Article

JIS

Journal of Information Science
2025, Vol. 51(3) 638–657

DOI: 10.1177/01655515231191213
journals.sagepub.com/home/jis
SAGE

Do the paper's connections to existing work disclose its citation impact? A study based on graph representation learning

Zhuoran Luo
Wuhan University, China

Jiangen He
The University of Tennessee, Knoxville, USA

Jiajia Qian
Wuhan University, China

Yuqi Wang
Wuhan University, China

Wei Lu
School of information management, Wuhan University, China

Abstract

Influential scientific papers tend to be primarily based on combinations of prior works. However, assessing the potential impact of a new scientific paper remains a challenging task. In this article, we introduce an innovative framework to investigate the relationship between the embedding of citation networks and a paper's future citation counts, based on the graph representation learning approach. First, we employ three Nobel Prize-winning topic papers from the Web of Science as our data source. Through data preprocessing and direct citation network modelling, we train the struc2vec model to obtain embeddings of papers' citation network structure. Then, we perform visualisation and analysis on two types of networks. One is the direct-citation network, in which we identify four patterns of linkage between newly published papers and existing knowledge, and the other is the co-citation network, where we measure three structural variation indicators of new papers based on existing research findings. Finally, a statistical test is used to examine the predictive potentials of network embeddings. The results demonstrate that the structural features captured by the graph representation learning model can be used to predict a paper's citation counts and impact. This article innovatively combines cluster analysis, visual analysis and statistical analysis to gain insights into the relationship between the hard-to-explain structural embeddings of newly published papers in a citation network and their future citations.

Keywords

Citation impact; graph representation learning; knowledge structural; network embedding

Corresponding authors:
Wei Lu, School of information management, Wuhan University No. 299, Bayi Road, Wuhan, 430072, China.
Email: weilu@whu.edu.cn

Jiangen He, School of Information Sciences, The University of Tennessee, Knoxville 1345 Circle Park Dr, Knoxville, TN 37996, United States.
Email: jiangen@utk.edu

1. Introduction

Science can be described as a complex, self-organising, evolving network of scholars, projects, papers and ideas that reveals new scientific paradigms in scientific research [1,2]. If only bare, truly innovative and highly interdisciplinary ideas are left, they may not have the greatest scientific impact. Instead, to enhance their impact, novel ideas should be placed in the context of existing knowledge [3]. Some researchers claim that the greatest impact is based primarily on traditional prior work with unusual combinations [3,4]. It is a sign of high-quality research that is embedded in previous studies. Thus, citation counts are regarded as an important metric of research evaluation as they reflect this embedding [5].

Papers' citation count becomes the de facto standard for predicting and evaluating the impact of an article [6], researcher or institution [7], and many studies have focused on measuring and predicting the citation of publications [8,9]. The information used for citation prediction in prior research includes the number of pages, number of references, h-index [10], journal factors [11,12], keywords [13], institute [5] or other altimetric indicator [14], such as access counts, download counts and citation counts. Regression models [5,15–18], differential equation models [19,20], machine learning models [21], neural network methods [22–25] and graph models [2,26] have been widely used in citation count prediction. For example, Yan et al. [27,28] consider several regression models to formulate the learning process and use a range of features to predict citations which include author, content, venue, venue rank, venue centrality, topic rank, diversity, h-index, author rank, productivity, sociality and authority. Using author, venue, and article characteristics, Chakraborty et al. [29] divided citations into six groups and then constructed a regression model to predict the citation count within each group. Wang et al. [19] developed a mechanistic model to predict the citation dynamics of individual papers; they indicate that all papers tend to follow the same universal temporal pattern. Huang et al. [30] proposed a fine-grained citation count prediction task to predict in-text citation count from each structural function of a paper separately. Davletov et al. [31] employed machine learning approaches to predict the citation performance of articles by using temporal features. Du et al. [32] proposed a method to predict the citation counts of papers by using a modified LSTM (Long Short-Term Memory) network. Abrishami and Aliakbary [24] proposed a sequence-to-sequence neural network model to predict the long-term citation count of a paper. Shi et al. [33] hold that the structure of networks and the network characteristics of science largely determine how science evolves. Shibata et al. [34] used citation network topology analysis to investigate the factors that influence academic papers' potential to be cited. Min et al. [35,36] showed that certain features in the citation structure can distinguish between breakthrough and non-breakthrough papers, and recently, they proposed to predict early breakthrough research from variations in knowledge structure from a citation structure perspective.

Overall, existing literature shows that researchers have considered many methods and information in predicting the impact and citation counts of research papers. Related studies have also suggested that, for most of the indicators considered, evaluating citation impact is a time-consuming game due to the lack of data from newly published papers [14]. Due to its high generalisation ability, representational learning technologies are increasingly being used in information science research. Network representation learning (NRL) is the bridge between the raw data of the network and the network application tasks [37]; it represents the characteristics of each node in the network as a low-dimensional vector. These node representations can then be used in tasks such as node classification [38] and link prediction [39]. In recent years, NRL has attracted the attention of researchers in the fields of scientometric and science of science research, and related scholars apply it to community clustering [40,41], scholar evaluation and scientific collaboration [42–44], paper impact prediction [45], topic identification [46–49] and journal evaluation [50].

