaging at -37.80 when $y = 2010$ and $d = 3$, a sharp decrease in the values in all other y and smaller d combinations with an average $R^2 = 0.5725$. Such low values with a specific *year* and *year distance* indicate there was an evolutionary fissure between 2010 and 2013 in the field of *mineralogy*, and structures of an emergent topic's ancestors in 2010, while showing high explainability to ancestors in 2011 and 2012, are no longer viable in predicting an emergent topic's ancestors in 2013. Sudden drops in R^2 were observed with specific combinations of y and d in several other domains including *civil engineering*. These findings support the idea that research domains sometimes experience a year of rapid evolutions, which cannot be effectively captured using continuous extrapolative approaches.

Table 5
Average structural similarity threshold α, based on two structural similarity threshold percentiles α' over all datasets, for both unweighted and weighted edges.

	$\sigma_y(e)$		$\sigma_y'(e)$	
	α'=0.90	α'=0.95	α'=0.90	α'=0.95
2001	0.3028	0.3834	0.0077	0.0193
2002	0.3015	0.3813	0.0075	0.0187
2003	0.2973	0.3756	0.0072	0.0179
2004	0.2917	0.3668	0.0066	0.0165
2005	0.2890	0.3635	0.0063	0.0158
2006	0.2863	0.3587	0.0061	0.0152
2007	0.2836	0.3552	0.0056	0.0141
2008	0.2806	0.3517	0.0052	0.0132
2009	0.2778	0.3466	0.0050	0.0127
2010	0.2733	0.3392	0.0045	0.0113
mean	**0.2884**	**0.3622**	**0.0062**	**0.0155**
var	**1.00E-04**	**2.18E-04**	**1.18E-06**	**7.19E-06**

Table 5 shows *structural similarity threshold* values calculated with two different percentiles over the years, averaged over 103 datasets used in the experiment. There are clear disparities between threshold values for $\sigma_y(e)$ and $\sigma_y'(e)$ with more than 30 times the differences in their respective average values of 0.3253 and 0.0108. Such differences are caused by multiplying *weighted vector similarity* and *normalized weight* to the unweighted structured similarity in Eq. (6), granulating initial *common neighbor size* values even further. Differences within $\sigma_y(e)$ and $\sigma_y'(e)$ are also statistically significant for different threshold percentile α', giving a higher average to a higher percentile with p = 3.17E-12, 1.86E-8 with two-tailed t-tests.

Differences in α show that the algorithm is filtering insignificant edges as proposed, while a larger variance in α for higher α' justifies the use of percentile thresholds instead of a fixed value over different topic networks exhibiting different levels of structural similarities. This is less pronounced in Table 5 as variance differences for $\sigma_y(e)$ are insignificantly different with p = 0.13 using a two-sample F-test, which is reduced to 1.07E-07 when the thresholds are divided by 103 datasets instead of years. A steady decrease in α over years with a correlation coefficient over 0.99 indicates that topic networks in general experience slow but steady evolution. Significant topics are becoming more dynamic with their roles in multiple communities with the introduction of novel topics to the research domains in question.

There can be many-to-many relationships between found clusters and golden answer sets, therefore modified definitions of precision and recall (Salatino et al., 2018) were used. Precision is defined as the fraction of clusters that resulted in positive matchings, while recall is defined as the fraction of golden sets that resulted in positive matchings.

Precision = |clusters(matched)| / clusters,

$$Recall = |ancestors(matched)|/ancestors \tag{12}$$

There is clear evidence that DAC outperforms existing clustering algorithms in terms of topic emergence prediction in both precision and recall while showing comparably fewer differences within DAC variants. Fig. 2 shows the average performance metrics for three existing algorithms along with an average of four DAC instances using two α' with $\sigma_y(e)$ and $\sigma_y'(e)$ when θ' = 0; the error bars mark standard deviations for each clustering algorithm. All graphs show the highest performance observed when the cluster similarity is most relaxed with θ=*0.01* as only 1% of common members warrant a positive match. Performance measures diminish with higher θs as stricter cluster matching is performed leading to fewer, but more similar, matches. The *Greedy* algorithm showed the worst result as a modularity-based clustering algorithm with no specific arrangement for topic evolution; it recorded F1 < 10% even when less than 5% of common members were required for a positive match. Its poor performance led to higher *std* when θ reaches 0.3 even though precision and recall showed little deviations from their mean values; the differences between higher precision and lower recall become too small with low performances, and random fluctuations dwarfed the performance metrics. The *LED-m* algorithm showed a slight improvement compared to the *Greedy* algorithm with minimal cluster threshold values but showed a jagged performance descent with large standard deviations. Precision was considerably lower for *LED-m* compared to its recall values, indicating that a small fraction of highly impactful clusters were matched to many ancestor groups. The *ACPM* algorithm retained much higher F1 values showing more than double the F1 values compared to the *LED-m*, with relatively steady precision and recall values. This is not far from the performance measured in Salatino et al. (2018), where the task-specific *ACPM* outperformed existing overlapping clustering algorithms with higher precision and lower recall. These results reflect the focus of clustering algorithms, where *Greedy* and *LED-m* algorithms are focused on cluster detections while *ACPM* is specifically designed to be used in the topic evolution domain. DAC is a dedicated clustering algorithm for network-based topic emergence prediction and shows the best curves. While four variations exhibited minimal differences compared to other existing algorithms, all of them showed F1 over 0.75 when clusters were needed to correctly capture at least 15% of ancestor group members. All variations showed statistically significant improvements over other algorithms, having 67% of the precision and recall points above the *ACPM*'s positive standard deviation range over $0.08 \leq \theta \leq 0.22$.

