Journal of Informetrics 14 (2020) 101057

Contents lists available at ScienceDirect

Journal of Informetrics

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

<![CDATA[Regular article]]>

Mining Temporal Evolution of Knowledge Graphs and Genealogical Features for Literature-based Discovery Prediction

Nazim Choudhury[a], Fahim Faisal[b,*], Matloob Khushi[c]

[a] Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA
[b] Dept. of Computer Science, George Mason University, Fairfax, VA 22030, USA
[c] School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 14 November 2019
Received in revised form 16 May 2020
Accepted 22 May 2020

Keywords:
Literature-based Knowledge Discovery
Dynamic Supervised Link Prediction
Keyword Co-occurrence Network (KCN)
Genealogical Community
Weighted Temporal Citation

ABSTRACT

Literature-based discovery process identifies the important but implicit relations among information embedded in published literature. Existing techniques from Information Retrieval (IR) and Natural Language Processing (NLP) attempt to identify the hidden or unpublished connections between information concepts within published literature, however, these techniques overlooked the concept of predicting the future and emerging relations among scientific knowledge components such as author selected keywords encapsulated within the literature. Keyword Co-occurrence Network (KCN), built upon author selected keywords, is considered as a knowledge graph that focuses both on these knowledge components and knowledge structure of a scientific domain by examining the relationships between knowledge entities. Using data from two multidisciplinary research domains other than the bio-medical domain, and capitalizing on bibliometrics, the dynamicity of temporal KCNs, and a recurrent neural network, this study develops some novel features supportive for the prediction of the future literature-based discoveries - the emerging connections (co-appearances in the same article) among keywords. Temporal importance extracted from both bipartite and unipartite networks, communities defined by genealogical relations, and the relative importance of temporal citation counts were used in the feature construction process. Both node and edge-level features were input into a recurrent neural network to forecast the feature values and predict the future relations between different scientific concepts/topics represented by the author selected keywords. High performance rates, compared both against contemporary heterogeneous network-based method and preferential attachment process, suggest that these features complement both the prediction of future literature-based discoveries and emerging trend analysis.

1. Introduction

Statistical bibliography or bibliometrics (Pritchard, 1969) has been supporting researchers to address the challenges related to the rapid growth of scholarly publications and scientific knowledge. Two network-based methods (i.e., co-citation and keyword co-occurrences network) (Kessler, 1963) in bibliometrics enabled researchers to

* Corresponding author.
E-mail addresses: nachoudhury@usf.edu (N. Choudhury), ffaisal@gmu.edu (F. Faisal), matloob.khushi@sydney.edu.au (M. Khushi).

https://doi.org/10.1016/j.joi.2020.101057
1751-1577/

explore the structure of scientific and technical knowledge. While the co-citation focuses on the structure of scientific communication by analyzing citation links, keyword co-occurrence network (KCN) or co-word network, focuses on both knowledge components and knowledge structure by examining co-appearances of keywords found in the scientific literature. Therefore, it is also known as the knowledge graph (Popping, 2003). Associated as metadata information within the scholarly publications, the author selected keywords are considered as the carriers of knowledge units, or knowledge entities (Su & Lee, 2010). These hand-picked signal words/terms/themes/conceptual keywords not only denote the authors' understandings of their work, the thematic context, and summarized topic of their research but also facilitate knowledge discovery (Song, Han, Kim, Ding, & Chambers, 2013). Further, these topical concepts are also used for indexing purpose in digital libraries. The co-appearance of two author selected keywords in an article defines a certain intrinsic relationship between two topics whereas multiple such instances denote the strength of their relationships (Yang, Wu, & Cui, 2011). Representing such co-occurring relationships among knowledge entities, KCN bears both theoretical and practical implications including literature-based discovery (Choudhury & Uddin, 2016).

Literature-based discovery (Kostoff, 2005), widely spelled as Literature Based Discovery (LBD), is the process that seeks to discover new knowledge (innovation) from the information embedded in published literature. In other words, the task of the LBD system is to exploit already known scientific knowledge to derive meaningful connections between scientific topics those are implicit and yet to be disclosed (Smalheiser, 2017). Hence, it is often also referred to as hypothesis generation. In both automated and semi-automated ways, it identifies the hidden (implicit) and important connections among knowledge components. The rapid growth of scholarly publications and associated scientific knowledge prompted scholars to restrict themselves within their narrow specialties (i.e., fragments within a broad domain) and cite only from the related articles (Ganiz, Pottenger, & Janneck, 2005). Consequently, useful connections between fragmented information remain unnoticed due to the lack of awareness of work from mutually exclusive fragments. LBD addresses this problem of knowledge overspecialization and strives to identify the implicit and novel connections from the concepts explicitly published in the scientific literature. Preiss et al. (Preiss, Stevenson, & Gaizauskas, 2015) reported different applications of this approach including identification of treatments for diseases, drug re-purposing, disease candidate gene discovery, and drug side effect prediction. Similarly, Henry et al. (Henry & McInnes, 2017) reported application areas outside the biomedical domain such as efficient water purification systems, development acceleration of developing countries, potential bio-warfare agents categorization, climate change studies, and identification of promising research collaborations.

Most methodologies addressing LBD are based on Swanson's 'ABC' model (Swanson, 1986) and predominantly rely on text analysis (Swanson & Smalheiser, 1997). Further, most studies in LBD used MEDLINE as the literature database. Sebastian et al. (2017a, Sebastian, Siew, and Orimaye (2017a,b) divided existing techniques of literature-based knowledge discovery from scientific literature into two categories: (i) traditional, and (ii) emerging approaches. The traditional approach dominates the current research landscape. These are mainly knowledge-based and comprised of lexical statistics or graph-theoretic methods that require domain knowledge. On the other hand, emerging approaches prompt new trends and unprecedented paradigm shifts in the knowledge discovery process. These methods consider the occurrence and co-occurrence frequencies of keywords, probability distributions, association rule mining, graph data structure, temporal features, relational attributes, and supervised classification approach. A literature review on these methods is presented in the Related Works section.

Being a knowledge network representation technique, KCN inherently fits best in the LBD process. However, despite their success in mapping scientific knowledge structure, the evolutionary dynamics of KCN, and metadata information (e.g., authors, citations) associated with scholarly articles (Ding et al., 2013) are underutilized in the existing LBD techniques. Temporal changes of metadata information, and communities of keywords extracted from dynamic KCN are non-trivial factors than mere frequencies or connectivity information in predicting the underlying implicit and complex relations between keywords. Most importantly, existing methods in LBD concentrates on identifying the implicit or undisclosed relations between keywords in the published literature instead of predicting their future relations those are yet to emerge. This can facilitate emerging trend prediction which can be supportive for both researchers and science policymakers. To this end, by formulating the LBD prediction problem as a dynamic supervised link prediction problem, we propose a recurrent neural network (RNN)-based method to predict the emerging associations (i.e., co-occurrences) between author defined keywords in the scientific literature. This prediction process incorporates features constructed by leveraging the temporal dynamics of KCNs, evolutionary metadata information (authors, citation) and genealogical communities of keywords. Due to the temporal nature of the constructed features and binary supervised classification task, we considered the Long Short Term Memory (LSTM) - an artificial RNN architecture, suitable for sequence classification.

The research objective of this study is to ***predict the emerging LBD and thus the co-evolution of knowledge entities (author selected keywords) found in the scientific literature***. This will facilitate the prediction of emerging scientific hypotheses. Since the prediction problem is framed as a supervised learning process, it requires the development of representative features to describe and successfully classify instances (non-connected keyword pairs). Therefore, the scientific contribution of this study is the ***development of some novel features by leveraging the temporal dynamics of KCNs, metadata information (i.e., authors, articles, citations), and finally, communities defined via deriving ancestral lineages of keywords extracted from the dynamic KCNs***. The prediction performances of the proposed features, applied over scholarly

datasets on two topics [1] (i.e., ***Obesity, Sleep Apnea***), including comparisons against a contemporary method based on heterogeneous bibliographic information network (HBIN) and preferential attachment, are also reported. These generic features are domain independent and can be applied to any multidisciplinary scientific disciplines to predict the emerging trends and scientific hypotheses.

Related Works

Literature-based discovery (LBD) identifies the implicit relations from the explicit information. It is principally used in biomedical research where running experiments is expensive. This knowledge discovery process broadly encompasses lexical and semantic text analyses of articles found in the MEDLINE database. Most of the LBD techniques employed frequency-based approaches. The underlying assumption is that discoveries are likely to arise from logical connections among source, intermediate, and target concepts (keywords) based on either their frequent/infrequent (co)occurrences in the literature, or common/rare connections to a knowledge base (Cameron et al., 2015). Gordon and Dumais (Gordon & Dumais, 1998) took the advantages of both co-occurrence frequencies and Latent Semantic Indexing (LSI) used in analyzing relationships between a set of documents. Hristovski et al (Hristovski, Stare, Peterlin, & Dzeroski, 2001) used association rule mining and Unified Medical Language Systems (UMLS) to discover the relationships between medical concepts. Yetisgen-Yildiz and Pratt developed (Yetisgen-Yildiz & Pratt, 2006) 'LitLinker' that incorporated knowledge-based methodologies with statistical approach considering the background distribution of keyword probabilities. With the help of Fuzzy Set Theory and probabilistic model of relationships, Wren et al. (Wren, Bekeredjian, Stewart, Shohet, & Garner, 2004) developed a computational method to identify large sets of relationships between unrelated items within scientific reports.

Ensemble-based approaches combined both statistical and temporal features to find the intermediate keywords connecting both source and target keywords. These approaches were explored to find meaningful links between two disparate sets of articles in MEDLINE (Smalheiser, Torvik, & Zhou, 2009). Relational techniques (Ahlers, Hristovski, Kilicoglu, & Rindflesch, 2007; Hu, Li, Yoo, Zhang, & Xu, 2005) used the explicit semantic relationships (i.e., predicates) between concepts where such predicates were typically obtained from structured background knowledge or known *a priori* via domain experts. Few approaches focused on graph data structures to discover knowledge from the literature by creating subgraphs based on the binary relationships between literary concepts. These relationships were also drawn from semantic predications which were extracted directly from assertions in MEDLINE literature. By weighting links using degree centrality, Wilkowski et al. (Wilkowski et al., 2011) developed such a graph-theoretic approach. Their approach used an iterative greedy strategy to create the 'best' subgraph with the help of semantic predications. Cameron et al. (Cameron et al., 2015) also used semantic predication graph and introduced a method to automatically find clusters of contextually similar paths in the graph. These clusters were then used to identify the latent associations between disjoint concepts in the literature to reconstruct eight scientific discoveries. Similar to most LBD methods, these graph-based data structures were also primarily constructed in the biomedical domain using predicative relations extracted from MEDLINE literature.

Contemporary approaches developed heterogeneous networks capable of encoding richer information and better semantics between various real world objects (Sun & Han, 2012). Known as Heterogeneous Bibliographic Information Network (HBIN), these networks represent a collection of scientific publications as a network of heterogeneous bibliographic objects (e.g., keywords, authors etc.). HBIN allowed various information to flow across different types of objects and links to capture the previously unknown associations between research articles. It also harnessed the meta-path features found in HBIN networks to discover the latent associations. Sebastian et al. (Sebastian, Siew, & Orimaye, 2017) used lexico-citation features of HBIN networks to predict the co-citation links between articles from previously disconnected research areas. Similarly, Ding et al. (Ding et al., 2013) used information from the literature in the form of an 'entitymetrics' graph. The objective was to model the latent relationships among biomedical entities (e.g. diseases, drugs) based on the existing citation relationships among the corresponding articles. Apart from predicting the implicit and/or hidden relationships between disjoint sets of articles, researchers (Liu, Yu, Guo, Sun, & Gao, 2014; Ren et al., 2014) also used HBIN to predict the citation count. HBIN-based LBD used simple statistics and does not require sophisticated and domain-specific NLP tools and ontologies. It also facilitated the utilization of metadata information in constructing features for prediction task. However, to predict future links between author selected keywords instead of links between articles, processing HBIN and calculating meta-path features will be computationally intensive.

One of the emerging approaches (Kastrin, Rindflesch, & Hristovski, 2016) conjectured the task of predicting implicit relationships as a classification task by leveraging the link prediction methodology of network science. It describes the associations between different concepts/keywords/topics using networks where the links represent their semantic or co-occurrence relationships. This approach was primarily used in predicting the implicit links within a co-occurrence network of Medical Subject Headings (MeSH) terms. Crichton et al. (Crichton, Guo, Pyysalo, & Korhonen, 2018) recently investigated how inputs from four node representation algorithms affect the performance of a neural link predictor on random- and time-sliced biomedical graphs (i.e., Drug-Target Interactions, Protein-Protein Interaction and LBD) containing information relevant to drugs, protein and literature. Katukuri et al. (Katukuri, Xie, Raghavan, & Gupta, 2012) used manually-created features in a

[1] datasets and codes are available in https://github.com/faisal-iut

supervised link prediction task on a large-scale biomedical network of concepts co-occurrences. The objective was to predict links those represented scientific hypotheses in a time-sliced corpus. The authors extracted relevant information from the biomedical corpus to generate a concept network and concept-author map. They also developed a set of heterogeneous features by exploiting random walk, network neighborhood and common authorship. Wang and Zeng (Wang & Zeng, 2013) used two-layer graphical model, called restricted Boltzmann machine, to perform similar link prediction task on Drug-Target Interactions (DTI) network. Their objective was to predict multiple types of DTIs, unknown drug-target relationships or drug's modes of action. However, Lu et al. (Lu, Guo, & Korhonen, 2017) pointed out the limitation of such machine learning models in DTI predictions despite their high performance. This included the absence of additional information about the characteristics of drugs, targets and DTIs, (e.g. chemical structure, genome sequence, binding types, causes of interactions, etc.). Therefore, the authors used topological similarity indices, such as Common Neighbours and Katz (Katz, 1953) index from complex network theory, to predict links in a DTI network. Our study is probably closer to this set of machine-learning based studies however it differs in different ways which is described later in this section.