It is worth noting that the above studies seldom perform well on the classification task of determining whether the local topology of two nodes is equivalent, and the intrinsic cause is that the nodes in the network are homogeneous, that is, two nodes have edges connected because they have some highly similar characteristics. To address it, a graph representation learning named struct2vec [51] was proposed to generate node vector representations on the network that preserve structural features, which learns low-dimensional embedding representations for nodes in a graph and can be used to detect nodes in networks with similar structural characteristics. An early sign of a newly published paper's potential impact or novelty is how the paper has constructed links between itself and the existing knowledge network [52]. To investigate such links in the context of citation networks, it is essential to characterise structural features that describe how a new paper is embedded in the existing knowledge space.

In this study, we investigated whether the structural features of a scientific publication relate to its future citation impact. First, we adopt the struct2vec algorithm to generate a vector for each new node (newly published paper) in the citation network to obtain its topological characteristics in the network. Through visual analysis of direct citation networks and co-citation networks, we explored the relationship between a paper's future citations and its connection to existing knowledge clusters and elements. Then, we identified four patterns of how newly published papers link to the previous intellectual network. Furthermore, we quantitatively validate the relationship between structural features and

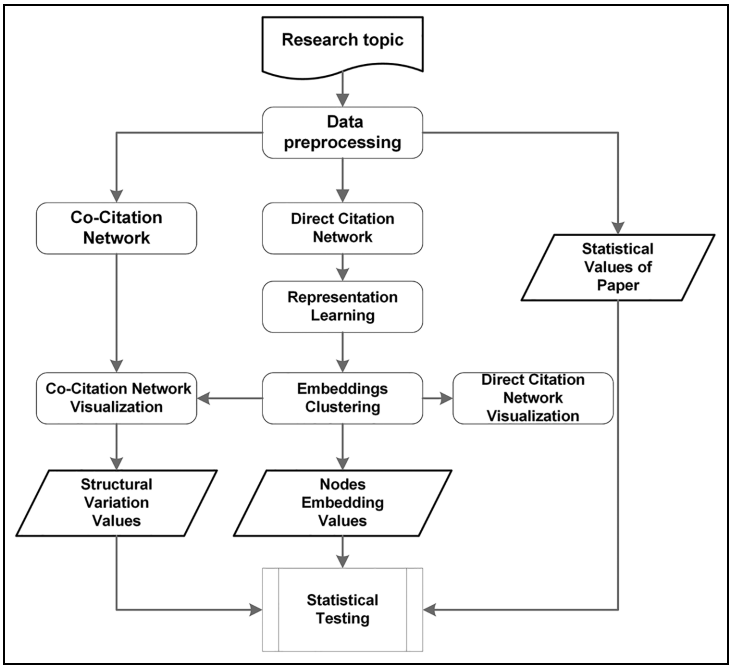

Figure 1. Overview of the key process for this study.

future citations using regression analysis, which method allows us to assess the significance of these features in predicting the impact and citation count of a paper within the academic community.

This work contributes to the research of science in the following ways: First, our research proposes a new framework to investigate how network embeddings relate to papers' future citation counts based on the graph representation learning approach. Second, we perform visualisation and analysis on two different types of networks to better illustrate the network embeddings and identify four patterns of linkage between newly published articles and existing knowledge that might influence their future citations. Our research offers a novel perspective on quantitatively analysing the hard-to-interpret network embeddings, which may provide practical implications for research on scientific impact in a broader sense by integrating deep learning with visual analysis and statistical methods.

The organisation of this article is as follows. In Section 2, we describe the overview of our research framework, followed by data strategies and detailed methodology used in this study. In Section 3, we present the case-study visual analytics and illustrate our findings of the analysis. In Section 4, we illustrate the statistical test result of our study. In Section 5, we discuss the results and their implication and conclude the article.

2. Methodology

As shown in Figure 1, we simply clarify the overall research framework of our study. For a given research topic, we first query the ISI Web of Science (WOS) Core Collection database to obtain bibliographic data related to the topic. After data preprocessing operations such as field parsing, null filtering and citation relationship extraction, then, we train the direct citation network with the struc2vec representation learning model, obtain the embedding vector representation of all nodes and use the K-means algorithm to cluster these vectors. For more visual observation of the variation in citation structure, we construct direct citation networks and co-citation networks by using Gephi and Citespace software, respectively. Finally, we use regression models to examine the relationship between the network structure of newly published papers and their future citations.