Among the four DAC variations, $\sigma_y(e)$ with α' = 0.95 showed the best result as shown in Table 6. Further analyses were made based on the selected parameters. The effect of year distance parameter d was statistically significant but not surprising; F1 steadily

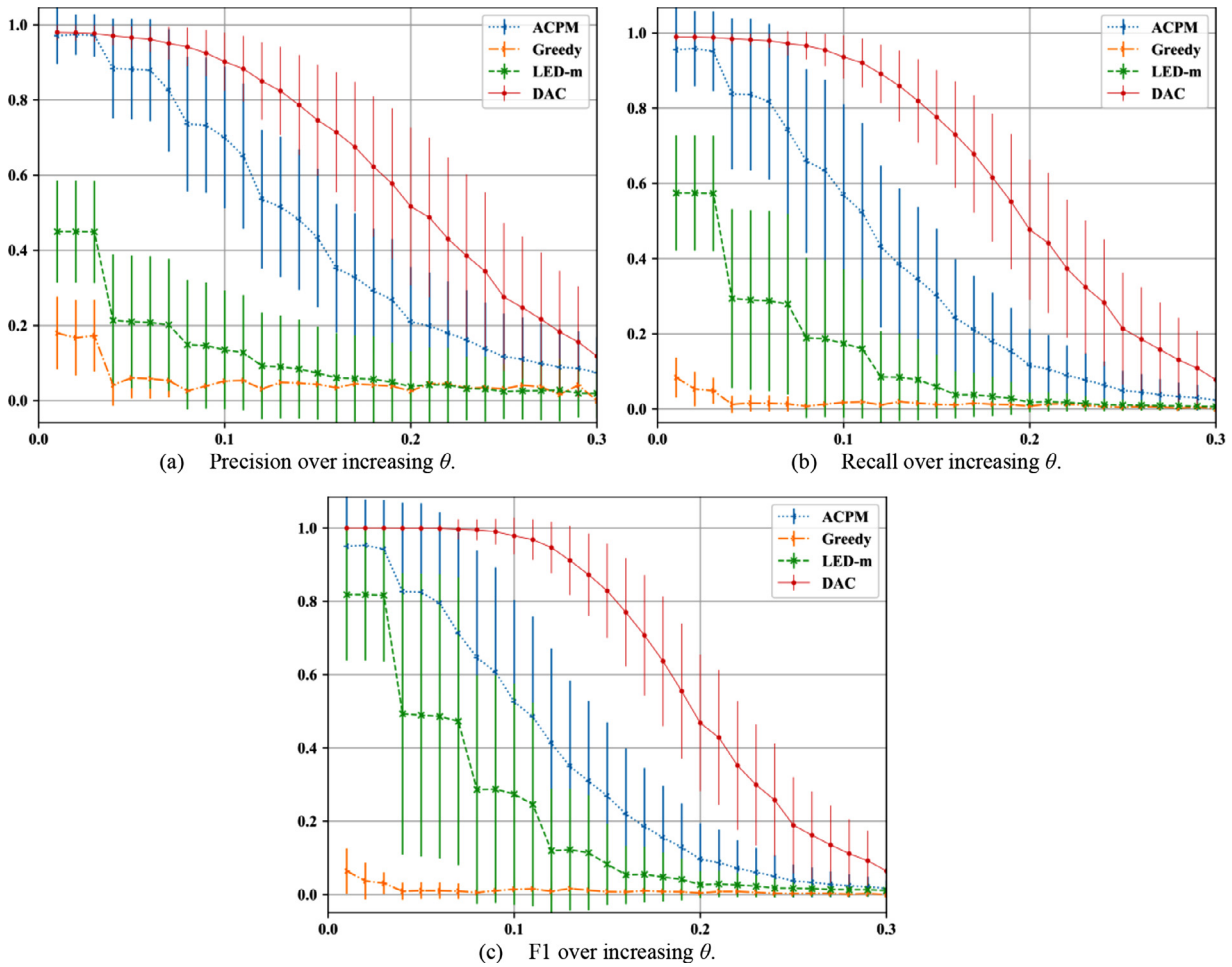

Fig. 2. Comparison between the existing algorithms and the proposed DAC algorithm with averaged prediction result for $y \sim y + 3$ over 103 datasets and 10 years. θ and (a) Precision, (b) Recall, and (c) F1 are respectively shown as the x and y axis.