In summary, existing research used combinations of lexical distributional statistics, graph-theoretic measures, heterogeneous network-based methods, and machine learning models. Predominantly capitulated on MEDLINE database, most existing methods require domain knowledge to interpret relations between knowledge entities (drug, target, protein, gene etc.). Graph-theoretic methods considered static networks which is unable to capture the temporal aspects of network evolution such as 'recency', genealogical traits (origins of relationships), and time-variant frequencies crucial for predicting the emerging associations in evolving networks. Despite the merits of existing approaches, there exists lack of generalized predictive features applicable to any research domain. These features should be free from any domain-specific knowledge-base for semantic-based predicative relationships or relax the requirements of domain experts for interpretation. It should accommodate metadata information which is commonly associated with any scientific literature like authors, keywords, affiliation etc, and integrate the temporal information since scientific knowledge structure is inherently dynamic. These facts are the motivation behind this study. Further, in most cases, these methods focused only on identifying implicit(hidden) relationships between scientific concepts which are yet to be recognized from their explicit relations instead of predicting the future relations those are yet to emerge. To this end, this study develops a recurrent neural network-based method to predict emerging LBD instead of identifying the hidden relations between keywords/concepts. It manifests the knowledge evolution as a dynamic process and leverages the evolutionary aspects of knowledge graphs, temporal recency and citation information to develop hand-crafted features suitable for the prediction task. Note that while co-occurrence associations between scientific concepts are commonly used in LBD (Kastrin, Rindflesch, & Hristovski, 2014) by the term emerging LBD, we denote the emergence of co-occurrence relationships between author selected keywords those are yet to be discovered.

2. Scholarly Datasets

We extracted our scholarly datasets on two topics and the source of these datasets is 'Scopus' -the largest abstract and citation database of peer-reviewed literature. The first search keyword was '**sleep apnea**', also spelled as sleep apnoea, and the second keyword was '**obesity**'. The first topic is related to the serious sleep disorder that occurs when a person's breathing is interrupted or a person experiences periods of shallow breathing during sleep. The second topic is related to overweight and represents a complex disorder involving the accumulation of excessive body fat to an extent such that it may have negative impact on health of individuals. Article metadata information including publication year, title, authors, affiliations, author selected keywords and citation count were extracted from the Scopus digital library by considering the following constraints: (i) article published in English journals, (ii) the search keywords are present in the article's titles and abstracts, and (iii) articles published within the duration 2007-2016. For the sake of brevity, G_s and G_o will be used to denote the dataset related to sleep apnea and obesity respectively for the rest of the article.

The author selected keywords are crucial to know about the thematic context, topics and related concepts of the corresponding scholarly articles. Recent advances in network science (Börner, Sanyal, & Vespignani, 2007) have prompted researchers to address the mapping and understanding of scientific knowledge evolution via different types of bibliometric networks. These networks consist of nodes representing different scholarly items like publications, journals, researchers, or keywords and edges indicating the relations between pairs of nodes. According to Van Eck and Waltman (Van Eck & Waltman, 2014), the most commonly studied types of relations are: (i) citation relations, (ii) keyword co-occurrence relations, and (iii) co-authorship relations. To predict the emerging relations between different scientific concepts/keywords, we took the advantage of temporal keyword co-occurrence relations and constructed dynamic KCNs in both G_s and G_o for each year. Therefore, as an integral part of dynamic KCN construction. we extracted author selected keywords appeared in more than one articles. Keywords failed to gain such minimum attention from the research community were considered to be irrelevant to the corresponding research and thus discarded. Keyword extraction phase was followed by the text processing phase that includes text cleaning and transformation. Authors used different spellings and acronyms to represent their keywords those represent same semantic meaning. Firstly, any unwanted space and other discriminators were eliminated from the keyword list with the help of NLTK text pre-processing tools (Bird, Klein, & Loper, 2009). This step also included lower-casing all keywords, removing singular-plural differences (e.g., epidemiological studies ↔ epidemiological study, dilator muscles ↔ dilator muscle) and lemmatization of some commonly used keywords (e.g., dreaming ↔ dream). Secondly, semantically related common keywords were uniformly presented by using one representative word for all (e.g., aspect ↔ feature). Most widely used abbreviated keywords were kept unchanged however others were changed to it's full forms (e.g., bmi ↔ body

Table 1
Basic statistics of two scholarly datasets

Dataset	# Articles	# Authors	# Keywords	Duration
G_s	29203	37268	10721	2007-2016
G_o	107745	210023	11643	2007-2016

Table 2
Yearly statistics of nodes (author selected keywords) and edges (co-appearances) in keyword co-occurrence networks (KCNs), and average citation count per article in G_s=Sleep Apnea and G_o=Obesity

		2007	2008	2009	2010	2011	2012	2013	2014
G_s	Nodes	388	459	470	496	536	584	634	583
	Edges	689	773	688	786	977	1054	1280	1144
	Citation	33.834	30.029	23.689	21.616	16.287	12.163	9.631	5.343
G_o	Nodes	2112	2355	2090	2299	2400	2634	2633	3216
	Edges	8286	8820	9478	10925	13096	14837	14380	16349
	Citation	25.645	20.188	15.907	10.901	6.727	3.471	0.882	6.952

mass index). There were some abbreviations with different full forms (e.g., egfr ↔ (estimated glomerular filtration rate, epidermal growth factor receptor)). In these cases, the content of the corresponding article was verified to identify the right full form. Besides, all numbers and their corresponding roman forms were used in number format. Basic dataset statistics including the number of these cleaned and transformed keywords are presented in Table 1. Due to the inherent bias associated with the search keywords, they were filtered out from the final pool of author selected keywords in both datasets. It is noteworthy that although the data collection period includes 2007, however, in the experiment, we considered 2008-2014 duration as our actual training period. This fact will be explained in the later section.

3. Background

3.1. Knowledge Evolution and Dynamic KCN

Currently, scientific knowledge creation is dynamic and interdisciplinary in nature where different avenues of research converge and new connections emerge among disjoint and existing areas of science (Pan, Sinha, Kaski, & Saramäki, 2012). This knowledge is generally incremental except few revolutionary and fundamental changes. The development and evolution of science and technology create new knowledge from the previously accumulated and ubiquitous information (Lee, Yoon, & Park, 2009). New hypotheses are being postulated by encompassing existing scientific concepts from multiple domains. Agusti Canals (Canals, 2005) pointed out that the diffusion of scientific knowledge can be mapped into a network structure where knowledge propagates via interactions among networked agents. A knowledge graph (network) is a specific kind of knowledge representation technique that uses a semantic network structure where nodes are keywords and edges represent their causal relationships (Popping, 2003). In addition to causal relationships, statistically significant and non-trivial co-occurrence patterns of keywords also represent their semantic affinity (Montemurro & Zanette, 2013) and relatedness (Schulz et al., 2014). To generate such knowledge graph in this study, using the extracted, cleaned, and processed keywords, as described above, dynamic Keyword Co-occurrence Networks (KCNs) were constructed for each year in 2008-2015.

KCN is defined as an undirected network $G(V, E)$ where V is the set of nodes representing the author selected keywords and E is the set of edges representing their co-appearances in the same article. Multiple such co-appearances define the edge weights which are ignored in this study. Instead of considering the crude co-occurrences between author selected keywords, this study considered a normalization index, defined by van Eck and Waltman, known as the ***assoication-strength*** (Eck & Waltman, 2009). By using this index, the similarity S_{ij} between two keywords i and j is defined as $S_{ij} = \frac{c_{ij}}{s_i \times s_j}$, where c_{ij} denotes the count of co-occurrences, s_i and s_j denote the total occurrences of keyword i and j respectively. After choosing a threshold value δ, we considered all co-occurrences where $S_{ij} \geq \delta$. In this study, we considered $\delta = \mu - \sigma$ where, μ σ are the mean and standard deviation of the association strength scores fro all co-occurrences. Since it is out of the scope of this study, we left for the future studies to explore methods to identify the optimal threshold value.

Dynamic KCNs are the temporal networks $G_t = (v_t, e_t)$ for time period $t = 1, 2, .., T$ where v_t is the set of nodes or keywords and $e_t \in E_t$ is the set of edges connecting the set of keywords $v_t \in V$ at time t. These edges at time t can be new or recurring. Table 2 provides statistics of the number of nodes (keywords) and edges (co-occurrences) per year in both datasets including average citation counts for articles containing the nodes as their author selected keywords. In Table 3, we present the basic statistics of the evolutionary patterns observed in dynamic KCN for both G_o and G_s. The dynamicity of keyword co-occurrences denotes that new research topics, hypotheses, or directions emerge over time through co-appearances of existing keywords. Three different scenarios can be observed in dynamic KCNs.

As observable in Table 3, firstly, new edges emerge each year when new keywords co-appear in articles. Secondly, new keywords form edge with old (existing) keywords (appeared in previous years). Finally, edges are formed in a year between two old keywords from the previous year(s), where these old keywords appeared in different articles but not co-appeared

Table 3
Evolutionary statistics of nodes (author selected keywords) and edges (co-occurrences in the same article) for two datasets in this study. V_t = keywords in year t, V_n = new keywords each year, V_o = old keywords from the previous year(s), E = edges, $E_{o\leftrightarrow o}$ = recurring edges between $v \in V_o$, $E'_{o\leftrightarrow o}$ = new edges between $v \in V_o$, $E_{n\leftrightarrow o}$ = edges between $v \in V_o$ and $v \in V_n$, $E_{n\leftrightarrow n}$ = edges between $v \in V_n$. The term 'old' in a particular year denotes the set of keywords appeared in the previous year(s)

	year	V_t	V_n	V_o	E	$E_{o\leftrightarrow o}$	$E'_{o\leftrightarrow o}$	$E_{n\leftrightarrow o}$	$E_{n\leftrightarrow n}$
G_s	2007	857	857	0	1486	0	0	0	1486
	2008	968	572	396	1690	144	530	777	239
	2009	978	438	540	1587	202	609	636	140
	2010	1000	332	668	1648	264	800	483	101
	2011	1172	345	827	2264	387	1166	599	112
	2012	1240	279	961	2399	271	1524	531	73
	2013	1328	272	1056	2715	424	1692	539	60
	2014	1367	291	1076	2728	500	1605	502	121
	2015	1262	157	1105	2347	512	1530	235	70
G_o	2007	3207	3207	0	15725	0	0	0	15725
	2008	3515	1217	2298	16826	3167	8995	4336	328
	2009	4206	1709	2497	21122	4515	9091	6695	821
	2010	4696	1152	3544	24093	6578	13154	4094	267
	2011	5059	658	4401	29261	6514	19962	2698	87
	2012	5426	364	5062	32663	9944	20900	1744	75
	2013	5418	77	5341	32076	11702	19987	382	5
	2014	4283	738	3545	22410	6522	12386	3251	251
	2015	4466	367	4099	23606	8043	13860	1548	155

in the same article. The term 'old' in any particular year t denotes the set of keywords appeared in year(s) prior to t. For example, in G_s for the year 2010, the number of V_o is 668 which denotes that out of $V_t = 1000$ keywords in 2010, 668 keywords appeared within 2007-2009. It is observable from the table that the number of new keywords normally decreases. When a new hypothesis gains considerable attention in the subsequent years, the related keywords become significant. This fact prompts expansion of these keywords' degree through new or recurring relations. Besides, new relations between old keywords $E_{o\leftrightarrow o}$ were found to be dominating which is the generic trend in inter-disciplinary research. Further, the growth of edges in $E'_{o\leftrightarrow o}$ implies that most new hypotheses emerge across existing keywords, topics, and/or concepts. Conversely, sporadic nature of new hypothesis generation across new keywords is observable via the decreasing number of edges in $E_{n\leftrightarrow n}$. Delayed consumption of new concepts by the scientific communities can be attributed as a cause. In case of $E_{n\leftrightarrow o}$, we observed that a lot of edges are formed with the most central keywords within the research domain which is true for knowledge network evolution (preferential attachment).

3.2. Preferential Attachment

The generation of author selected keywords is governed by the inherent rules of scholarly communication which is also known as preferential attachment (Zhao, Mao, & Lu, 2018). In citation-based scholarly communication, few articles within a research area will draw most attention and subsequently acquire most citations. These articles are considered as the representatives of the corresponding research area. Essentially, authors select keywords for their new articles either from the existing pool of keywords previously used or generate new ones conjoined with the existing ones. Therefore, few representative keywords having high degree centrality can always be found in KCNs (Choudhury & Uddin, 2016).

This phenomenon is demonstrated in Fig. 1 using two keywords ***Life course*** and ***Endothelial dysfunction*** (green coloured) from G_o (a-c) and G_s (d-f) respectively in three different times. In these figures, the red and navy coloured keywords are the existing keywords where the former denote the representative keywords in the respective research domain. The size variances of the keywords represent their degree of connections in the KCN of the corresponding time. The edge thickness represents the edge weights. From these network snapshots, it is observable that although new keywords emerge in association with less-acquainted (relatively unfamiliar) keywords in the corresponding research area in the beginning, however, gradually, they tend to co-appear with the representative keywords (e.g., cardiovascular disease in G_o and metabolic syndrome in G_s). For example, the endothelial dysfunction formed association with the keyword *erectile dysfunction* in the year 2008. In the year 2011, it formed associations with the high degree keywords such as *inflammation, hypoxia, metablic syndrome*. In the year 2014, it acquired connections with more divergent high degree keywords of the domain (e.g., *hypertension, sleep*). These figures demonstrate how a new keyword emerge by connecting with unfamiliar keyword(s) at first and giving rise to new hypotheses (Henry & McInnes, 2019) and later on achieve endorsement from the representative topics (high degree keywords). It is also observable that over time, some keywords (e.g., epidemiology in G_o and metabolic syndrome in G_s) achieve further new connections which are visible by the number of nodes connected to them and their edge weights in the year 2014. Hence, the rich club phenomena is also evident in dynamic KCN which can be captured by the degree centrality of the corresponding representative keywords.