2.1. Data acquisition and definition

The purpose of this study is to investigate how newly published papers are embedded in the intellectual networks constituted by past papers. More specifically, newly published papers are the objects we want to study the structural variation of the network, papers published before the publication year are historical papers and the cut-off year for citation status

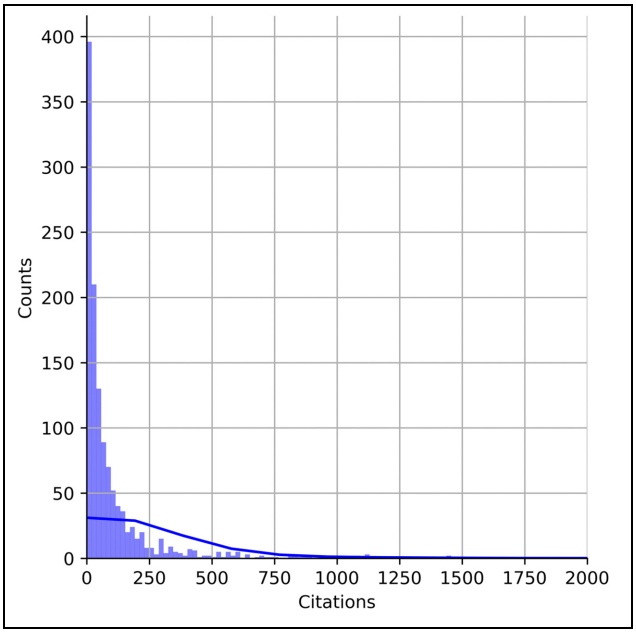

Figure 2. Citation counts distribution of three topics.

statistics is defined as the observation year. For example, if we consider papers published in 2005 as 'newly published', thus 2005 is the publication year, and then all relevant papers published before 2005 are our historical papers. If we want to study the citations of those papers in the publication year up to 2019, thus the year 2019 is the observation year, and then we will count the number of citations of those papers in the publication year up to 2019 and use them to test the prediction effect of network structure variation.

We first collected papers published before the publication year in the WOS Core Collection database on three Nobel Prize-winning topics: Autophagy, *Helicobacter pylori* and Graphene. We excluded publications that are not articles, letters or reviews. The first topic – Autophagy, or autophagocytosis, is also known as type II cell death [53]. In 2016, the Nobel Foundation announced the award of the Nobel Prize in Physiology or Medicine to Japanese scientist Yoshinori Ohsumi for discovering the principal mechanisms underlying this cellular autophagy process. For this topic, the publication year is 2005 and the observation year is 2019. The query we used for this topic is 'autophagy', 'autophagocytosis' and 'cellular autophagy'. For the second topic – *Helicobacter pylori*, the Nobel Prize in Physiology or Medicine for 2005 jointly to Barry J. Marshall and J. Robin Warren for their discovery of the bacterium *Helicobacter pylori* and its role in gastritis and peptic ulcer disease, and the related work was published around 1997. The search terms are 'helicobacter pylori' and 'H. pylori'. In this case, the publication year is 1997 and the observation year is 2019. For the third topic – Graphene, the Nobel Prize in Physics 2010 was awarded jointly to Andre Geim and Konstantin Novoselov for their groundbreaking experiments regarding the two-dimensional material graphene; we used 2004 as the publication year and 2019 as the observation year and use 'graphene' as a search query to collect data.

The number of papers until each publication year of three topics is 216, 245 and 789, respectively. The distribution of the number of citations for papers on the three topics is shown in Figure 2.

2.2. Citation NRL

We generate direct citation networks using citation relationships between papers in the original bibliographic records dataset for a certain research topic. Building on these networks, we propose a method for applying struct2vec to capture a citation network's structural features. In contrast to node2vec, which optimises node embeddings so that nearby nodes in the graph have similar embeddings, struc2vec captures the roles of nodes in a graph. Even if structurally similar nodes are far apart in the graph, the representations of the nodes learned by struc2vec would be similar. Struc2vec learns low-dimensional representations for nodes in a graph, generating random walks through a constructed multilayer graph starting at each graph node. The distance between the latent representations of nodes is strongly correlated to their structural similarity.