diminished as predictions are made further in the future. The F1 value started with 0.5129 when the topics were detected in a given year ($d = 0$), which then showed diminishing values of 0.4909, 0.4810, and 0.4685 with $d = 1$, 2, and 3. Performance differences over d are universal in all datasets. A number of case studies showed that while the overall performance diminishes greatly with higher θ, positively matched clusters show better semantic explanations. *Lawlessness* is an emergent topic in the domain *development economics* in the year 2008, previously having been internally used at least 5 times. 25 *Ancestor* topics with the highest co-occurrence intensity in (11) were then compared against DAC clusters built by using data up to the year 2005, successfully performing a 3-year prediction with one cluster showing a Jaccard similarity of 0.5172. Comparing topics within two groups shows that the *ancestors* and *cluster* shared an idea of lawlessness being an outcome of geopolitical economics (*geography, politics, economics*) in deteriorating conditions (*terrorism, corruption, poverty*). *Cluster* members predicted topics such as *organized crime* and *public sector* would be relevant to a future emergent topic, while the actual *ancestors* instead had more demographic-specific topics such as *Islam, Somali,* and *colonialism*. Using the same experimental conditions, a topic *musical acoustics* within the *human-computer interaction* domain also showed interesting differences between the actual *ancestors* and the predicted *cluster* with θ=*0.4138*. Agreement on the topic's technical source is shown by topics shared by two groups (*artificial intelligence, multimedia, speech recognition, user interface*), while the prediction was made that the emergent topic would be more application-focused (*musical composition, pop music automaton, the internet* in a cluster). The *ancestors* were more technology-aware instead, having topics such as *auditory display, haptic technology, scalability, sonifications,* and *visual servoing*. Table 9 shows the full list of *ancestors* and *clusters* for the above cases.

Performance changes over time y with cluster similarity threshold values θ were analyzed as shown in Fig. 3 to reduce the number of outcome dimensions. Performance changes with θ are similar among variations showing drastically different outcomes per θ. F1 values over ten years with $\theta = [0.1, 0.2, 0.3]$ showed statistically significant difference over y with $p = 1.99\text{E-}7$ only when $\theta = 0.1$ and the differences became statistically obsolete with higher θ reaching $p = 0.026$, 0.837. 40.57% of the result had no positive matched clusters with $\theta = 0.4$ and therefore thresholds at or greater values were not considered in further analysis. This result shows that the

Table 6
Average F1 for four DAC variants ($\sigma_y(e)$ and $\sigma_y'(e)$ with α' = 90, 95) over all y and d, when θ'=0.

	σ_y(e)		σ_y'(e)	
θ	α'=90	α'=95	α'=90	α'=95
0.01	0.9895	0.9913	0.9854	0.9906
0.05	0.9834	0.9862	0.9773	0.9847
0.10	0.9426	0.9443	0.9402	0.9372
0.15	0.7927	0.7973	0.7901	0.7686
0.20	0.5044	0.5164	0.4930	0.4705
0.25	0.2345	0.2495	0.2264	0.2122
0.30	0.0876	0.0965	0.0850	0.0778
0.35	0.0314	0.0339	0.0318	0.0264
0.40	0.0111	0.0112	0.0114	0.0093
0.45	0.0055	0.0052	0.0052	0.0047
0.50	0.0035	0.0031	0.0028	0.0026

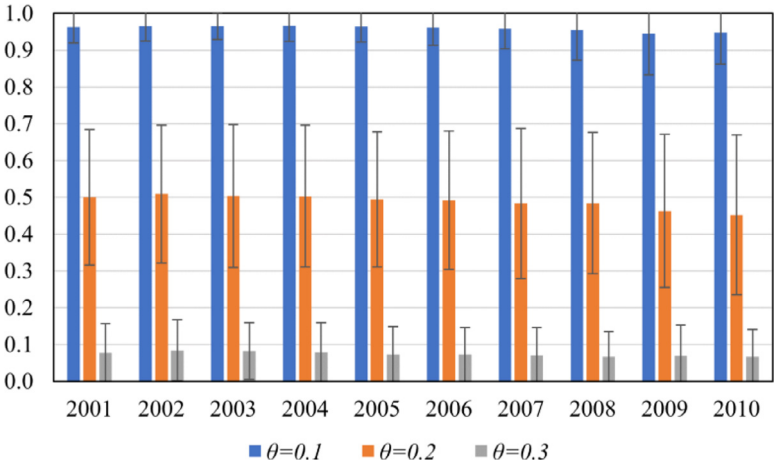

Fig. 3. Average F1 for DAC variants with standard deviation bars over ten years with three sets of θ ($\sigma_y(e)$, α' = 0.95).