In Fig. 2, we present the degree distribution (Fig. 2a & 2 c) by fraction of keywords in KCN built upon their co-occurrences (normalized) found in two different durations, 2007-11 and 2012-15 respectively. The right 'heavy-tailed' distributions in

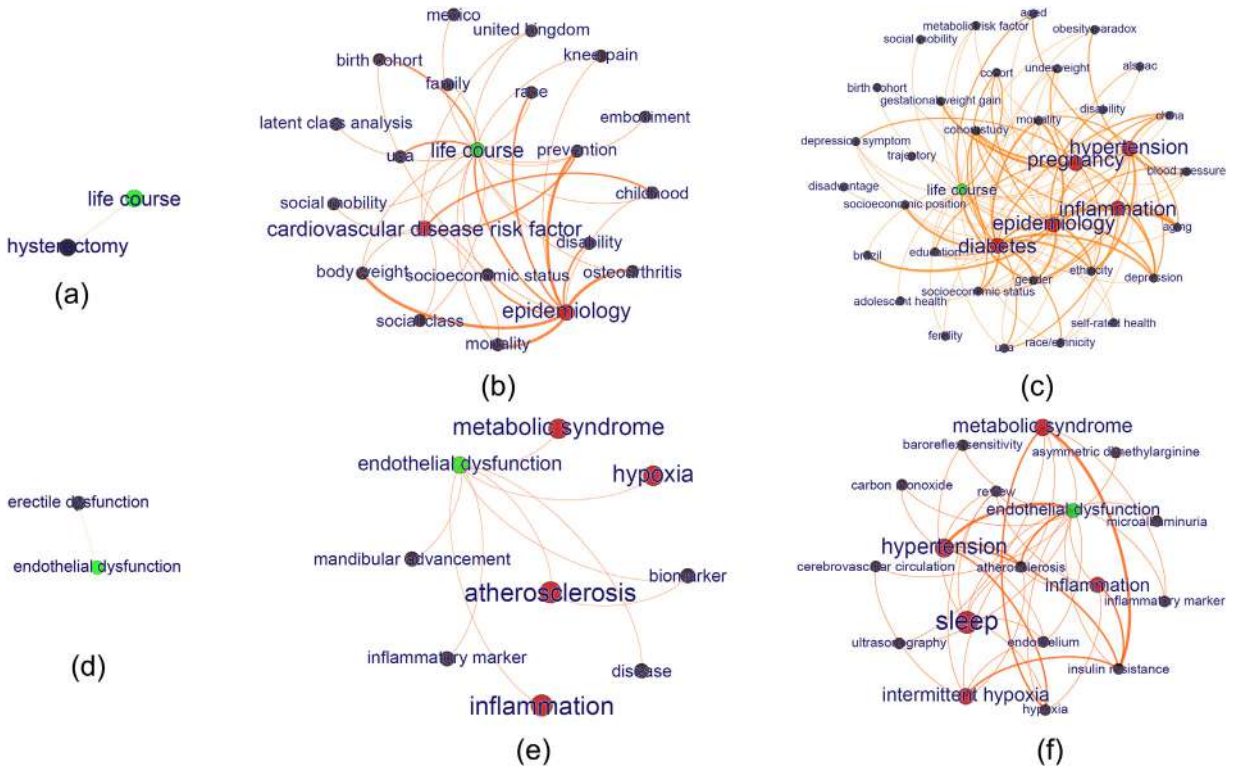


Fig. 1. Temporal patterns of co-occurrences between new and existing keywords demonstrated by two example keywords ***Life course*** and ***Endothelial dysfunction*** in G_o and G_s respectively. All network snapshots are timestamped and present the evolving connection patterns of the example keywords (green coloured) with the existing keywords (red and navy coloured) of the respective research domain. The red coloured keywords are the representative keywords of the domain those acquire more connections over time. The size of the node represents to the corresponding keyword's degree of connections. The thickness of the edges represents corresponding edge weights.

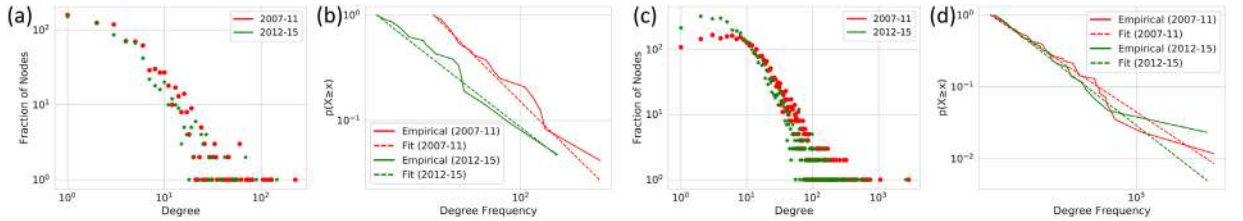


Fig. 2. The degree distribution (a,c) and powerlaw fit (b, d) computed by considering the complementary cumulative distribution function over degree distribution of KCN during two intervals (2007-2011) and (2012-2015) in both G_s and G_o respectively.

Fig. 2a and 2 c denote resemblance to the Barabasi-Albert model (Barabási & Albert, 1999) which assumes preferential attachment for the underlying network growths and explains the formation of power laws in the degree distributions of networks. Networks with power law distribution have many nodes with small degree and few nodes with very large degree. In two different durations, we observe that the underlying network growths follow preferential attachment model. In (Fig. 2b & 2 d), we present the powerlaw-fit with the help of ***powerlaw*** package (Alstott & Bullmore, 2014) which supports the complementary cumulative distribution function (CCDF). The package also allows to calculate the α of the formula $p(x) = Cx^{\alpha}$ for which we observed that the range was $\alpha \cong 2.56 - 3.5$ for both scenarios.

3.3. Recursive Centrality

Alongside the classical degree centrality measure, researchers attempted to define custom metrics to denote nodal importance. In one such study, Klimek et al. (Klimek, Jovanovic, Egloff, & Schneider, 2016) developed a bipartite relations using term-document matrix and proposed a recursively defined document centrality measure to denote the importance of scientific documents. According to their assumption, a list of documents is considered central if these consume a large number of central terms those in turn also are consumed by a large number of central documents. We adopt the concept of this recursive centrality for the keywords in KCNs instead of document centrality. Following the conceptualization of recursive

Table 4
Frequently Used Notations

Notation	Description
k_{gp}, k_p, k_c, k_g	Four keyword types: Grandparent, Parent, Children, Guest
v_t^{au}, v_t^{at}	Two variants of recursive centrality values for keyword v at time t (year) calculated from keyword-author and keyword-article bipartite relations
N_t^{au}, N_t^{at}	Sets of top N central keywords in year t *by* considering the bipartite network centrality measures

document centrality measures, two recursive keyword centrality measures were developed for each keyword v in each year t of the training period. Before jumping into the detailed definition of this recursive centrality, Table 4 provides descriptions of some notations those will be used frequently in the sections below.

The first centrality measure considers the keyword-author bipartite relations whereas the second considered the keyword-article relation. In the case of the first bipartite network, the intuition behind a keyword's importance is defined by the number of central authors who, in turn, also use a large number of central keywords. We considered all keywords from the normalized KCN developed in section 3.1. We constructed a keyword-author adjacency matrix $M_{A\times K}$ for each year t in the training period. It is worth mentioning that we considered the first name and last name of the authors and manually verified with the affiliation information to perform author name disambiguation. Although the objective of this manual author name disambiguation was to find all publications that belong to a given author and distinguish them from publications of other authors who share the same name, however, exploration of other methods for the same purpose is left for future studies. The adjacency matrix M is a binary matrix of size $|A| \times |K|$ where A and K denote the set of authors and keywords respectively in year t. An entry in matrix M is 1 if the author $a \in A$ uses the keyword $k \in K$ in his/her article. Starting with an example author a_i in year t, for each keyword k_i used by this author, we can reach all other authors who also used k_i in their articles. Iterating this procedure twice, we reach all the authors in two-hop distances from author a_i. These authors used keyword(s) that are also used by author(s) sharing some keywords with author a_i. This measure accumulates the number of different paths available among authors through their keyword usages. Higher number represents the fact that the corresponding keyword is pervasive across the research domain and thus more central and familiar to the key authors. To calculate recursive keyword centrality in this way, two vectors ψ_k and ψ_a are defined recursively:

$$\psi_k(n,t) = \frac{\sum_a M(t)\psi_a(n-1,t)}{\psi_k(0,t)}, \quad \psi_a(n,t) = \frac{\sum_k M(t)\psi_k(n-1,t)}{\psi_a(0,t)} \tag{1}$$

with n represents the number of iterations. $\psi_k(0,t)$ and $\psi_a(0,t)$ denote the initial conditions as $\psi_k(0,t) = \sum_{a\in A} M(t)$ and $\psi_a(0,t) = \sum_{k\in K} M(t)$. The values of ψ_k can have different interpretations. The initial condition $\psi_k(0,t)$ denotes the degree centrality of a keyword. For different values of $n = 1, 2, 3. ..$, ψ_k assigns weights to individual keywords considering the number of authors consuming them in their articles. Therefore, high values of ψ_k correspond to the keywords selected by a large number of authors who used large number of such keywords. Conversely, low values represent the keywords' specificity and relevance to limited number of research issues. Similar observations are also true in case of psi_a. The number of iterations converged and found stable at $n = 20$. We also attempted higher value of n, however, similar to the values at $n = 20$ were returned by the algorithm at higher values. In this way, for each year t in the training period, we calculated a vector of z-score normalized recursive keyword centrality values C_t^{au} by following the algorithm, from the keyword-author bipartite relationships. Likewise, another recursive keyword centrality measure C_t^{at} was computed from the keyword-article bipartite relationships in each year t of the training interval. In addition to these recursive centrality measures, to compare our analyses with a traditional centrality measure, we also calculated degree centrality values of keywords C_t^d, extracted from the normalized unipartite keyword-keyword relations in KCN (please see section 3.1) for each year t. Future studies can calculate other centrality values to compare the results further.

3.4. Genealogical Typology

As mentioned earlier, due to prevailing preferential attachment in evolution of dynamic KCNs, there exists a set of most central keywords which exercise greater influence over the structure of the network. It was observed that other keywords generally tend to form relations with these rich keywords by co-appearing in same articles. To predict the future co-evolution of keywords, this fact is required to take into consideration. In this study, we capitalize the idea of preferential attachment, but in a different way.

We defined a keyword's genealogical community membership in each year to denote it's current ancestral relation with the central keyword(s) of the previous year. The underlying objective here was to capture the evolution of a keyword's lineage to its ancestry. Genealogy is the study of family tree and a genealogy graph is used to portray the complex evolving relation originating from the ancestors. We defined a keyword's family relationship in a particular year t based on it's relation with the central keywords from previous year $t-1$ where the centrality of keywords was measured by considering the centrality measures defined above. This approach would label keywords according to the type of their ancestral relationships (e.g.,

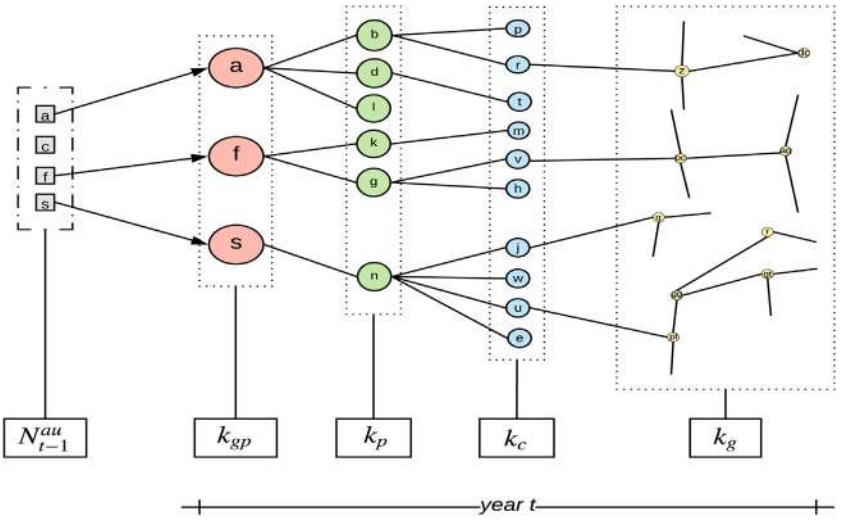


Fig. 3. Genealogical communities of keywords by considering a set of keywords with high recursive centrality values (i.e., N^{au}_{t-1}) computed from keyword-author bipartite relationships. Keywords with high centrality values at timestamp $t-1$ are designated as grandparents (k_{gp}) at timestamp t. Direct neighbours of these grandparents belong to parents (k_p) community, and two-hop neighbours are designated as children (k_c). Rest of the keywords are designated as guests (k_g).

grandparents, parents, child or guest) and help us understand the impact of these relationship types in their co-occurrences. The construction of the genealogical communities is described below:

3.5. Genealogical Community and KCN Evolution

In our dynamic KCN, we defined four types of communities for keywords, namely, grandparents (k_{gp}), parents (k_p), children (k_c) and guests (k_g). ***Grandparents*** k_{gp} in year t are the top-N central keywords from year $t-1$ according to a centrality measures. Here, the word 'central' denotes higher values of a chosen centrality measure from the three aforementioned centrality measures (i.e., two recursive and one degree centrality). In other words, these are the most frequent and influential keywords both in regards to the contents of metadata information (e.g.,author and article) and in forming relations with other keywords. We experimented with the number of grandparents ranging from $10-200$ by considering two recursive centrality measures. Interestingly, we found that having $N \geqslant 20$ grandparents in each year does not change the prediction performance (described later) significantly in both datasets. Therefore, we designated the top 20 keywords as grandparents by considering the chosen centrality measures. In this way, we got three sets of top-20 central keywords based on our three centrality measures (i.e., C^{au}_{t-1}, C^{at}_{t-1}, and C^{d}_{t-1}). Secondly, we define the parents, children and guests. ***Parents*** (k_p) are the keywords having direct relations (i.e., an edge is observed between them) with k_{gp} in the year t. ***Children*** (k_c) are the neighboring nodes of k_p keywords but not K_{gp}. Alternatively, k_p keywords are considered as the common neighbours between k_{gp} and k_c. Once these three communities of keywords (i.e. k_{gp}, k_p, k_c) are labeled, rest of the were designated as ***guests*** (k_g). These guests keywords had maximum distance from the grandparents. In this way, four genealogical communities of all keywords appearing in a particular year were defined. In our experimental setup, we use top-20 keywords from year ($t=0$) (i.e., 2007) for the first training year ($t=1$) (i.e., 2008), though this year ($t=0$) is not included within our training period. This fact was mentioned in the earlier section. 2. We also assigned a score for each keyword according to its genealogical community membership. The successive order of keyword's genealogical community score is: $score(k_{gp}) \gg score(k_p) > score(k_c) > score(k_g)$. Being most influential in the previous year, the grandparents (i.e., k_{gp}) were assigned the highest score in a year. Scores for the other three keyword types (i.e., k_p,k_c,k_g) were defined according to their distance from the grandparents k_{gp}. Being the most distant ones, k_g (guests) are assigned with lowest community score. The process of assigning a keyword's genealogical community scores is depicted in Fig. 3. The relative size of each keyword in this figure represents the weight of the score (i.e. greater sizes for higher values). In this figure, N^{au}_{t-1} denotes the set of ton-N central keywords by considering recursive centrality values v^{au}_t extracted from the keywords-authors network.