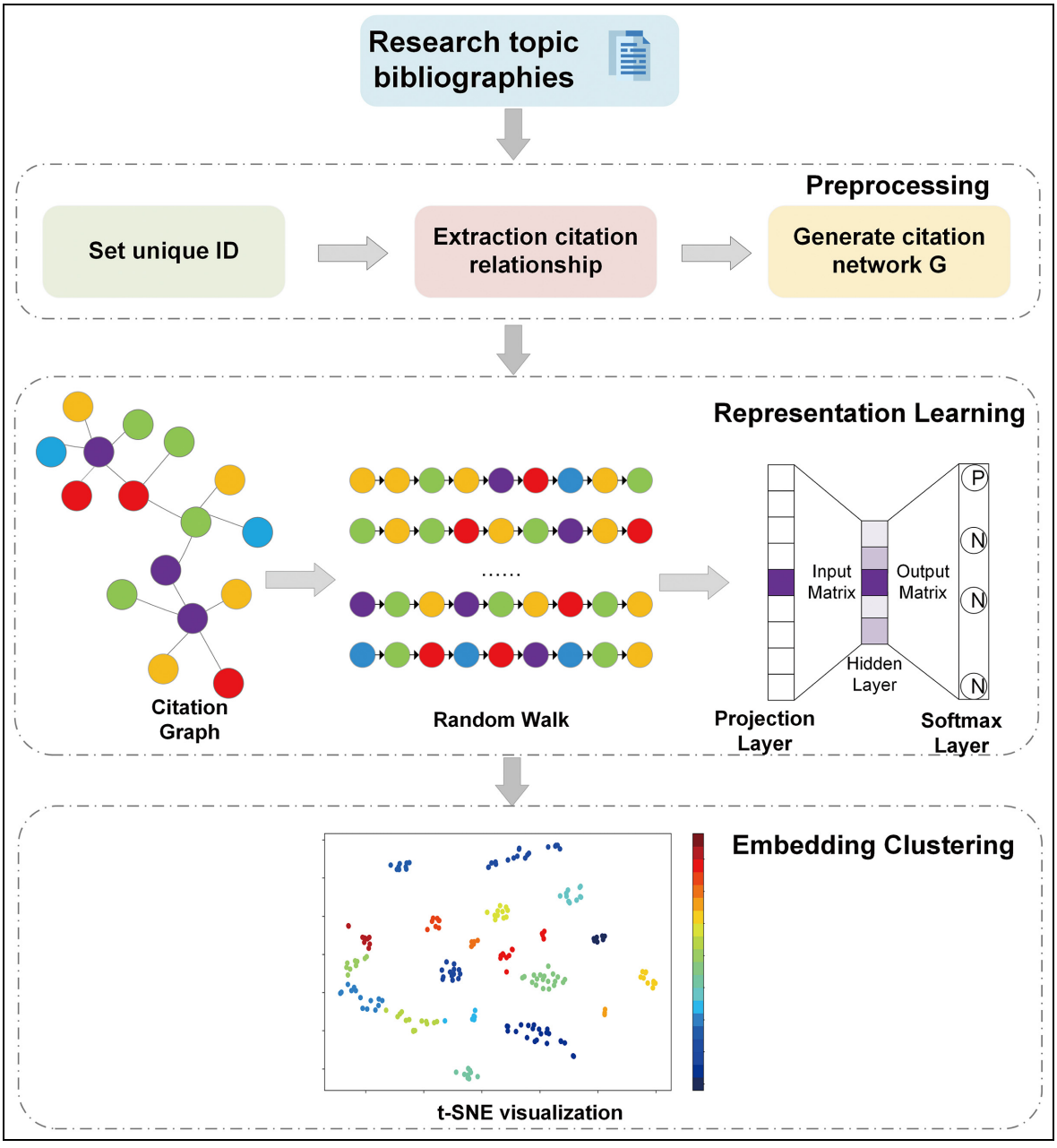

Figure 3. Flow chart of citation NRL process.

In our approach, as depicted in Figure 3, we first create a direct citation network G using the collection of papers on the selected research topic. An embedding representation on G is performed by the struc2vec model. Specifically, a multilayer graph is generated for each node context and a sequence of nodes is generated via biased random wandering on a multilayer graph. Finally, we use the skip-gram model to learn the node representation and generate a 128-dimensional embedding vector representation of each node in G.

2.3. Embedding results clustering and visualisation

For the node embedding vectors trained by the struct2vec model, the more similar the vectors represent, the more similar the structural features of the nodes in the network. To learn papers with similar structural features, we feed the vectors

of newly published papers to the k-means algorithm to identify the clusters of papers that build connections to existing knowledge similarly.

We use high-dimensional vectors that represent the structural features of newly published papers to identify clusters. To present the patterns of the clusters, we perform a dimension reduction and visualisation of those embedding vectors. The technique we applied is t-distributed stochastic neighbour embedding (t-SNE) [54], which is used for visualising high-dimensional data by nonlinear reduction to two dimensions while preserving some features of the original data. With the results of t-SNE, all the papers are presented in two-dimensional visualisation where the distance between the papers presents the structural similarity between them. We can also integrate the results of k-means clustering into the visualisation.

2.4. Network visualisation

Although the clustering analysis can help us learn the clusters of papers built with similar connections to existing knowledge, it is hard for us to distinguish what type of connections are made by the papers since those embeddings built by struc2vec lack interpretability. Network visualisations can provide an intuitive presentation of networks and enable informative interpretation. Thus, we adopt visualisation to explain the results from both direct citation network and co-citation network perspectives.