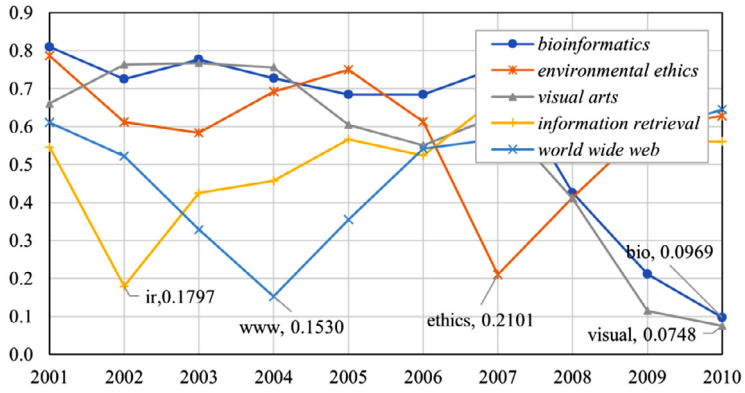

Fig. 4. Five outlier domain-specific datasets showing significant F1 changes over the years.

effect of a specific year is mostly not relevant to the performance over most of the datasets therefore results over y are averaged for further analysis.

While most datasets showed rather consistent performances, there are five outliers out of 103 domain-specific datasets showing significant time-sensitive performance changes. Fig. 4 shows F1 changes over years with θ = 0.2 for those outliers, where three of them exhibit a valley-shaped evolution. *Information retrieval, World Wide Web*, and *environmental ethics* showed a rapid drop in their F1 values in 1∼2 years, only to return to their former levels at similar speeds. This indicates that these three datasets experienced

Table 7
Average Precision, Recall, and F1 for DAC($\sigma_y(e)$, α'=95, θ=0.2) over 103 datasets over ten years.

d	θ'	Precision	Recall	F1
0	**0**	0.4905	0.4822	0.4705
	1st	0.6186	0.4764	0.5228
	2nd	0.7202	0.4545	0.5434
	3rd	0.8008	0.3953	0.5152
1	**0**	0.4767	0.4587	0.4517
	1st	0.5986	0.4534	0.5006
	2nd	0.6973	0.4328	0.5202
	3rd	0.7727	0.3757	0.4915
2	**0**	0.4679	0.4470	0.4418
	1st	0.5901	0.4426	0.4908
	2nd	0.6898	0.4233	0.5110
	3rd	0.7651	0.3659	0.4807
3	**0**	0.4587	0.4338	0.4312
	1st	0.5787	0.4291	0.4783
	2nd	0.6791	0.4098	0.4977
	3rd	0.7550	0.3535	0.4672

a discrete change in their topic evolution in different years, which could not be captured with continuous topic evolution tracking approaches. *Bioinformatics* and *visual arts* showed lowering F1, and the authors assume that the two datasets also had fissures in their topic evolution trend in or around 2010 and that DAC's performances will re-grow after the newer continuous trends are captured.

Comparison between the unfiltered outcomes and outcomes filtered by *predicted similarity* values indicates that removing less-performing clusters benefits overall performances. Clusters with maximum Jaccard similarity to ancestors in the lowest quartile are removed when θ' = 1st, which resulted in the F1 increasing by 0.0493 on average as shown in Table 7. The precision constantly increased with larger threshold quartiles up to average precision of 0.7734 as clusters with lower predicted similarities are removed from the process. A lower number of higher quality clusters resulted in fewer false positives; higher θ' is also expected to cause a higher ratio of false negatives, resulting in lower recall values. Using only the top quartile with θ' = 3rd resulted in slight performance drops in terms of F1, as a steady decline in the recall values begins to outweigh the increased precision when only a quarter of the detected clusters were compared against ancestors.

Conclusion

Topic evolution captures changes in topic semantics as well as their relationships over time by analyzing various data sources, from unstructured documents to metadata in bibliographical datasets. Traditional topic models measure topical similarity with contents; therefore predicting emergent topics with no known content is not well studied. Existing research primarily focused on detection instead, capturing the emergence of new topics as they appear in the timelined dataset. The descendant-aware clustering algorithm is proposed to tackle the problem of topic emergence prediction before topics materialize. This is done by converting the problem into *predicting topic groups that will have an emergent topic as their common descendant*, employing a network-based approach. In this problem, an emergent topic is predicted when a cluster positively matches one of the emergent topics' ancestor groups in the future.

The DAC algorithm is proposed to focus on identifying possible common ancestors of new descendant topics within a given topic network. Topical qualities such as popularity, importance, and maturity are taken into consideration with various parameters such as topic co-occurrence frequencies and pace of collaborations. A series of filtrations and expansions are done based on structural similarities, then regularized by edge filtration and merging heavily overlapping clusters. The algorithm process is designed to operate on large dense networks, extracting non-exhaustive and overlapping topic clusters which are regarded as candidate ancestor groups for topics that will emerge in the future.

Linear regression analyses on 103 domain-specific datasets were first done to measure the prediction capabilities, comparing observed cluster similarities versus the predicted similarity values based on the results in previous years. Results showed that up to 59% of the cluster similarities can be explained by linear regression models, showing the average mean squared error of 0.002 for 103 datasets. The generalizability of the DAC algorithm is shown over different research domains each with a different research focus and interests. It is also worth noting that the DAC is not a specialized algorithm for a specific dataset such as the MAG's FoS used in the experiment. Co-occurrence networks as input data are loosely defined and other forms of topics such as word embeddings can be used to build the input network. DAC algorithm can also be applied to non-topical networks in order to perform a *node* prediction as opposed to topic emergence prediction.