Considering three sets of centrality measures of keywords (i.e., v^{au}_t, v^{at}_t and v^{d}_t), a keyword belonging to one genealogical community (e.g., grandparent) in one set, might belong to different community (e.g., parent) in another. This effectively means that if a keyword becomes grandparent by considering the v^{au}_t, it may not necessarily be true in case of other centrality measures (i.e., v^{at}_t or degree centrality). Likewise, considering the same centrality measure, a keyword can belong to different genealogical community in different year(s). In Fig. 4, a comparative representation of these variations is presented for some keywords in G_s dataset. This figure presents network snapshots in G_s considering six keywords (i.e., ***atherosclerosis, morbid obesity, diagnosis, excessive daytime sleepiness, behavior, and compliance***) in different years. It is evident that there exists some commonality of keywords in both datasets since some of these keywords resemble to those in the G_o dataset. Since the

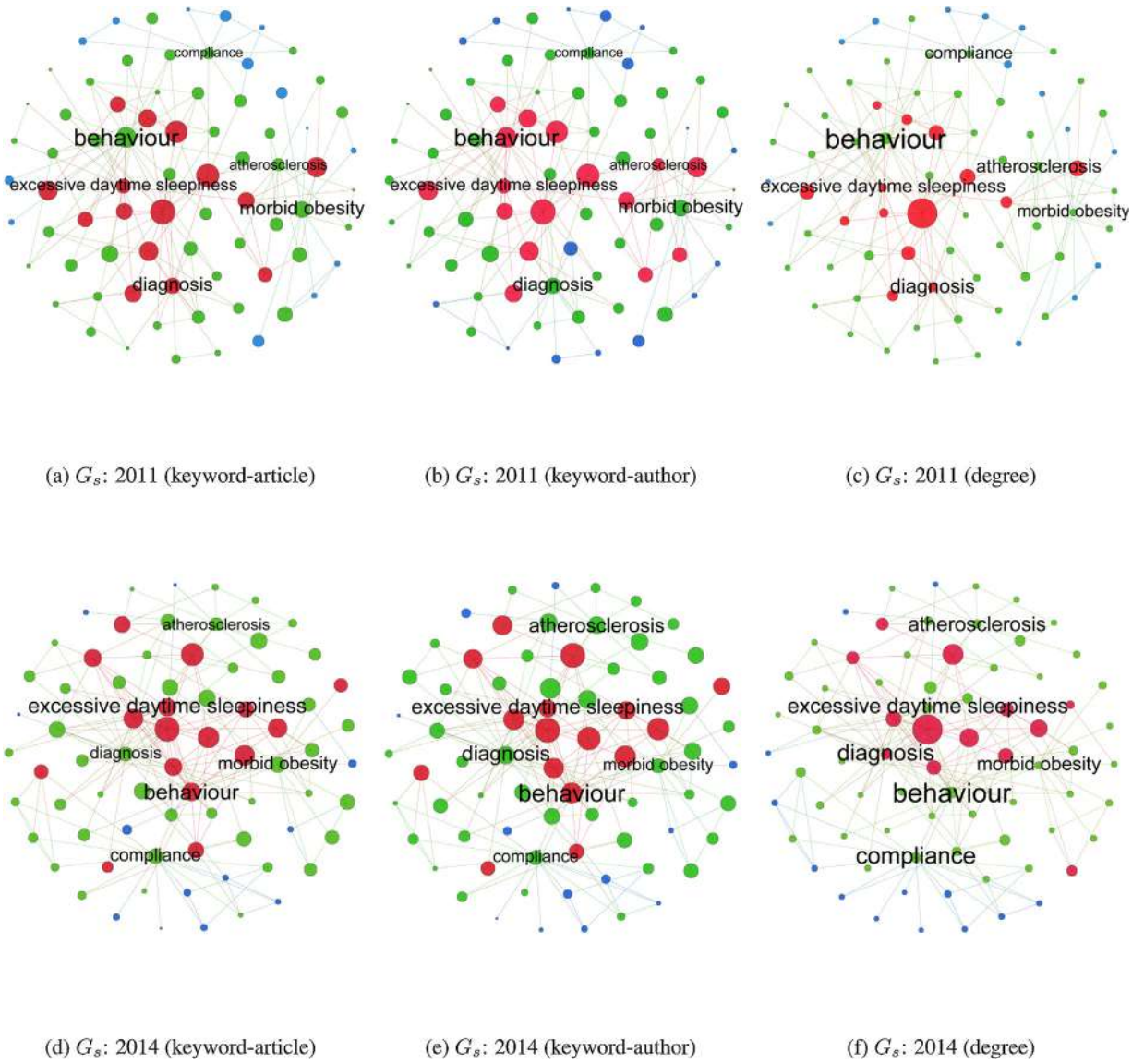


Fig. 4. Temporal variations of genealogical traits demonstrated by keywords depending on the centrality measures used in G_o dataset. All network snapshots are timestamped including the centrality measure used to identify the genealogical traits. The color codes represent the genealogical communities of keywords: red (grandparent), green (parent) and blue (child). The size of the node represents to the keyword's corresponding centrality measures.

commonality is out of the scope of this study, we leave it for future studies where domain experts can contribute towards literature based discovery related to these two domains together. However, our initial investigation revealed that excess weight and increased BMI are the strongest risk factors associated with the development of obstructive sleep apnea. The first row in this figure presents the network snapshots in the year 2011 and the second presents snapshots in the year 2014. In each row, the left snapshot presents the genealogical typologies (i.e., communities) of keywords identified by considering the recursive centrality (i.e., ν_t^{at}) computed from keyword-article bipartite relations. The snapshots in the middle column present the genealogical typologies of keywords identified by considering the recursive centrality (.i.e., ν_t^{au}) computed from keyword-author bipartite network. The right column presents the similar by considering the keyword's degree centrality. Color codes represent different types of the keywords: grandparent (red), parent (green), and child (blue). The sizes of the nodes and labels denote corresponding centrality values. It is observable that keywords can belong to different communities in different years by considering different centrality measures used in this study. For example, consider the genealogical communities of the keyword *diagnosis*. In the year 2011, it was designated as '*grandparent*' by considering ν_t^{at}, whereas considering the same centrality measure, it belonged to the '*parent*' community in 2014. Surprisingly, the later is also true (parent) when recursive centrality measure ν_t^{au} was considered in the year 2011. Similar observations are evident in case of the keyword *behaviour*

Table 5
Number of nodes (keywords) according to different genealogical typologies (i.e., communities) defined in this study and edges (keyword pairs) between different typologies of keywords. *GP* represents the grandparents keywords, *P* represents the parents, *C* represents the children and *G* represents the guests keywords. Similarly, E_P denotes the edges between parent keywords, and $EGP \leftrightarrow P$ denotes the edges between grandparent and parent keywords.

Year	P	C	G	$E_{GP\leftrightarrow P}$	$E_{GP\leftrightarrow C}$	$E_{GP\leftrightarrow G}$	E_P	$E_{P\leftrightarrow C}$	$E_{P\leftrightarrow G}$	E_C	$E_{C\leftrightarrow G}$	E_G
G_s												
2008	197	95	147	66	27	52	91	70	86	15	46	26
2009	179	84	187	65	32	56	89	62	95	13	30	32
2010	189	98	189	70	29	58	97	48	101	9	34	26
2011	225	116	175	81	29	52	112	64	115	13	41	25
2012	243	127	194	103	51	49	151	94	104	19	49	23
2013	273	167	174	99	45	52	156	121	114	23	49	24
2014	235	156	172	100	53	35	141	140	66	31	64	27
G_o												
2008	1198	811	326	425	288	102	2206	1853	590	289	219	24
2009	1122	716	232	430	276	68	2252	1729	413	346	167	11
2010	1286	756	237	485	264	82	2828	1813	402	299	126	15
2011	1403	766	211	515	248	52	3138	1865	409	339	137	12
2012	1560	864	190	610	285	46	4130	2373	438	305	96	5
2013	1585	853	175	540	298	71	4005	2533	427	357	104	5
2014	1528	1376	295	562	435	66	4553	3578	532	658	195	20

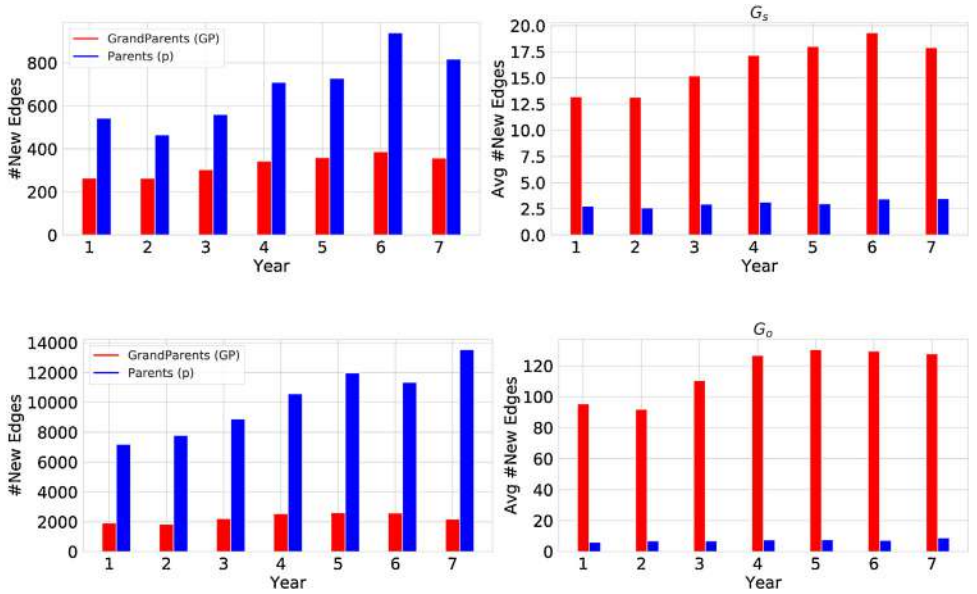


Fig. 5. Total number of new edges (left) and average number of new edges (right) acquired by both grandparent (GP) and parent (P) keywords collectively and individually over 7 years (2008-2014).

In Table 5, basic statistics of nodes (keywords) and different types of edges (keyword co-occurrences) formed between keywords belonging to different genealogical communities are presented. In this table, we ignored the grandparent keywords since for each year it was the top-20 keywords from the previous year. Consequently, edges among the grandparent keywords were also ignored since they were trivial in comparison to the other types of edges. In both datasets, we found that edges among keywords belonging to the parent community dominate in numbers. This fact is correlated with the increasing number of edges between grandparents and parents over time. Considering the sizes (i.e., number of articles and keywords) of both datasets, G_s harbours more guest keywords and edges. However, the ratio of edges between children and guest keywords are similar in both datasets. The children and guest keywords tend to form edges with parents more than the grandparents.

This chronology of descendants not only helped us understand the temporal trends of topics and research hypotheses developments but also attribute specificity in preferential attachment. For example, in absence of such typologies, we would consider grandparents and parent keywords not only in the same category and but also the richest since these would acquire most of the emerging edges. From this table, we can also observe the fact that there exists some emerging parent keywords competing with the grandparents in accruing new links over time. However, on average, the grandparents were still the most appealing collectively in accumulating most new relations formed each year. This phenomenon is depicted in Fig. 5 where we observe the total number of links acquired by both grandparent and parent keywords during 2008-2014 in both

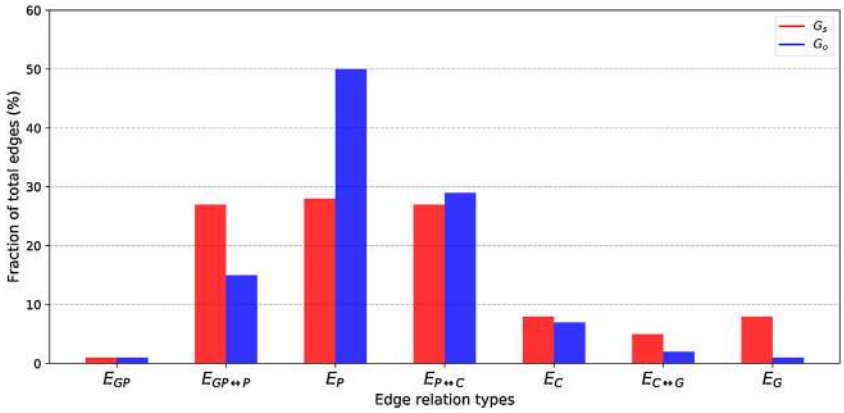


Fig. 6. Percentage of different types edges among keywords from different genealogical communities with respect to the total number of edges in the test year (i.e., 2015).

datasets. It also presents the average number of new edges acquired by each individual keywords from both categories. These figures demonstrate that while parent keywords compete with grandparents in attracting new links, however, the representative keywords - the grand parents - are still the biggest achiever on average. While this signifies the preferential attachment, however, such classification using genealogical traits will allow us to comprehend the evolutionary growth of KCNs better and thus help in science mapping.