2.4.1. Direct citation network. We use a direct citation network to examine how newly published papers connect to the existing body of knowledge. Direct citation considers only links within a collection, that is, if a paper cites another paper, it will be linked. Direct citation networks are the fastest and best means of detecting emerging research frontiers [55]. To conduct visual analytics of the direct citation network, we use Gephi's Force Atlas algorithm for network layout and Louvain's algorithm for association modular division. Finally, we identify the clusters of papers with high, medium and low citation counts on average and analyse the network characteristics of those papers.

2.4.2. Co-citation network. Co-citation networks enable us to investigate the combinations between existing knowledge elements made by newly published papers. Co-citation is defined as the frequency with which two documents are cited together by other documents, changes in the co-citation model have been suggested to allow for a more objective way of simulating the knowledge structure of scientific knowledge [56] and co-citation network's structural features have also been used in publication's citation prediction. We construct a paper co-citation network for a certain research topic by using the bibliographic dataset. We also apply the structural variance analysis model of Citespace to analyse the co-citation network. The structural variation algorithm (SVA) is based on the co-citation network and emphasises the role and impact of novel recombination in creative thinking [57]. We calculate the modularity change rate (MCR), cluster linkage (CL) and centrality divergence (CD), three measures that characterise the degree of structural variation of the network.

2.5. Statistical testing

Citation network mapping is often equated with a visual representation of the scientific structure, yet the visual representation reflects only the layout and partitioning of bibliographic units, rather than the mathematical output behind the mapping [58]. For this concern, we perform a regression analysis to examine how the embedding values that characterise the structural features of papers relate to future citations of the papers. The embedding data trained by the struct2vec are used as a set of predictive variables to examine whether the network structure features captured by the network embedding model can predict the number of future citations of scientific publications. In our data, we count papers' citation counts and find that their general distribution follows an exponential pattern. We use Poisson regression models to test whether structural feature embedding of a paper's citation network predicts its citation counts, after conditioning on the numbers of authors, numbers of references and network metrics of structural variation that have been proved to have predictive effects on citation counts [57]. The Poisson regressions model number of citations associated with each paper. Our Poisson regressions take the following form

$$E(Citation_i|\mathbf{S_i},\ \mathbf{X_i}) = \exp(\mathbf{S_i}\beta + \ \mathbf{X_i}\gamma) \tag{1}$$

$Citation_i$ is the number of citations to paper i. The key independent variable of interest is $\mathbf{S_i}$ which is the structural features of the paper i derived from the embeddings trained by struc2vec. $\mathbf{X_i}$ is a vector of control variables that vary by

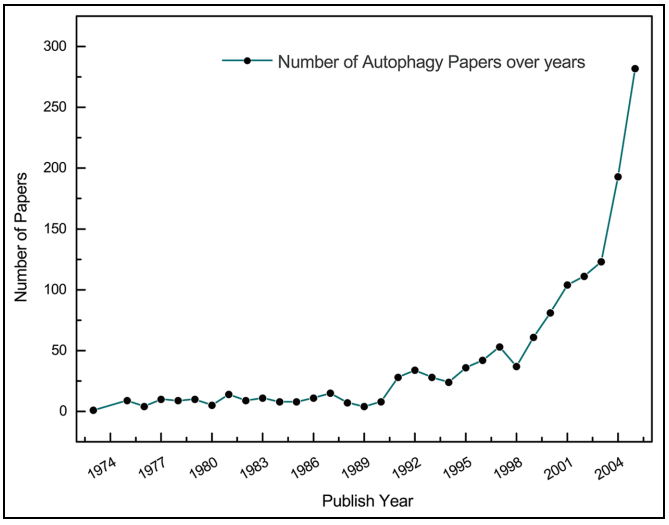

Figure 4. Autophagy papers published from 1973 to 2005.

regression specification, including *References*, *Authors*, *MCR*, *CL* and *CD*. Those five control variables have been found to be related to the dependent variable in our study. We report the models' effectiveness with the Akaike information criterion (AIC) [59] and pseudo-R^2.

3. Visual analytics

To explore the relationship between the citation network structure brought about by newly published scientific papers and their future citations, we applied our approach to a case study on the topic of autophagy.

3.1. Topic and data introduce

Figure 4 shows the publication status of 1343 papers on the topic of autophagy from 1973 to 2005. It can be seen from Figure 4 that the annual research results on cellular autophagy are very few before 1990, and the publication amount from 1990 to 1997 shows a stagnant upward trend until 2005 when there was a burst in the number of publications. After removing records with missing fields such as references, the number of citations, WOS ID and so on, we got 1180 papers for the network analysis. In this case study, the 'newly published' papers are those published in 2005, we investigated how papers published in 2005 were embedded into the historical citation network consisting of papers published from 1973 to 2004, that is considering 2005 as the publication year and 2019 as the observation year.