Multiple variations of DAC showed superior performance compared to three existing algorithms which undergo the same cluster regularization process as DAC. Three algorithms represented exhaustive clustering, overlapping clustering, and clustering focused on topic emergence prediction showing incremental performances as their goal is closer to topic evolution. DAC's clusters resulted in positive matches with larger cluster similarity threshold values and showed a higher ratio of positive matches at each threshold while exhibiting slow performance descent with increasing thresholds. While producing the best outcome, DAC was also shown to be fast and memory-efficient compared to existing algorithms, capable of working on networks with more than 100 million edges.

Positive matches were detected made up to three years in the future, predicting the likelihood of a cluster having an emergent topic as its descendant. Predictions made for three years in the future still showed an average F1 of 0.4685, indicating that structural patterns related to emergent topics are steady enough in most research domains to perform multi-year predictions. When their topic evolution trend abruptly altered, DAC was able to capture the new patterns after one to three years; DAC's performance rebounded after experiencing temporary drops in prediction performances.

Modifying the DAC algorithm will be done in future work. Different regression methods with more training features will be evaluated to enhance the *cluster qualities*, building an accurate model to rank the DAC clusters in the likely order of being the ancestors for emergent topics. Clusters with varying common structures will be extracted with dynamic similarity modifications in the edge filtering process to capture sudden changes in topic evolution patterns. The *connected component expansion* process will consider the edge frequency as well, allowing clusters to utilize more topic co-occurrence frequency information. Improved performances with predicted similarity quartile parameters will be incorporated into the clustering algorithm as well, with structural similarity rank and cluster size limitation parameters. Modularized evaluations would be done along with the algorithm improvements, analyzing the effects of each step and removing irrelevant steps to further reduce the time and resources required .

CRediT authorship contribution statement

Sukhwan Jung: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Aviv Segev:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Appendix

Table 8
List of datasets with their core topics and the number of membership topics and topic co-occurrence frequencies, followed by the number of emergent topics and the average size of their neighbors in the years 2005 and 2010.

			2005		2010	
Core Topic	# of Topics	# of Co-occurrences	Count	Avg. size	Count	Avg. size
actuarial science	35,167	2397,765	212	24.90	239	24.78
advertising	54,572	4423,987	304	24.73	472	24.79
aesthetics	30,517	2271,009	137	25.00	214	24.92
agroforestry	43,226	2998,769	139	24.71	167	24.93
algebra	31,280	2817,917	141	24.99	157	24.96
ancient history	38,793	3633,754	111	24.33	221	24.28
animal science	58,735	2778,387	145	25.00	193	24.78
anthropology	35,457	3003,186	121	24.74	167	24.22
applied mathematics	41,099	2588,396	146	24.72	164	24.98
archaeology	62,556	3790,023	126	24.61	236	23.89
astronomy	41,224	5076,883	73	24.53	115	24.72
astrophysics	31,260	5212,637	77	24.84	91	24.67
atmospheric sciences	30,646	2695,013	78	25.00	102	24.96
bioinformatics	92,142	3767,221	381	24.99	496	24.98
biomedical engineering	72,127	3472,455	268	24.72	390	24.86
biophysics	53,056	2632,646	127	24.99	129	24.84
cartography	95,012	5482,134	288	24.41	497	24.26
civil engineering	40,542	2382,278	213	24.93	263	24.62
classical mechanics	52,404	6625,401	185	24.90	177	24.98
classics	31,963	3414,987	73	24.47	159	24.06
climatology	35,705	2930,897	117	24.99	158	24.77
clinical psychology	52,174	6484,065	202	24.91	255	25.00
cognitive psychology	45,507	3434,011	182	25.00	430	24.52
combinatorics	47,835	4783,880	193	24.99	218	24.72
computational chemistry	35,357	3008,257	82	25.00	54	25.00
computer graphics images	46,601	3763,998	238	24.73	190	24.45
crystallography	59,842	6413,452	208	24.44	181	24.73
demography	57,107	3239,925	149	25.00	215	24.84
development economics	33,024	3722,942	181	24.99	266	24.90
discrete mathematics	55,017	6078,114	269	24.92	211	24.90
distributed computing	57,437	5675,803	510	25.00	318	24.83
econometrics	50,243	4104,364	223	24.88	277	24.98
economic growth	37,273	5365,714	267	24.91	411	24.95
economy	42,093	4151,910	283	24.96	305	24.98
engineering drawing	62,324	5648,804	261	24.72	274	24.52

(*continued on next page*)

Table 8 (*continued*)