In Fig. 6, we observe how this ancestral relationships helped us to classify the edge types in the test year (i.e., 2015). In this figure, we present the percentage of different types of edges, as presented in Table 5, with respect to the total number of edges. It is evident from this figure that the keywords belonging to the parent community play crucial roles in emerging relations than any other keyword types. In both datasets, the parent keywords dominate in attracting both descendants and antecedents to form emerging relations.

4. Research Methodology

We frame the future LBD prediction problem as a learning-based link prediction task. Predicting co-evolution of author selected keywords or identifying implicit future relationships between them can be mapped to the dynamic supervised link prediction model from network science. In this section, we describe our research methodology to construct such a predictive model that incorporates both nodal and edge features, and a recurrent neural network for both feature forecasting and classification.

4.1. Dynamic Supervised Link Prediction

For the purpose of link prediction, the total duration in each dataset was split into two non-overlapping intervals, T and $T+1$, known as the training and test phase. The primary objective of dynamic supervised link prediction framework is to analyze the temporal topological structure and nodes' attributes in the network of training phase $G_T(V_T, E_T)$ to predict the emergence of future edges in the network of test phase $G_{T+1}(V_T, E_{T+1})$. In dynamic network perspective, G_T is sampled using an aggregation granularity to generate a time series of network snapshots. Due to the data collection granularity this study used yearly window size to sample the network G_T and generate time series of network snapshots. It is impractical to predict the edges between nodes that are absent in the training network. Therefore, V_T is the set of nodes appearing in both phases. Then, we prepared our classification dataset consists of the instances of non-connected keyword-pairs in the training phase. Each instance is labeled either positive or negative based on its appearance as a true link in the test network. The supervised method for link prediction problems needs to predict the emerging edges by successfully discriminating the positive and negatively labeled keyword-pairs within the classification dataset. Hence, supervised link prediction is considered as a binary classification task by learning positive and negative instances with the help of interesting features describing each instance. A classification model requires effective features for training purposes. To perform this task, we constructed novel features (described later) by incorporating temporal information, network structure, and genealogical community memberships of the author selected keywords. The feature values for both positive and negative instances were computed by considering each network snapshot at timestamps $t_0, t_1, t_2, \ldots t_n \in T$. Here, each timestamp represents a year in the training phase T (i.e., 2008-2014). A recurrent neural network (described below) was also employed to forecast the feature values for each instance in the test phase. These forecasted feature values were input into a classifier for the classification purpose.

To measure the classification performance, there are mainly two categories of evaluation metrics in a supervised learning problem: (i) Fixed threshold metrics like accuracy, precision, recall (ii) k-equivalents and Threshold curves like precision-recall (P-R) curve and the area under the ROC curve (AUC) (Yang, Lichtenwalter, & Chawla, 2015). The threshold curves like

ROC curve and P-R curve are two dimensional curves. In ROC curve, true positive rates are plotted on the Y-axis and false positive rates are plotted on the X-axis. In P-R curve, Y-axis is for precision and X-axis is for recall. P-R curve provides better measurement in case of class distribution with large skewness. It also performs better in providing more insight regarding the exposure of class differences.

4.2. Feature Engineering

To support the supervised link prediction task, we constructed different features by taking advantage of different recursive centrality measures and genealogical communities extracted from temporal KCN. These features assessed the influence of keywords with regards to the authors and articles and leveraged the temporal significance of evolving networks. Before generating features for keyword-pairs (edge-level), we first identified different characteristics of individual keyword's importance. The rationale behind this is to contemplate different levels of significance each keyword carries with regards to both network importance and ancestral lineages. We also considered accumulated citation scores over time for each keyword to construct citation-based features. Since, these features only pertain to one keyword, some aggregation functions are needed to combine/aggregate the individual feature value of the corresponding keywords in a keyword pair. To illustrate further, consider the fact that if either(or both) of the keywords are prolific or belong to the same community, it is more likely that they will co-occur. Before aggregation, the individual measure denotes the proliferation rate of individual keyword - its community membership, or temporal activeness (number of co-occurrences). Aggregation of these individual features yields the aggregated features which are meaningful for each individual keyword pairs in dynamic link prediction. Hence, aggregation of keyword-specific features develops features for each keyword pairs. In this study's context, we assume that the higher the aggregated feature values for a non-connected keyword pair, the more likely that those two keywords will co-occur. In the following section, we describe the features used in the classification task.

4.2.1. Keyword centrality

For each keyword, we considered three centrality measures as node-level features. Two recursive centrality measures, computed over keyword-author (v_t^{au}) and keyword-article (v_t^{at}) bipartite networks as defined in section 3.3, and degree centrality (v_t^{d}) for comparison's sake.

4.2.2. Temporal community importance

This is the first edge-level aggregated feature we computed for each non-connected keyword pair for the link prediction purpose. The objective of this feature was to capitalize both the origin information (genealogical community) and current network importance (centrality measures) of both keywords in a keyword pair. We first multiply individual keyword's genealogical community score (section 3.5) with its corresponding network importance computed by a chosen centrality value for each year in the training period. The former denotes the keyword's relation with previous year's central keyword (i.e., keyword's family lineage or origin information) and the later denotes its degree on connections. This way generates the individual temporal community importance score for each keyword in a year. Individual temporal community score put more values on the grandparent and parent keywords than those of children and guest keywords. The final temporal community score for a keyword pair in each year t was the total of individual temporal community score for each keyword in that pair. Hence, for a keyword pair (a, b) in each year t of the training period, the temporal community importance score was computed as:

$$score_t^H(a, b) = g_t^a \cdot v_t^a + g_t^b \cdot v_t^b \tag{2}$$

where g_t^a and v_t^a denote the genealogical community score (3.5) and the chosen centrality score (3.3) of keyword a. By considering the temporal sequences of $score_t^H(a, b)$, we built a time series of this aggregated feature for each instance of non-connected keyword pair (a, b). Since we considered three centrality values (i.e., v_t^{au}, v_t^{at} and v_t^{d}) for each keyword v, the following notations will be used to denote three variants of this feature value: $score_{au}^H(a, b)$ to denote the temporal community importance score for an edge (a, b) where recursive centrality measures of keywords a and b were extracted from the keyword-author bipartite network. Similarly, $score_{at}^H(a, b)$ will denote the same feature value where recursive centrality measures were extracted from the keyword-article network. Finally, $score_{d}^H(a, b)$ will be used to denote the same feature value where degree centrality values of keywords a and b were considered.

4.2.3. Citation-weighted Recency

An individual keyword's citation count represents its influence factor over other keywords. There exist few considerable facts in temporal citation-based feature computation. Firstly, domain-specific and representative keywords generally achieve higher citation counts than the others. Secondly, old keywords accumulate more citations over time then new keywords. Conversely, some new keywords can gain significance quickly and acquire relatively high citation then other insignificant but old keywords. Third, considering our training period (i.e., 2008-2014), an article published in 2008 gets more time to be cited than an article published in 2014. Finally, training year closer to the test period (e.g., 2014) is more significant than earlier years in computing the feature score since recently appeared keywords are more likely to come together in near future, a fact known as 'recency' in link prediction task (Yang, Chawla, Sun, & Hani, 2012).

Therefore, we considered temporal factors of keyword's appearances including current or distant appearances over time and introduced a relative influence factor to compute citation weighted recency. Since citation count and temporal recency are applicable to individual keywords, it requires to aggregate these counts for keyword-pairs. For each keyword pair in year t, we first aggregated their total citation counts. Then, we assigned a temporal recency score depending on their appearances in a particular year. The recent their appearances are, or alternatively, the closer their co-appearances were to the test year, the higher their recency score. For example, consider a non-connected keyword pair (a, b) in the test phase. For this keyword pair (a, b), let τ denotes the chronological index of the training years $\tau \in [1, 2, 3, .. . T]$ where 1 denotes year 2008, 2 denotes year 2009, 3 denotes year 2010 and so on. Let h_t^a = total citation count of keyword a in year t and h_t^b = total citation count of keyword b in year t. The weighted-citation recency is calculated as follows:

$$score_t^W(a, b) = (h_t^a + h_t^b) * \gamma * \tau \tag{3}$$

Here, γ amplifies the appearance effect and τ amplifies the recency effect. For every non-connected keyword pair (a, b) in year t, $\gamma = 2$ if both a and b appear in a year t, $\gamma = 1$ if either a or b appears, other wise $\gamma = 0$. The feature value will be amplified depending on the current appearances of the keywords (i.e., multiplied by value of τ). For example, if a and b appear in 2008, $score_t^W(a, b)$ will be multiplied by 1, if they appear in 2014, then the feature score is multiplied by 7. The assumption here is that, keywords having high citation in the recent years will have high probability to appear together since they represent the recent trends. In this way, we build a time series of feature score $score_t^W(a, b)$ for the keyword pair $(a, b$.

4.3. Feature Forecasting and Classification

Considering temporal KCN $G_t(V_t, E_t)$, constructed for each year t during the training period, we constructed time series of features for non-connected keywords pairs. Like other supervised dynamic link prediction studies (da Silva Soares & Prudêncio, 2012), we employed a deep-learning framework to forecast the future values of constructed features during the test phase (i.e., 2015). We use a Long Short Term Memory network (Hochreiter & Schmidhuber, 1997) which is a special kind of Recurrent Neural Network (RNN) for both forecasting and classification tasks. An LSTM takes sequential data as input and is considered well-suited in classifying temporal sequence. The LSTM used in this study consisted of two blocks of memory-cells with two different layers of hidden units. A simple LSTM cell unit takes three inputs (X_t, h_{t-1}, C_{t-1}). X_t is the input of the current time step, h_{t-1} is the output from the previous LSTM unit and C_{t-1} is the "memory" of the previous unit. As for outputs, h_t is the output of the current unit and C_t is the memory of the current unit.

In our experiment, time series of feature values were input into a 2-layer LSTM network. The number of timesteps for each training sample is seven since the training period is seven years (2008-2014) long. To forecast numerical features values, *linear* activation function and MSE (Mean Squared Error) loss function were used. Categorical features were encoded as one-hot vectors. In this case, *softmax* activation function and categorical cross-entropy (loss function) were used. In all cases, adam optimization technique was used. Further, these forecasted feature values were then fed as the training samples for our classification task. This classification task was implemented by adding another LSTM netwrok in the pipeline which includes a dense layer with a single neuron as output layer. A dense layer is a fully connected layer which means each neuron here receives input from all the neurons in the previous layer, thus densely connected. In this layer, we used a logistic activation function named *sigmoid* which is ideal for assisting in binary mutual exclusive classification problem. This output layer takes the forecasted feature values and predicts the class (positive/negative) of the keyword pairs.

5. Results

In this section we present our feature forecasting and link prediction results. We first present the feature forecasting performance followed by the classification performance demonstrated by the LSTM.

5.1. Feature Forecasting Performance

We used LSTM to forecast the constructed feature values (4.2) for the test period. For each keyword-pair, time series of both keyword features (4.2.1) and aggregated features (4.2.2 and 4.2.3), computed during the training period, were input in LSTM. To evaluate the performance of this forecasting model, we computed the RMSE (root mean squared error) values by considering the actual feature values in the test year against the forecasted values. Table 6 present the RMSE values to measure the performance of LSTM in forecasting. We trained the LSTM model for different number of iterations (i.e. epochs). The effect of iteration number is visible in the RMSE values. In most cases, a higher number of iterations results in more robust model training. The range of actual values for different features can be varied significantly. So we presented the normalized RMSE values in the range of 0-1. RMSE values in G_o denote that in this dataset, the forecasting errors were smaller than those in G_s which can be attributed to the size of the dataset. In this table, forecasting performance for both the keyword feature and aggregated 18 features are presented. For comparison's sake, we also calculated RMSE in forecasting the *preferential attachment* score which is a widely used metric in the link prediction task. It is noteworthy that preferential attachment is an aggregated network feature that is computed by multiplying the number of neighbors of each keyword in a non-connected keyword-pair.

Table 6
Normalized RMSE values (0-1) calculated on forecasted feature values against true feature values in the test year. Features include both node-level (keyword features) and edge-level (aggregated features for keyword-pairs). $score_{PA}(a,b)$ represents the preferential attachment score. v_t^{at}, and v_t^{au} denote two recursive centrality values of keywords and v_t^d denotes degree centrality values of keywords.

Datasets	G_s			G_o		
Number of Iteration	100	500	1000	100	500	1000
			Aggregated Features (Keyword-pairs)			
$score_{at}^{H}(a,b)$	0.112	0.085	0.079	0.036	0.021	0.019
$score_{au}^{H}(a,b)$	0.115	0.080	0.085	0.032	0.018	0.022
$score_{d}^{H}(a,b)$	0.114	0.014	0.016	0.033	0.023	0.029
$score_{t}^{W}(a,b)$	0.110	0.011	0.010	0.018	0.016	0.016
$score_{PA}(a,b)$	0.114	0.014	0.014	0.009	0.009	0.009
$score_{PathSim}(a,b)$	0.132	0.123	0.127	0.082	0.083	0.088
			Individual Features (Keyword)			
v_t^{at}	0.127	0.125	0.125	0.062	0.066	0.066
v_t^{au}	0.107	0.100	0.101	0.049	0.048	0.048
v_t^{d}	0.029	0.023	0.072	0.013	0.011	0.011
Citation	0.005	0.006	0.003	0.006	0.006	0.005

Table 7
Dynamic supervised link prediction performance using AUCROC (AUC) and Accuracy (Acc%) values.