3.2. Citation network representation and clustering analysis

We model the citation relationship set N to generate the citation graph $G = (V, E)$, where V represents the set of vertices, each of which represents a paper, and E is the set of edges between vertices, each edge representing the citation relation between two papers. For a given graph $G = (V, E)$, the purpose of the network representation is to learn a mapping function $f = v_i \rightarrow y_i \in \mathbf{R}d$, where the dimension d of the vector is much smaller than the total number of nodes $|\mathbf{V}|$. The goal of the function f is to map vertex v_i to a low-dimensional space and to enable the network representation y_i to explicitly characterise vertex v_i in this space.

Poor interpretability is one of the drawbacks of deep learning at present [60]. For this reason, we use the t-SNE algorithm to downscale the embedding vectors into two dimensions and visualise them, and scatter plot results are shown in Figure 5. The dots in Figure 5 represent papers published in 2005, and we have coloured the dots according to the number of citations, where a redder node represents a higher number of citations for the paper and a bluer node indicates a lower number of citations. According to the features of the struct2vec model captured in the learning process, when the dots are clustered close to each other, it means that the paper represented by them has a similar structure in the citation network.

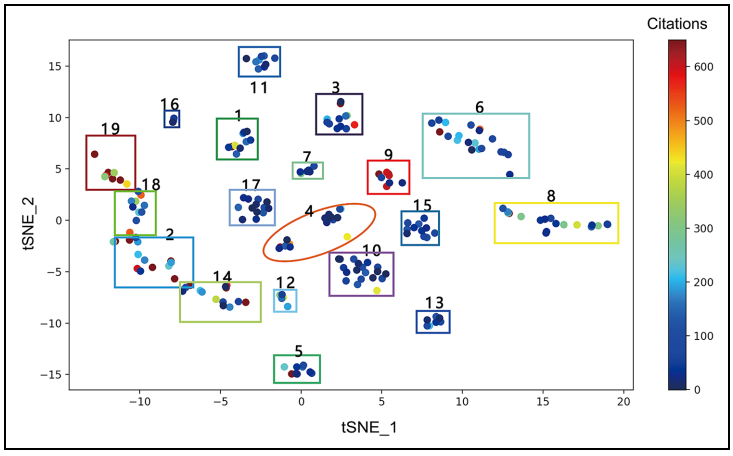

Figure 5. t-SNE embedding of a network representation of papers published in 2005. The colours of the dot (paper) mean the corresponding citation counts it earned until December 2019.

Table 1. Citation statistics for papers in 19 clusters.

Cluster id	Paper amount	Mean value	Standard deviation	Minimum	25% percentiles	50% percentiles	75% percentiles	Maximum
1	9	199.44	150.60	29	109	161	227	524
2	15	394.27	459.30	5	69	183	609	1515
3	12	196.58	256.48	4	42.25	59.5	230.25	826
4	15	181.33	311.53	4	14.5	30	176.5	1162
5	9	161.33	251.91	20	33	57	147	806
6	27	243.96	457.32	4	57.5	100	223	2430
7	5	563.80	786.02	118	173	267	297	1964
8	13	120.77	127.38	29	38	56	138	403
9	8	501.13	476.23	23	50	603	617.75	1441
10	19	71.68	95.25	0	11	47	98	417
11	9	89.11	66.93	5	30	69	139	184
12	6	80.00	106.39	21	23.5	34	67	293
13	7	83.00	69.33	12	28	85	109.5	209
14	10	172.40	222.68	7	38	92	169	706
15	11	65.45	38.74	13	37.5	66	89	130
16	4	149.50	188.18	10	38.5	82	193	424
17	13	45.38	50.83	0	11	27	66	148
18	14	437.07	364.09	40	186.75	250	698	1291
19	10	699.90	401.66	175	368.75	651.5	953.5	1337

We use the k-means algorithm to cluster the embedding values of all paper nodes. Through several comparative experiments on the k-means algorithm, it is found that clustering with the k value of 19 yielded the best results in terms of reducing the loss of original spatial clustering features. Table 1 reports the number of citations of the nodes in each of the 19 clusters, along with the mean, standard deviation and five-number summary for the citation counts of each cluster we labelled the clusters identified by k-means clustering by numbered rectangles or ovals in Figure 5. Note that there are more than one clusters in which the majority are highly-cited, mediumly-cited or lowly-cited papers. For example, cluster 9 and cluster 19 contain nodes of highly-cited papers, while the embedding values of these two clusters differ significantly, indicating that two different patterns of citing existing papers may explain why papers are highly-cited. It can be seen that papers in the same cluster tend to have a similar number of citations. This observation implies that paper nodes with similar network structure feature display patterns relevant to their future citation counts.