Core Topic	# of Topics	# of Co-occurrences	2005		2010	
			Count	Avg. size	Count	Avg. size
engineering ethics	34,499	2385,121	144	25.00	243	24.70
engineering management	37,691	2995,708	219	24.53	279	24.70
environmental chemistry	57,123	4468,251	135	24.87	202	25.00
environmental engineering	61,037	5374,966	220	25.00	421	24.93
environmental ethics	39,355	2353,948	108	24.79	209	24.50
environmental health	55,325	3851,309	190	25.00	257	24.99
environmental planning	30,892	2855,897	182	24.83	266	24.78
environmental protection	48,124	2256,015	162	25.00	230	24.85
epistemology	44,110	4851,536	242	24.94	304	24.89
ethnology	43,498	3170,213	75	24.08	129	23.97
fishery	71,576	3406,750	163	24.92	216	24.95
forensic engineering	51,718	2445,868	168	24.95	181	24.72
forestry	68,974	2661,963	126	24.46	236	24.44
gender studies	33,694	4782,842	224	24.76	254	24.79
geochemistry	26,396	3320,336	68	24.43	81	25.00
geometry	60,296	4477,009	153	24.56	200	24.79
geomorphology	42,353	3458,390	81	24.70	157	24.36
geotechnical engineering	45,438	5840,508	241	24.99	248	25.00
gerontology	51,878	3944,056	174	24.97	257	24.98
human computer interaction	46,325	2900,975	238	25.00	244	24.83
hydrology	48,765	4956,760	182	25.00	212	24.92
information retrieval	48,977	3049,638	238	24.84	216	24.59
library science	59,260	5675,004	172	24.51	468	23.06
linguistics	40,193	5463,720	184	24.87	255	24.75
management	48,477	3513,747	184	24.90	362	24.72
marine engineering	36,650	3895,780	115	24.88	274	24.92
marketing	51,027	6416,905	365	24.97	557	24.96
media studies	39,288	4457,398	207	24.89	298	24.65
medical education	40,947	4963,492	162	24.75	292	24.87
medicinal chemistry	35,003	3013,725	57	24.12	43	24.16
meteorology	45,651	3541,140	142	24.86	204	24.65
microeconomics	33,566	2458,070	178	25.00	147	24.93
mineralogy	58,769	4575,444	144	24.52	184	24.86
neuroscience	68,721	5933,957	184	24.87	301	24.92
nuclear chemistry	61,373	5633,411	189	24.61	256	24.70
nuclear engineering	27,728	2310,540	60	25.00	99	24.73
nuclear magnetic resonance	57,969	5440,971	133	25.00	115	25.00
nuclear physics	22,080	3790,254	58	24.88	40	24.58
oceanography	46,845	3270,096	105	24.50	122	24.07
operations management	63,496	4105,735	346	25.00	467	24.96
operations research	65,399	3076,770	267	25.00	341	24.74
paleontology	63,417	2710,317	95	25.00	107	25.00
particle physics	17,917	3467,573	52	24.90	44	24.93
petroleum engineering	38,434	4403,087	124	24.56	229	24.93
political economy	31,155	4707,239	167	24.26	205	24.72
psychoanalysis	32,207	2246,569	78	24.59	126	24.37
psychotherapist	32,775	2516,913	87	24.98	122	24.78
public administration	42,084	6337,588	243	24.58	374	24.53
pulp and paper industry	39,493	3124,851	126	24.63	251	24.89
pure mathematics	27,364	2928,460	89	24.94	139	24.94
quantum electrodynamics	22,296	3440,328	80	25.00	38	25.00
quantum mechanics	43,893	5610,778	190	25.00	130	25.00
religious studies	22,446	3146,961	96	23.91	126	24.09
remote sensing	47,035	3932,035	207	24.94	236	25.00
simulation	89,549	6456,106	552	24.95	730	24.97
speech recognition	47,573	3300,950	213	24.84	248	24.67
statistics	77,319	4999,099	230	24.92	325	24.76
systems engineering	45,938	2904,975	215	24.98	216	24.93
theology	34,429	4450,785	103	23.11	193	23.66
theoretical computer science	53,430	3561,597	305	25.00	258	25.00
thermodynamics	45,324	5110,995	131	24.88	150	24.98
topology	53,617	4406,534	232	25.00	275	24.99
traditional medicine	85,056	4430,631	389	24.78	552	24.61
transport engineering	37,133	2698,597	150	24.84	249	24.82
veterinary medicine	70,733	2543,243	194	25.00	281	24.78
visual arts	39,341	3605,958	155	24.57	307	24.13
world wide web	62,275	5814,558	339	24.74	397	24.48
zoology	150,091	3958,030	149	24.86	267	24.53

Table 9
Examples of DAC clusters with positive matching to the actual ancestor groups exceeding θ=0.5. Data in 2005 is used to predict emergent topics in 2008 ($y = 2005$ and $d = 3$).