	G_s		G_o	
	AUC	Acc(%)	AUC	Acc(%)
$score_{at}^{H}(a,b)$	0.738	87.9	0.767	88.2
$score_{au}^{H}(a,b)$	0.743	85.5	0.778	88.0
$score_{d}^{H}(a,b)$	0.748	86.4	0.772	87.8
$score_{t}^{W}(a,b)$	0.685	78.9	0.757	87.7
$score_{PA}(a,b)$	0.676	88.9	0.754	86.4
$score_{PathSim}(a,b)$	0.722	84.6	0.766	87.1

5.2. Link Prediction Performance

The supervised link prediction framework is subject to highly imbalanced data with a very large number of instances with negative labels. In practice, only a few pairs of keywords participate in true emerging edges out of all possible pairs. Therefore, following other studies (Choudhury & Uddin, 2016), the ratio of the positive and negative class instances were set to 1:10. We also used 30% of the instances for validation and evaluation purposes. Similar to forecasting performance, G_o dataset was found to demonstrate high performance in comparison to G_s. The AUCROC (i.e., Area Under ROC Curve) is commonly used for evaluating such imbalanced classification problems. It introduces a probability value to quantify the uncertainty associated with the classifiers. In case of binary classification, AUCROC enforces a larger weight on smaller class by using this threshold value. In Table 7, we present both the accuracy and AUCROC scores computed in both G_s and G_o datasets. For comparison's sake, we also present the performance of ***Preferential Attachment*** metric which is a well-known topological similarity metric widely used in supervised link prediction tasks. It is evident from Table 7 that the features constructed in this study outperformed the traditional and widely prevalent metric - the preferential attachment, for link prediction in KCNs. Better performance was observed in G_o rather than G_s which can be attributed to the greater number of instances. It is also evident that the first aggregated feature (temporal community importance) performed better than the citation weighted recency scores.

Considering other evaluation metrics, the P-R (precision - recall) curve depicts the precision-recall trade-off for a classifier. This measurement is widely used in information retrieval. Reviewing both precision and recall is useful in cases where there is an imbalance in the instances between the two classes. The reason for this is that typically a large number of negative class instances mean we are less interested in the skill of the model at predicting negative class instances correctly, (i.e., high true negatives). The most crucial thing of P-R curve is that this calculation does not make use of the true negatives. It is only concerned with the correct prediction of the minority class, the positive class instances. A P-R curve is a plot of the precision in y-axis and the recall in x-axis. In ROC curve, the goal is to have a model be at the upper left corner, which is basically getting no false positives. Whereas, in P-R curve, the goal is to have a model be at the upper right corner, which is basically getting only the true positives with no false positives and no false negatives. In Fig. 7, we present both the ROC and P-R curves of our features (solid lines) including the Preferential Attachment (dashed lines) metrics. The top row represents P-R and ROC curves in G_s and the bottom row represents the same in G_o. Like the aforementioned performance measurements, in most cases of this figure, the features constructed in this study outperformed the preferential attachment metric in both datasets.

In section 3.5, we observed that, parent keywords (k_p) in both datasets are notable contenders not only in accruing emerging links among themselves, but also in forming links with the other types of keywords (i.e., grandparents, children and guests). While the accretion of new relationships by the grandparent keywords (k_{gp}) remains significant that represents the preferential attachment phenomenon in the emergence of scientific hypotheses via keyword co-occurrences, we tend

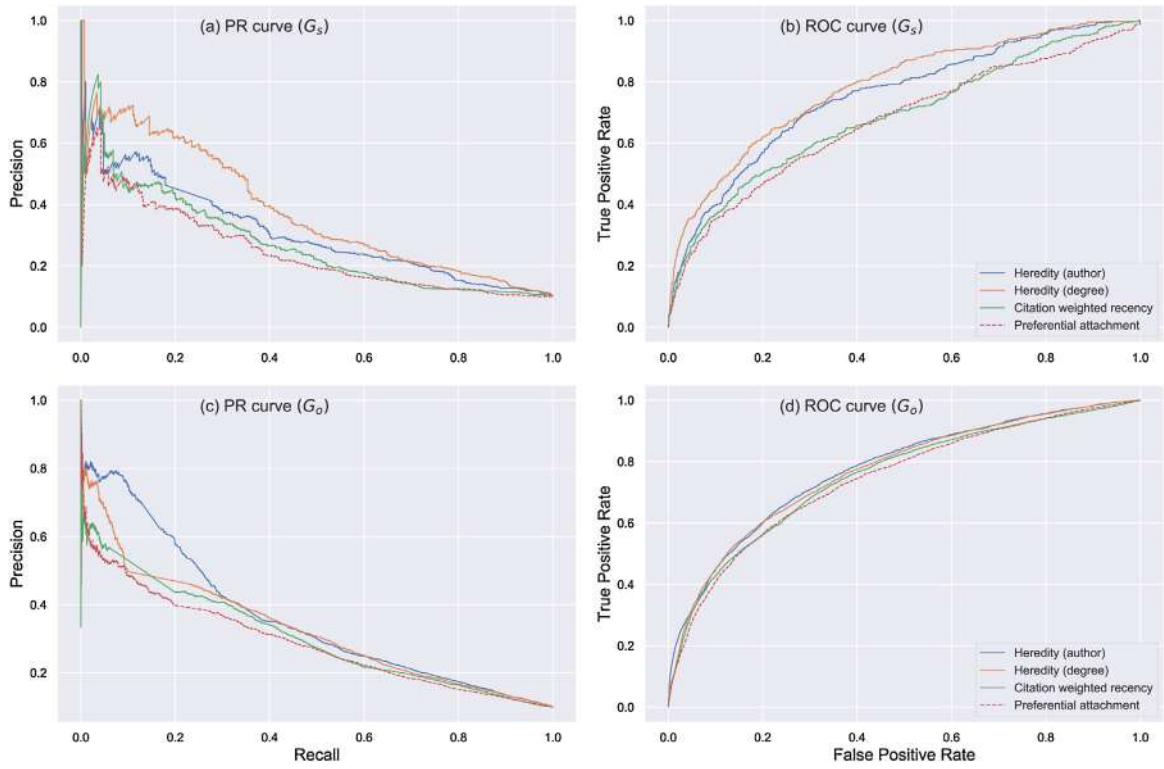


Fig. 7. P-R and ROC curves in both datasets (G_s and G_o) to demonstrate the classification performance of the LSTM classifier using the feature values constructed in this study. Traditional topological similarity metric 'Preferential Attachment' was also used to compare the performance with the constructed features. *Heredity (author)* denotes the edge-level aggregated feature temporal community importance $score^H_{au}(a, b)$. Similarly,*Heredity (article)* denotes the temporal community importance $score^H_{at}(a, b)$.

to observe the impact of parent keywords in predicting future co-occurrence among keywords. Similar to Fig. 7, we present the prediction performance via ROC curve in Fig. 8, however, this figure demonstrates the performance degradation due to the gradual elimination of the parent keywords including the keyword-pairs associated with them from the classification dataset. Note that in dynamic supervised link prediction, described in section 4.1, we constructed classification datasets comprised of non-connected keyword-pairs both positively and negatively labeled. A keyword-pair is either positive or negative depending on its presence or absence in the test period. In this figure, we gradually removed different percentages (i.e, 5%, 10%, 25%, 50%, and 100 %) of parent keywords and their associated keyword-pairs from the classification dataset. It is observable that with the removal of all (100%) keyword-pairs associated with the parent keywords in the classification dataset, the classification performance by our LSTM classifier degrades closer to the level of a random classifier (diagonal line from the bottom-left corner to the top-right corner of the ROC curve). Note that this observation is valid for the genealogical features constructed by considering both types of recursive centrality values (ν^{at} and ν^{au}). Two features considered in these ROC curves are $score^H_{at}(a, b)$ Fig. 8(a-b) and $score^H_{au}(a, b)$ Fig. 8(c-d). This signifies the impact of genealogical classification of keywords, proposed in this study, on the growth of keywords co-occurrences networks. Through this classification, we introduce a set of keywords (parent) other than the representative keywords (grandparents) in a research domain that not only contributes in preferential attachment but also supports the prediction of future growth of bibliographic networks.

5.3. PathSim

Heterogeneous knowledge networks often imply rather different structures from that in homogeneous networks. Edges in heterogeneous networks indicate the interactions between various types of nodes in a network, and usually imply similarity or influence among these nodes. Information is propagated across various kinds of nodes in a network, via various kinds of relationships (i.e., heterogeneous edges). In Fig. 9, we present two snapshots of HBIN constructed for G_s in the year 2008 and 2014. The green nodes denote the articles (documents), the blue nodes denote the authors and the red nodes denote the keywords. Correspondingly, the blue edges define the relationships between articles and authors, the orange nodes represent the relationships between keywords to articles, and the red colored edges define the relationships between keywords and authors. Note that these snapshots only represent 5-10% of the actual number of nodes and edges due to the size of the entire datasets.

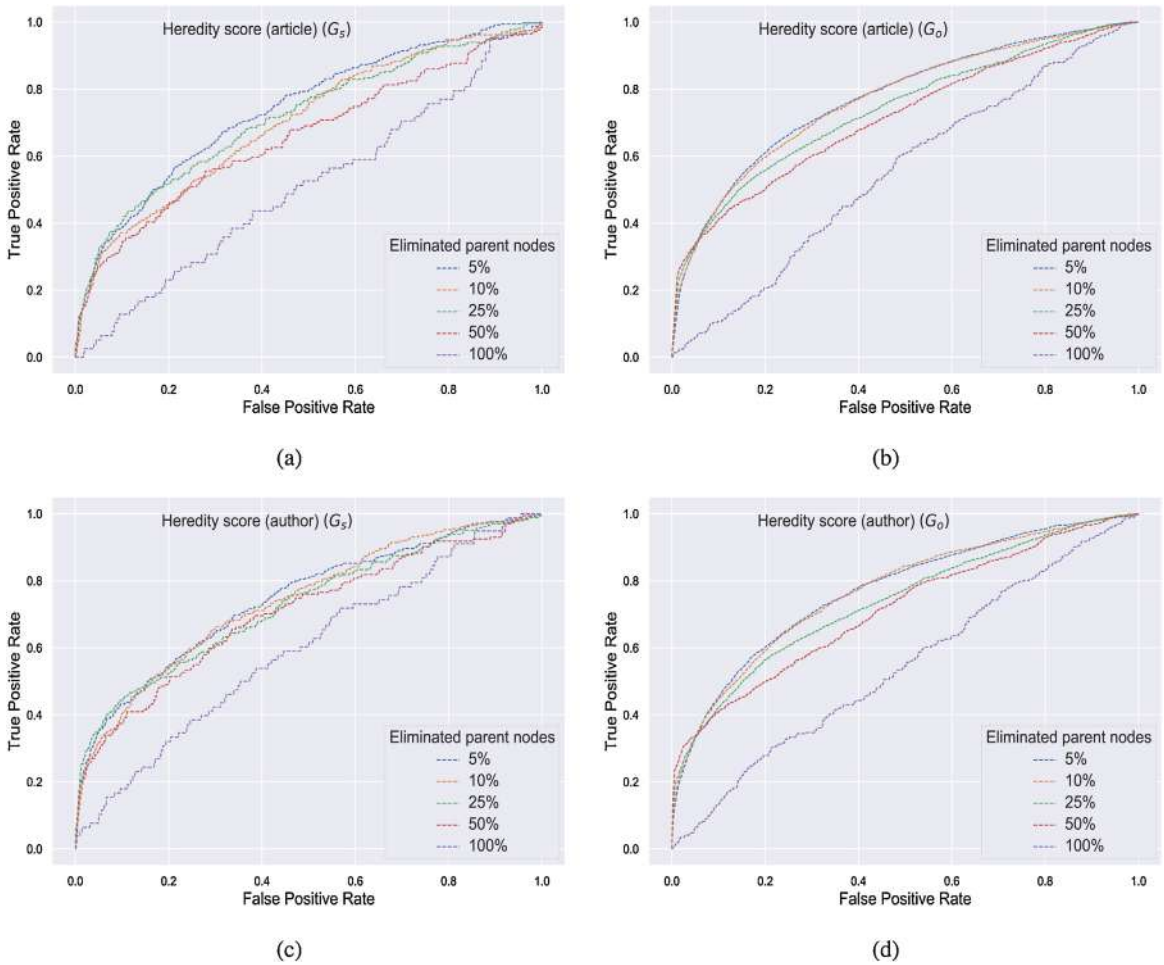
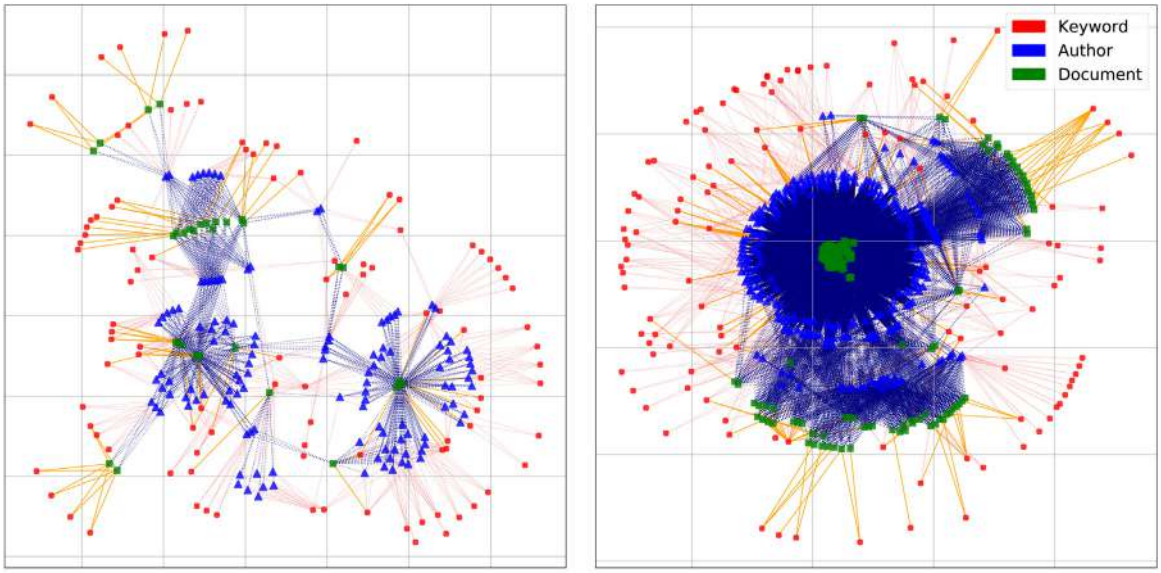


Fig. 8. ROC curves in both datasets (G_s and G_o) to demonstrate the classification performance downturn by the LSTM classifier with the reduction of parent keywords (k_p) and their associated keyword-pairs from the classification dataset. In all cases, ROC curves were generated by discarding different percentages (5%, 10%, 25%, 50% and 100 %) of parent nodes including their associated keyword-pairs from the test year (2015). *Heredity scores (article)* (a-b) denote the aggregated feature named temporal community importance $score^{H}_{at}(a, b)$. Similarly, Heredity scores (author) (c-d) denote the feature temporal community importance $score^{H}_{at}(a, b)$.