Table 1 shows a significant difference in citation counts between clusters. For example, the mean value of cluster 19 is 699.90, whereas cluster 17 shows a mean value of only 45.38. We visualise the distribution of paper citations in each cluster in a box-line graph in Figure 6. Using the length of each box as a guide, the distribution of data in clusters 2, 7

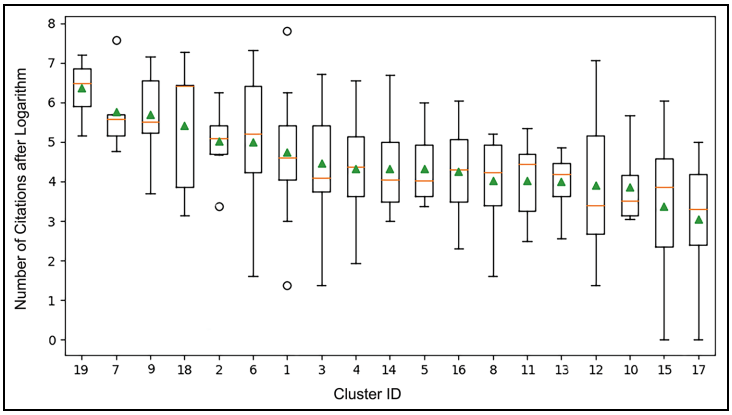

Figure 6. Box-line graph of citations for each cluster of papers.

and 19 is more concentrated, and the mean and median scores are higher, reflecting the more concentrated distribution of paper citations and higher citations score, while clusters such as 15 and 17 had lower mean, upper quartile and maximum values, suggesting that the citations in these clusters are less cited overall. When drawing the box-line graph, the citations are too disparate to allow for direct numerical comparison in box-plot form, so we take the logarithm (ln (2)) of the citation values for all papers.

3.3. Network visualisation

The embedding representation of the network structure lacks interpretability; to address it, we use both direct citation networks and co-citation networks to visually analyse the network patterns formed by papers within the same cluster.

3.3.1. Direct citation network construction and analysis. We conduct a visual analysis of direct citation networks that consisted of papers before 2005 and in 2005. The direct citation network forms a total of four communities with 1043 nodes and 11,806 edges. We group clusters into three clusters in terms of the citations received by the papers published in 2005 for each cluster: high-citation clusters, medium-citation clusters and low-citation clusters. For each cluster, four representative papers are selected to analyse the connections between papers published in 2005. We highlight the citation links of selected papers in yellow and identified different communities in the network with different colours.

Figure 7 shows the patterns of how papers published in 2005 from clusters 7, 9, 18 and 19 cited papers published before 2005. These papers in 2005 are highly-cited compared with the ones in other clusters. Although the patterns vary across the clusters, we can see certain common patterns. As can be seen from the citation network, high-citation papers generally tend to cite papers broadly across communities and have a large span between communities of papers published before 2005.

Figure 8 demonstrates the pattern of how papers published in 2005 from clusters 1, 5, 8 and 14 cited papers published before 2005. Overall, these papers are mediumly cited. The nodes for these papers in 2005 tend to be located in the middle of the community, which are less likely to combine old and new knowledge as extensively as those in Figure 7. These papers are more likely to cite papers that are in the same community but still cited papers across communities.

Figure 9 displays the patterns of how papers published in 2005 from clusters 11, 12, 15 and 17 cited papers published before 2005. The papers from these four clusters are relatively lowly-cited. It can be seen that lowly cited papers introduced fewer citation links, indicating that they cited fewer previous papers. One explanation is that it may be difficult for a paper that makes few citations to clearly illustrate its research basis, and this may limit the impact of the newly published paper.

In terms of the positions and connections of papers in the direct citation network, we qualitatively identified four common patterns from the visualisation results in Figures 7–9. Table 2 summarises these patterns.

Pattern 1 has been identified in only one highly-cited cluster (19). The high citation counts of this pattern may be explained from two perspectives. On the one hand, highly-cited papers tend to cite more unusual combinations of prior work, leading to higher value and citation. On the other hand, in terms of the cohesiveness of the community genre [61], papers at the margins of the community genre tend to be more valuable and novel than less marginal papers of