(a) Topic Lawlessness in domain Development Economics.	
Ancestors	Clusters
Colonialism	Corruption
Corruption	Democracy
Democracy	Development economics
Development economics	Economic growth
Economics	Economics
Geography	Geography
Globalization	Globalization
Government	Government
Insurgency	Law
International community	Organised crime
Islam	Political corruption
Law	Political economy
Legitimacy	Political science
Political economy	Politics
Political science	Poverty
Politics	Public sector
Population	Rule of law
Poverty	Sociology
Rule of law	Terrorism
Scholarship	
Sociology	
Somali	
Spanish Civil War	
Terrorism	
Wage	
(b) Topic *Musical acoustics* in domain *Human–computer interaction.*	
Ancestors	Clusters
Artificial intelligence	Artificial intelligence
Auditory display	Computer science
Computer music	Distributed computing
Computer science	Engineering
Digital audio	Gesture
Distributed computing	Human–computer interaction
Gesture	Input device
Haptic technology	Multimedia
Human–computer interaction	Musical
Input device	Musical composition
Multimedia	New Interfaces for Musical Expression
Musical	Pop music automation
Musical instrument	Software
New Interfaces for Musical Expression	Speech recognition
Scalability	The Internet
Scientific method	User interface
Simulation	
Software	
Software development	
Sonification	
Speech recognition	
Telecommunications	
User interface	
Virtual reality	
Visual servoing	

References

Allan, J., Carbonell, J.G., .Doddington, G., Yamron, J., & Yang, Y. (1998).Topic Detection and Tracking Pilot Study Final Report, https://doi.org/10.1184/R1/6626252.v1

Balili, C., Lee, U., Segev, A., Kim, J., & Ko, M. (2020). TermBall: Tracking and predicting evolution types of research topics by using knowledge structures in scholarly big data. *IEEE access : practical innovations, open solutions, 8*, 108514–108529. 10.1109/ACCESS.2020.3000948.

Battistella, C. (2014). The organisation of corporate foresight: A multiple case study in the telecommunication industry. *Technological Forecasting and Social Change, 87*, 60–79. 10.1016/j.techfore.2013.10.022.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2022).A Neural Probabilistic Language Model, (n.d.) 19.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). New York, NY, USA: ACM. 10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res., 3*, 993–1022.