The similarity between two nodes in HBIN using link-based similarity function is determined by how the nodes are connected in a network that can be described using paths. Two keywords can be connected via different paths in a HBIN (e.g., keyword1-article-keyword2", keyword1-article-author-article-keyword2 etc.). Formally, these paths are called meta-paths which forms the building block of metapath-based similarity measure, called PathSim (Sun, Han, Yan, Yu, & Wu, 2011). Considering the relation between peer nodes should be symmetric, PathSim takes symmetric metapaths into accounts. Given a symmetric meta-path P, PathSim calculates similarity between two nodes a and b as:

$$score_{pathsim}(a, b) = \frac{2 \times \left|\left\{p_{a\to b}; p_{a\to b} \in P\right\}\right|}{\left|\left\{p_{a\to a}; p_{a\to a} \in P\right\}\right| + \left|\left\{p_{b\to b}; p_{b\to b} \in P\right\}\right|} \tag{4}$$

Similar to the approach followed in this study, Sebastian et al. (Sebastian, Siew, & Orimaye, 2017) also viewed the LBD process as a link prediction problem and proposed a new LBD method harnessing the lexico-citation information found in an HBIN using metapaths. In this study, to compare the prediction performance of our constructed features, we calculated similarity between non-connected keyword-pairs via PathSim over time-sliced HBIN constructed for each year in G_s and G_o. In a similar fashion, followed for the constructed features in this study, we computed a time series of PathSim scores for each non-connected keyword-pairs in each year 2008-2014. Then a PathSim score was forecasted using the LSTM for the year 2015 which was the input to the dense layer (please see section 4.3) for the classification purpose. In Table 6 and 7, we present a performance comparison for the PathSim score against the constructed features in this study. From these tables, it is observable that PathSim performed well in G_o and against the temporal citation-based features constructed in this study.



(a) G_s: 2008 (b) G_s: 2014

Fig. 9. Heterogeneous bibliographic information network snapshots for G_s in 2008 and 2014. The red colored nodes denote the keywords, the green-colored nodes denote the articles (documents) and the blue colored nodes represent the authors. The orange, red and light blue edges represent the inter-layer relationships between keywords-articles, keywords-authors and authors-articles respectively. These visualizations include 5-10% of the original nodes and edges found in the actual HBIN for those two years inG_s.

However, our recursive centrality-based features demonstrated competitive performance against PathSim both in G_o and G_s.

5.4. Distribution of Aggregated Feature Values

In Fig. 10, we present the distributions of feature values for the proposed aggregated features (i.e., temporal community importance $score^{H}_{au}(a, b)$ and citation weighted recency). For all positive and negatively-labeled keyword-pairs (samples), the normalized feature values for both aggregated features are presented as kernel density plots. Distribution of the feature values for the positive samples (true keyword-pairs in the test year) are presented in green color and the negative samples are presented in red color. Two observations evident from these figures are described below. It is noteworthy that, for temporal community importance score, we only considered the recursive centrality values extracted from the keyword-author relations.

Firstly, since the overlap between the red and green colored regions may trigger classification errors, the lower the overlap, the higher the classification performance. The amount of overlap signifies that the aggregated features computed in this study have non-trivial discriminatory characteristics. However, in Fig. 11, we present the distributions of the same feature without considering the keyword-pairs associated with the parent keywords in the classification datasets. We observe that the overlapped region increases due to the removal of the keyword-pairs associated with the parent keywords in the classification dataset. This is similar to the observation, described in 5.2, where we found the link prediction performance downturn demonstrated by the LSTM classifier. The reduction of the non-overlapped region between feature distributions of positive and negatively-labeled samples (keyword-pairs) leads to the degraded classification performance demonstrated by the classifier.

Secondly, in contrast to topological similarity metrics where the high value of the metrics corresponds to the higher similarity between a pair of nodes (e.g., having more common neighbors between two nodes in a network), we found that positively labeled keyword-pairs have high density in lower feature values. This denotes that instead of higher values of the features, comparatively lower features values have a high probability in forming emerging relationships. Although, in G_o, we observe high feature values for positively labeled edge instances in the test period. This fact is contrary to our initial assumption, mentioned in 4.2 that higher values denote higher probability of keyword co-occurrences. Although, exploring the reason behind this is out of the scope of this study, however, in Fig. 12, we present the normalized keyword feature values (recursive centralities and degree centrality) computed for the test year (i.e., 2015) in both datasets. The lower feature values can be attributed to the number of children and guest keywords including the variances in normalized centrality values of the parent keywords. It is noteworthy that from the aforementioned figures and tables, we have observed that the parent community dominates the emerging link formations.

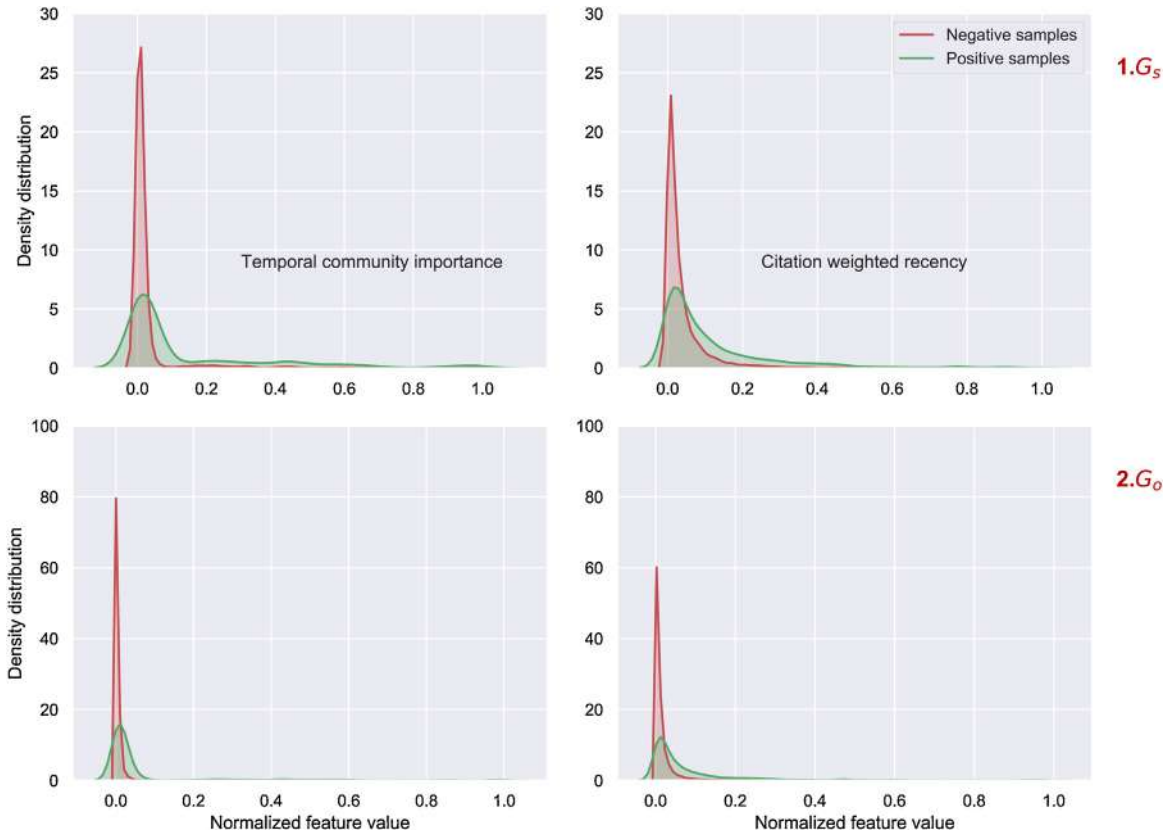


Fig. 10. Positive and negative class density of two aggregated features: temporal community importance score $score^{H}_{au}(a, b)$ where the recursive centrality measures were extracted from the keyword-author relations and Citation weighted recency $score^{W}_{t}(a, b)$ in both datasets.

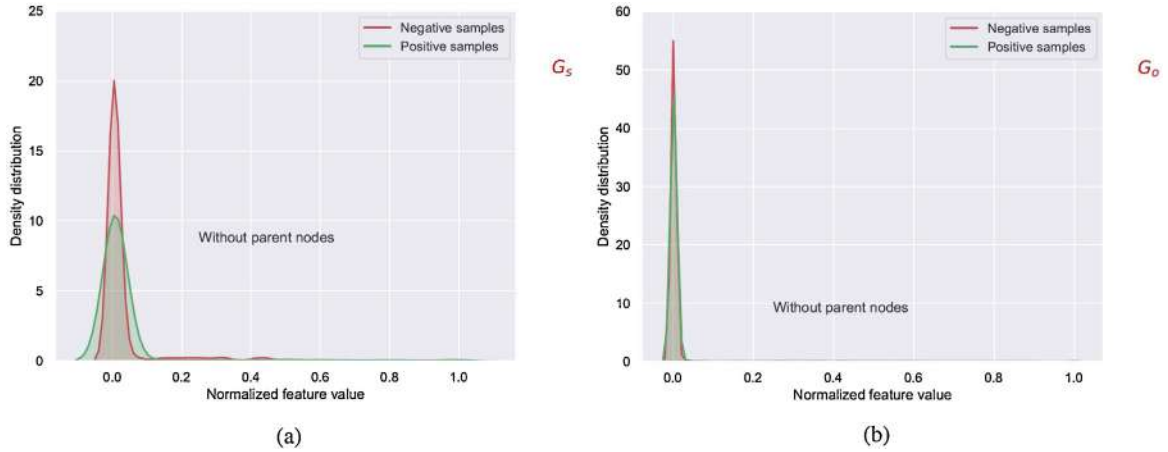


Fig. 11. Distribution of feature values for both positive and negative labeled keyword-pairs by considering the aggregated feature "temporal community importance score" $score^{H}_{au}(a, b)$ where the recursive centrality measures were extracted from the keyword-author relations in G_s (left) and G_o (right). In both cases, feature distributions were computed by discarding the keyword-pairs related to the parent (k_p) keywords.

6. Discussion and Conclusion

Scientific progress depends on formulating verifiable and deductible hypotheses generation. This requires both understanding and informed inferences from existing knowledge and information (Choi et al., 2018). The rapid growth of scientific knowledge and over specializations (domain-specific fragmentation) may engender opportunities to derive solutions from one domain to address problems in another, although the underlying relationship may remain implicit or the concerned groups from both domains are unaware of the work of each other (Hristovski, Peterlin, & Dzeroski, 2001). However, the

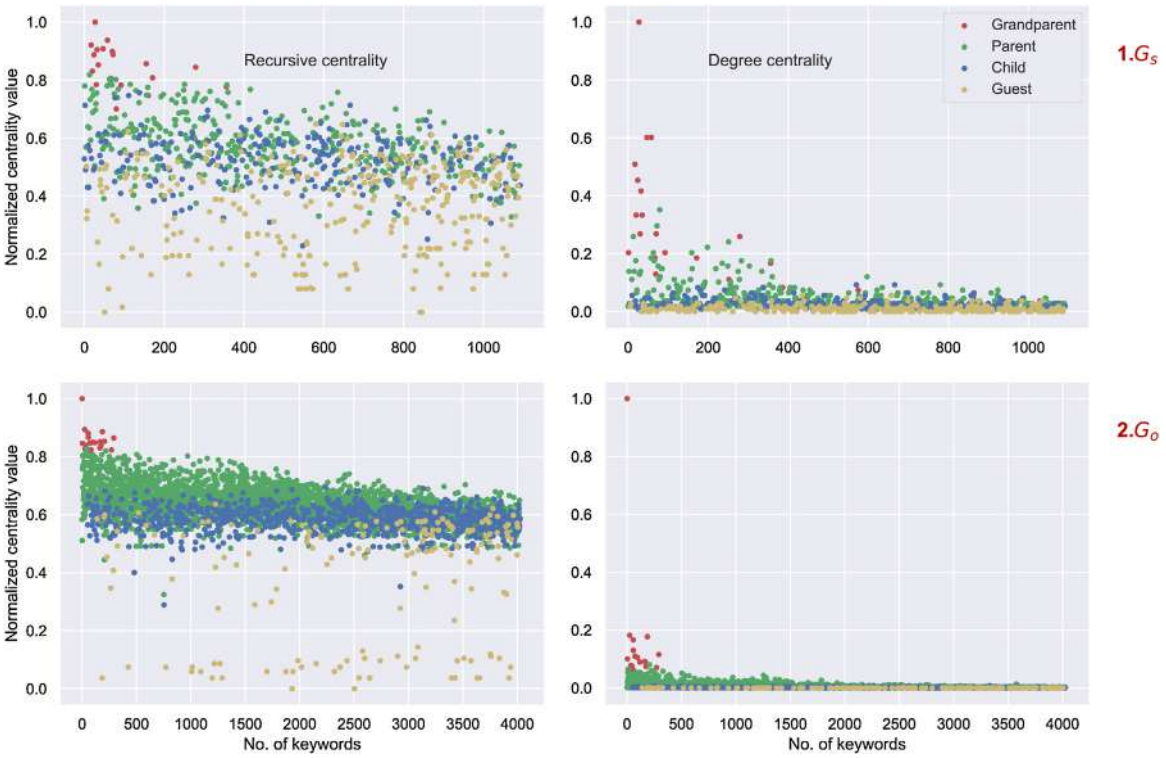


Fig. 12. Distribution of two different centrality values (i) recursive centrality values extracted from the keyword-author relations (left) and (ii) degree centrality of the keywords participating in the positively labeled edges during the test year in both datasets. Color codes represent the genealogical communities of keywords.

continuous surge in the number of published scientific literature limits the scope of analyses an individual can accomplish to extract these novel and implicit relationships between disjoint concepts, topics and domains (Wren et al., 2004). Proliferation of scientific production inhibits scientists and policymakers to detect trending subject areas and the linkages among these areas in their research fields, and mapping the dynamics of science to plan for research progress (He, 1999). human comprehension of such massive information and knowledge is challenging when it exceeds the scale of human analytical skills. For example, as mentioned by Spangler et al. (Spangler et al., 2014), it is inconceivable for a scholar to possibly assimilate, recall and accurately process all the known facts relevant to protein functions, relationships between proteins and identifications of roles of a particular protein related to a disease while there are over 70000 articles on a single protein - the tumour suppressor 'p53'. Thus, there is a great difference between what is known, and what we know as individuals from the collective and multidisciplinary knowledge within a given domain. Further, the specialization or fragmentation of literature may promote poor communication between specialties since scientists tend to communicate more within their fragments than the broader community engaged within the domain (Swanson, 2001). According to Ganiz et al. (Ganiz et al., 2005), Literature Based Discovery (LBD) addresses the challenge of seemingly boundless increases in scientific knowledge including knowledge overspecialization faced by the scientific communities. However, existing LBD models suffer from the lack of generalized predictive model to successfully predict the emerging trends in such discoveries. Despite their success, LBD techniques including text analysis, information retrieval and natural language processing are deprived of the benefits of bibliometrics, specially the analytical advantages of dynamic keyword co-occurrence networks (KCNs). These KCNs and network analysis methods are found to be supportive in identifying technological trends, analyse research topics and follow their evolution and track the development of innovation system research (Dotsika & Watkins, 2017). Further, temporal dynamics of KCN and community-aware features are underutilized in the process of literature-based hypotheses generation (Henry & McInnes, 2019). To this end, this study proposed a predictive model for LBD prediction that integrates temporal evolution of KCN, genealogical communities, and citation counts of keywords to construct predictive features, and a recurrent neural network for forecasting and prediction tasks.

The KCNs in this study were comprised of author selected keywords which best describe the research themes referred by the corresponding authors and are also considered as knowledge entities. The feature engineering process built novel features for both keywords (node) and keyword-pairs (edge defined by co-occurrences of keywords). Similar to this study, previous studies (Sebastian, Siew, & Orimaye, 2017) also viewed the LBD process as a link prediction problem to predict citation by using HBIN (heterogeneous bibliographic information networks). This study differs from the previous studies

by considering different recursive cemtrality measures computed from both unipartite and bipartite networks and the genealogical communities developed by considering the temporal network evolutions.

In this study, we developed two recursive centrality measures by considering two types of bipartite relations: keyword-author and keyword-article. We also considered the degree centrality of keywords extracted from temporal unipartite KCNs. These centrality measures were used to define temporal communities depending on genealogical relationships among keywords. The relative importance of temporal citation counts of keywords was also used as keyword features. Bipartite recursive centrality values including the degree centrality, genealogical community information and relative importance of temporal citation counts were used to construct edge-level (keyword-pairs) features. Seven years of training period was used to built time series of both node and edge-level features those were input into an LSTM network to forecast the feature values in the test year. With relatively trivial forecasting error, the forecasted feature values were used in supervised link prediction to classify both positive and negatively labeled non-connected keyword-pairs. The performance of the LSTM classifier was measured using well-known performance metrics and also compared against prevailing network topological similarity metric used in link prediction. High performance of the constructed feature indicates that these features are not only supportive in dynamic supervised link prediction but also can be beneficial in predicting literature based knowledge discovery or emerging trend detection. We also compared the prediction performance of the constructed features against features (i.e., PathSim) extracted from heterogeneous bibliographic information networks (HBIN) comprised of different types of relations (i.e., keywords-authors, keywords-articles, and authors-articles relationships). PathSim considered the symmetric metapaths in HBIN.

Despite the better performance measurements, this study is not free from any limitation. The first limitation comes from the lack of accommodating domain experts in identifying semantic similarity of keywords and various forms of abbreviations. Secondly, in this study we used manual name disambiguation instead of any standard methods to clean the author names including which can further be explored in future studies. Finally, the semantic relationships between textual contents were not considered in this study. The reason behind is that we only considered the author defined keywords those are not semantically presented in the scientific literature but only co-appears if the corresponding authors recognize them as relevant. To be more precise, author selected keywords appear as metadata information in the scientific literature. These keywords were not extracted from the actual textual contents of the literature (e.g., abstract or introduction) where the semantics matter.

This study can further be extended in various ways. Firstly, this study considered previous year's information to identify communities of keywords in the current year. However, future studies can consider further historical information (more than one previous year's information) to determine such communities. Other centrality measures and network community detection algorithm can be used to compare the performance of link prediction using community-aware features. Further, some weighting strategies can be followed to assign weights on edges between keywords belonging to either same community or different communities.

Author contributions

Nazim Choudhury conceived the manuscript idea, collected data, design experiments and contributed in writing, and Fahim Faisal cleaned the datasets, conducted the experiment(s), developed features, analysed the results, and wrote the research methodology section. Matloob Khushi contributed in research design.

7. Dataset and Code

The dataset and programming codes used in this study is available a https://github.com/faisal-iut/linkPrediction

References

Pritchard, A., et al. (1969). Statistical bibliography or bibliometrics. *Journal of documentation*, *25*(4), 348–349.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10–25.
Pan, R. K., Sinha, S., Kaski, K., & Saramäki, J. (2012). The evolution of interdisciplinarity in physics research. *Scientific reports*, *2*, 551.
Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information*, *42*(1), 91–106.
Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight. *Scientometrics*, *85*(1), 65–79.
Song, M., Han, N.-G., Kim, Y.-H., Ding, Y., & Chambers, T. (2013). Discovering implicit entity relation with the gene-citation-gene network. *PloS one*, *8*(12), e84639.
Yang, Y., Wu, M., & Cui, L. (2011). Integration of three visualization methods based on co-word analysis. *Scientometrics*, *90*(2), 659–673.
Choi, B. K., Dayaram, T., Parikh, N., Wilkins, A. D., Nagarajan, M., Novikov, I. B., Labrie, J. L., et al. (2018). Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by nek2. *Proceedings of the National Academy of Sciences*, *115*(42), 10666–10671.
Choudhury, N., & Uddin, S. (2016). Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics*, *108*(2), 745–776.
Smalheiser, N. R. (2017). Rediscovering don swanson: The past, present and future of literature-based discovery. *Journal of Data and Information Science*, *2*(4), 43–64.
Ganiz, M. C., Pottenger, W. M., & Janneck, C. D. (2005). Recent advances in literature based discovery, Journal of the American Society for Information Science and Technology. *JASIST (Submitted)*.
Preiss, J., Stevenson, M., & Gaizauskas, R. (2015). Exploring relation types for literature-based discovery. *Journal of the American Medical Informatics Association*, *22*(5), 987–992.

Henry, S., & McInnes, B. T. (2017). Literature based discovery: models, methods, and trends. *Journal of biomedical informatics*, *74*, 20–32.
Swanson, D. R. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, *30*(1), 7–18.
Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, *91*(2), 183–203.
Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PloS one*, *8*(8), e71416.
Börner, K., Sanyal, S., & Vespignani, A. (2007). Network science. *Annual review of information science and technology*, *41*(1), 537–607.
Cameron, D., Kavuluru, R., Rindflesch, T. C., Sheth, A. P., Thirunarayan, K., & Bodenreider, O. (2015). Context-driven automatic subgraph creation for literature-based discovery. *Journal of biomedical informatics*, *54*, 141–157.
Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, *49*(8), 674–685.
Hristovski, D., Stare, J., Peterlin, B., & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in medline and umls. *Studies in health technology and informatics*, *2*, 1344–1348.
Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical informatics*, *39*(6), 600–611.
Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., & Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, *20*(3), 389–398.
Smalheiser, N. R., Torvik, V. I., & Zhou, W. (2009). Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in medline. *Computer methods and programs in biomedicine*, *94*(2), 190–197.
Ahlers, C. B., Hristovski, D., Kilicoglu, H., & Rindflesch, T. C. (2007). Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annual Symposium Proceedings, Vol. 2007*, 6.
Hu, X., Li, G., Yoo, I., Zhang, X., & Xu, X. (2005). A semantic-based approach for mining undiscovered public knowledge from biomedical literature. *2005 IEEE International Conference on Granular Computing, Vol. 1*, 22–27.
Wilkowski, B., Fiszman, M., Miller, C. M., Hristovski, D., Arabandi, S., Rosemblat, G., & Rindflesch, T. C. (2011). Graph-based methods for discovery browsing with semantic predications. *AMIA annual symposium proceedings, Vol. 2011*, 1514.
Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *3*(2), 1–159.
Sebastian, Y., Siew, E. G., & Orimaye, S. O. (2017a). Learning the heterogeneous bibliographic information network for literature-based discovery. *Knowledge-Based Systems*, *115*, 66–79.
Sebastian, Y., Siew, E. G., & Orimaye, S. O. (2017b). Emerging approaches in literature-based discovery: Techniques and performance review. *The Knowledge Engineering Review*, *32*.
Ren, X., Liu, J., Yu, X., Khandelwal, U., Gu, Q., Wang, L., & Han, J. (2014). Cluscite: Effective citation recommendation by information network-based clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 821–830.
Liu, X., Yu, Y., Guo, C., Sun, Y., & Gao, L. (2014). Full-text based context-rich heterogeneous network mining approach for citation recommendation. *IEEE/ACM Joint Conference on Digital Libraries*, 361–370.
Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2016). Link prediction on a network of co-occurring mesh terms: towards literature-based discovery. *Methods of information in medicine*, *55*(04), 340–346.
Crichton, G., Guo, Y., Pyysalo, S., & Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC bioinformatics*, *19*(1), 176.
Katukuri, J. R., Xie, Y., Raghavan, V. V., & Gupta, A. (2012). Hypotheses generation as supervised link discovery with automated class labeling on large-scale biomedical concept networks. *BMC genomics, Vol. 13*, S5.
Wang, Y., & Zeng, J. (2013). Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics*, *29*(13), i126–i134.
Lu, Y., Guo, Y., & Korhonen, A. (2017). Link prediction in drug-target interactions network using similarity indices. *BMC bioinformatics*, *18*(1), 39.
Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.
Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2014). Link prediction in a mesh co-occurrence network: preliminary results. *in: MIE*, 579–583.
Van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In *Measuring scholarly impact*. pp. 285–320. Springer.
Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1st Edition). O'Reilly Media, Inc.
Kostoff, Ronald N. (2005). Method for data and text mining and literature-based discovery. *Google Patents*, Article 6,886,010.
Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*(6-7), 481–497.
Canals, A. (2005). Knowledge diffusion and complex networks: a model of high-tech geographical industrial clusters. *Proceedings of the 6th European conference on organizational knowledge, Learning, and Capabilities*, 1–21.
Montemurro, M. A., & Zanette, D. H. (2013). Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PloS one*, *8*(6), e66344.
Schulz, S., Costa, C. M., Kreuzthaler, M., Mi narro-Giménez, J. A., Andersen, U., Jensen, A. B., & Maegaard, B. (2014). Semantic relation discovery by using co-occurrence information. *Proceedings of BioTxtM 220*.
Eck, N. J. v., & Waltman, L. (2009). How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American society for information science and technology*, *60*(8), 1635–1651.
Zhao, W., Mao, J., & Lu, K. (2018). Ranking themes on co-word networks: Exploring the relationships among different metrics. *Information Processing & Management*, *54*(2), 203–218.
Henry, S., & McInnes, B. T. (2019). Indirect association and ranking hypotheses for literature based discovery. *BMC bioinformatics*, *20*(1), 425.
Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.
Alstott, J., & Bullmore, D. P. (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, *9*(1).
Klimek, P., Jovanovic, A. S., Egloff, R., & Schneider, R. (2016). Successful fish go with the flow: citation impact prediction based on centrality measures for term-document networks. *Scientometrics*, *107*(3), 1265–1282.
Yang, Y., Lichtenwalter, R., & Chawla, N. (2015). Evaluating link prediction methods. *Knowledge and Information Systems*, *45*(3), 751–782. http://dx.doi.org/10.1007/s10115-014-0789-0
Yang, Y., Chawla, N., Sun, Y., & Hani, J. (2012). Predicting links in multi-relational and heterogeneous networks. *2012 IEEE 12th international conference on data mining*, 755–764.
da Silva Soares, P. R., & Prudêncio, R. B. C. (2012). Time series based link prediction. *The 2012 international joint conference on neural networks (IJCNN)*, 1–7.
Hochreiter, S., & Schmidhuber, J. (1997). Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, 473–479.
Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, *4*(11), 992–1003.
Hristovski, D., Peterlin, B., & Dzeroski, S. (2001). Literature-based discovery support system and its application to disease gene identification. *Proceedings of the AMIA Symposium*, 928.
He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, *48*(1), 133–159.
Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C. R., Comer, A., Myers, J. N., et al. (2014). Automated hypothesis generation based on mining scientific literature. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1877–1886.

Swanson, D. R. (2001). Asist award of merit acceptance speech: On the fragmentation of knowledge, the connection explosion, and assembling other people's ideas. *Bulletin of the American Society for Information Science and Technology*, *27*(3), 12–14.
Dotsika, F., & Watkins, A. (2017). Identifying potentially disruptive trends by means of keyword network analysis. *Technological Forecasting and Social Change*, *119*, 114–127.