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics, 115*, 35–49. 10.1007/s11192-018-2654-5.
Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*, 1175–1189. 10.1016/j.joi.2017.10.003.
Chen, C., & CiteSpace, I. I. (2006). Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*, 359–377. 10.1002/asi.20317.
Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E., 70*, Article 066111. 10.1103/PhysRevE.70.066111.
Cordasco, G., & Gargano, L. (2010). Community detection via semi-synchronous label propagation algorithms. In *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)* (pp. 1–8). 10.1109/BASNA.2010.5730298.
Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*, 498–514. 10.1016/j.joi.2011.02.006.
Fiscus, J. G., & Doddington, G. R. (2002). Topic Detection and Tracking Evaluation Overview. In *Topic detection and tracking* (pp. 17–31). Boston, MA: Springer. 10.1007/978-1-4615-0933-2_2.
Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science, 47*, 117–132. 10.1287/mnsc.47.1.117.10671.
Fung, G. (2022).A Comprehensive Overview of Basic Clustering Algorithms, (n.d.) 37.
Furukawa, T., Mori, K., Arino, K., Hayashi, K., & Shirakawa, N. (2015). Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions. *Technological Forecasting and Social Change, 91*, 280–294. 10.1016/j.techfore.2014.03.013.
Garner, J., Carley, S., Porter, A. L., & Newman, N. C. (2017). Technological emergence indicators using emergence scoring. In *2017 Portland international conference on management of engineering and technology (PICMET)* (pp. 1–12). 10.23919/PICMET.2017.8125288.
Gohr, A., Hinneburg, A., Schult, R., & Spiliopoulou, M. (2009). Topic evolution in a stream of documents. In *Proceedings of the 2009 SIAM international conference on data mining, society for industrial and applied mathematics* (pp. 859–870). 10.1137/1.9781611972795.74.
Henriques, R., Bacao, F., & Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems, 36*, 218–232. 10.1016/j.compenvurbsys.2011.11.003.
Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with microsoft academic. *Scientometrics, 111*, 371–378. 10.1007/s11192-017-2247-8.
Jung, S., Datta, R., & Segev, A. (2020). Identification and prediction of emerging topics through their relationships to existing topics. In *2020 IEEE International conference on big data (Big Data)* (pp. 5078–5087). 10.1109/BigData50022.2020.9378277.
Jung, S., Lai, T. M., & Segev, A. (2016). Analyzing future nodes in a knowledge network. In *2016 IEEE International congress on big data (BigData Congress)* (pp. 357–360). 10.1109/BigDataCongress.2016.57.
Jung, S., & Segev, A. (2013). 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* (pp. 15–22). New York, NY, USA: ACM. 10.1145/2513549.2513553.
Jung, S., & Segev, A. (2014). Analyzing future communities in growing citation networks. *Knowledge-Based Systems, 69*, 34–44. 10.1016/j.knosys.2014.04.036.
Jung, S., & Segev, A. (2021). Analyzing the generalizability of the network-based topic emergence identification method (Accepted), special issue on deep learning and knowledge graphs. *Semantic Web Journal*.
Jung, S., & Yoon, W. C. (2020). An alternative topic model based on common interest authors for topic evolution analysis. *Journal of Informetrics, 14*, Article 101040. 10.1016/j.joi.2020.101040.
Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*, 373–397. 10.1023/A:1024940629314.
Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*, 1464–1480. 10.1109/5.58325.
Kulis, B., & Jordan, M.I. (.2012).Revisiting k-means: New Algorithms via Bayesian Nonparametrics, ArXiv:1111.0352 [Cs, Stat]. http://arxiv.org/abs/1111.0352 (accessed January 30, 2022).
Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E., 78*, Article 046110. 10.1103/PhysRevE.78.046110.
Levy, O., & Goldberg, Y. (2022).Neural Word Embedding as Implicit Matrix Factorization, (n.d.) 9.
Li, M., & Chu, Y. (2017). Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *Journal of Information Science, 43*, 725–741. 10.1177/0165551516661914.
Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*, 451–461. 10.1016/S0031-3203(02)00060-2.
Ma, T., Wang, Y., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., et al., (2016). LED: A fast overlapping communities detection algorithm based on structural clustering. *Neurocomputing, 207*, 488–500. 10.1016/j.neucom.2016.05.020.
Mei, Q., & Zhai, C. (2005). 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* (pp. 198–207). New York, NY, USA: ACM. 10.1145/1081870.1081895.
Newman, N. C., Porter, A. L., Newman, D., Trumbach, C. C., & Bolan, S. D. (2014). Comparing methods to extract technical content for technological intelligence. *Journal of Engineering and Technology Management, 32*, 97–109. 10.1016/j.jengtecman.2013.09.001.
Osborne, F., Mannocci, A., & Motta, E. (2017). Forecasting the spreading of technologies in research communities. In *Proceedings of the knowledge capture conference* (pp. 1–8). New York, NY, USA: Association for Computing Machinery. 10.1145/3148011.3148030.
Ozmutlu, H. C., & Çavdur, F. (2005). Application of automatic topic identification on Excite Web search engine data logs. *Information Processing & Management, 41*, 1243–1262. 10.1016/j.ipm.2004.04.018.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. Stanford InfoLab.
Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*, 814–818. 10.1038/nature03607.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change, 49*, 237–255. 10.1016/0040-1625(95)00022-3.
Rasmussen, C. E. (2000). *The infinite gaussian mixture model*. MIT Press n.d..
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*, 1827–1843. 10.1016/j.respol.2015.06.006.
Salatino, A. A., Osborne, F., & Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Comput. Sci., 3*, e119. 10.7717/peerj-cs.119.
Salatino, A. A., Osborne, F., & Motta, E. (2018). AUGUR: Forecasting the emergence of new research topics. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries* (pp. 303–312). New York, NY, USA: Association for Computing Machinery. 10.1145/3197026.3197052.
Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019). The CSO classifier: ontology-driven detection of research topics in scholarly articles. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital libraries for open knowledge* (pp. 296–311). Cham: Springer International Publishing. 10.1007/978-3-030-30760-8_26.
Schumpeter, J. A. (1939). *Business cycles*. New York: Mcgraw-hill.
Shen, Z., Ma, H., & Wang, K. (2018).A Web-scale system for scientific knowledge exploration, ArXiv:1805.12216 [Cs]. http://arxiv.org/abs/1805.12216 (accessed June 23, 2020).
Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., (Paul) Hsu, B.-. J., et al., (2015). An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246). Florence, Italy: Association for Computing Machinery. 10.1145/2740908.2742839.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emergent topics in science and technology. *Research Policy, 43*, 1450–1467. 10.1016/j.respol.2014.02.005.
Steyvers, M., & Griffiths, T. (2007). *Probabilistic topic models, in: Handbook of latent semantic analysis* (pp. 427–448). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Wang, K., Shen, Z., Huang, C., Wu, C.-. H., Eide, D., Dong, Y., & Rogahn, R. (2019). A review of microsoft academic services for science of science studies. *Front. Big Data., 2*. 10.3389/fdata.2019.00045.

Yamaguchi, Y., & Hayashi, K. (2017). When does label propagation fail? A view from a network generative model. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, international joint conferences on artificial intelligence organization* (pp. 3224–3230). 10.24963/ijcai.2017/450.

Zhang, J., Ghahramani, Z., & Yang, Y. (2004). A probabilistic model for online document clustering with application to novelty detection. In *Proceedings of the 17th international conference on neural information processing systems* (pp. 1617–1624). Vancouver, British Columbia, Canada: MIT Press.

Zhang, Y., Wu, M., Miao, W., Huang, L., & Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Journal of Informetrics, 15*, Article 101202. 10.1016/j.joi.2021.101202.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*, 1925–1939. 10.1002/asi.23814.

Zhou, X., Huang, L., Zhang, Y., & Yu, M. (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. *Scientometrics, 121*, 699–737. 10.1007/s11192-019-03218-5.