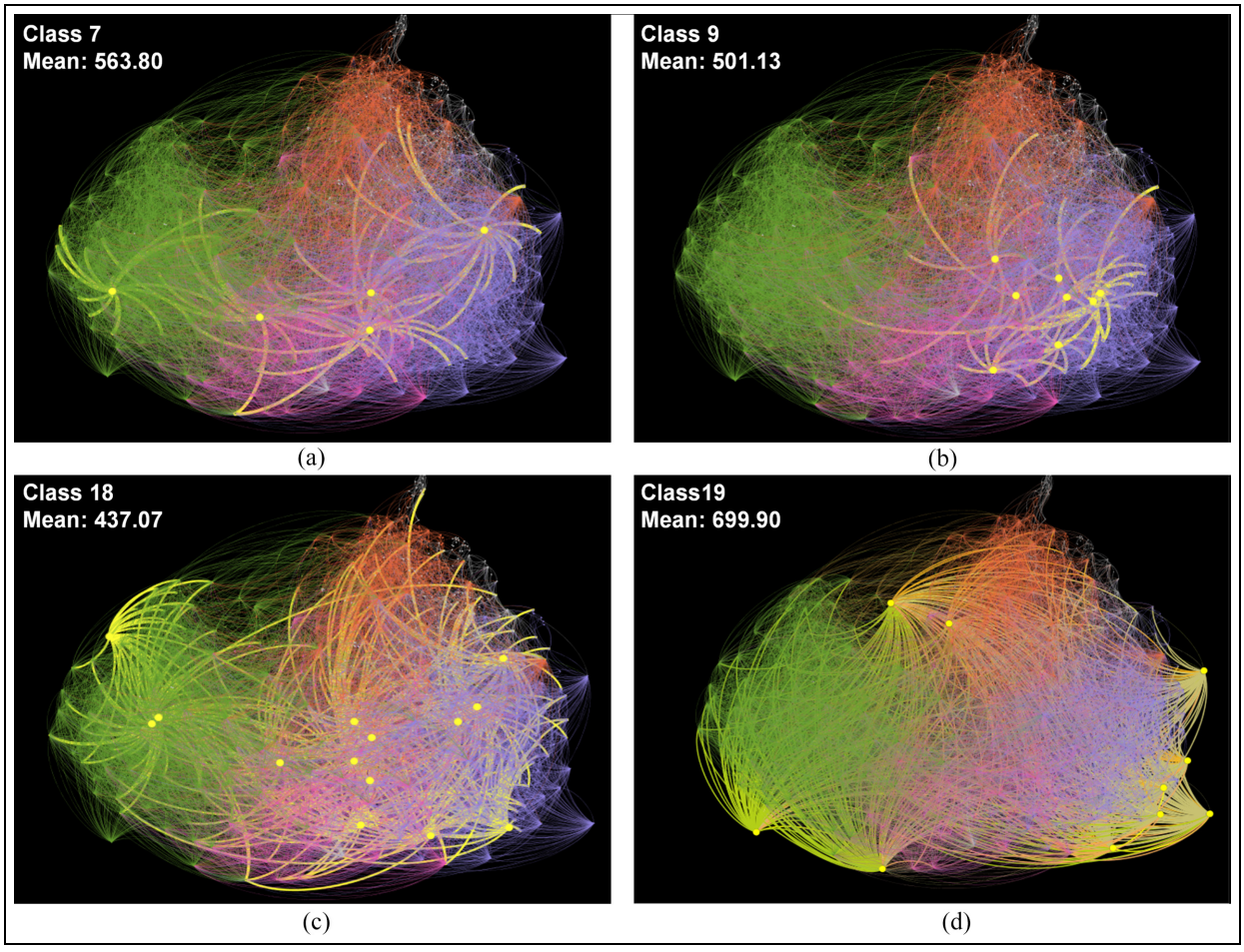

Figure 7. Direct citation networks of high-citation clusters. (a)–(d) Visualisations of four representative clusters: 7, 9, 18 and 19.

comparable citation count. This is due to the semi-marginal status of the association, which provides the researcher with a knowledge base without limiting the researcher's perspective on the innovation paradigm. Instead, more cross-disciplinary innovations and combinations are allowed.

Pattern 2 can be found in highly-cited cluster 18, along with moderately cited clusters 8, 1 and 5. Most of these papers belong to a specific existing knowledge community, but they are also connected to other knowledge communities, that is, most of the papers' references are from the cluster the papers belong to, but papers from various clusters are also cited. We consider the papers in these clusters not to be as radically novel as those in cluster 19. In Kuhn's [62] view of the structure of scientific revolutions, these papers might be the regular work of scientists theorising, observing and experimenting within a settled paradigm or explanatory framework instead of paradigm-shifting discoveries.

Pattern 3 includes the highly-cited clusters 7 and 12, in which papers are connected to a few existing communities but do not strongly belong to a specific community. They usually lie in boundary zones between communities. They exhibit either high or low, but not moderate, citation counts. In light of the theory of structural holes [63], these papers represent boundary zones, the potential location of holes in the existing intellectual structure. Bridging these gaps is a high-risk and high-return activity. In positioning their research in the boundary zones, scientists take the risk of garnering few citations in exchange for the chance of a high return (i.e. many citations).

In pattern 4, found in clusters 9, 11, 14, 15 and 17, papers belong to a specific existing community and mostly connect to other works within this community. When these papers refer to combinations of research from different communities, future available citations are still typically lower due to the excessive span of references and lack of relevance, reflecting the lack of a strong research base.

3.3.2. Co-citation network construction and analysis. In this section, we examine how each new paper connects to existing knowledge via co-citation networks and present a visual presentation. More specifically, we explored, from a visual

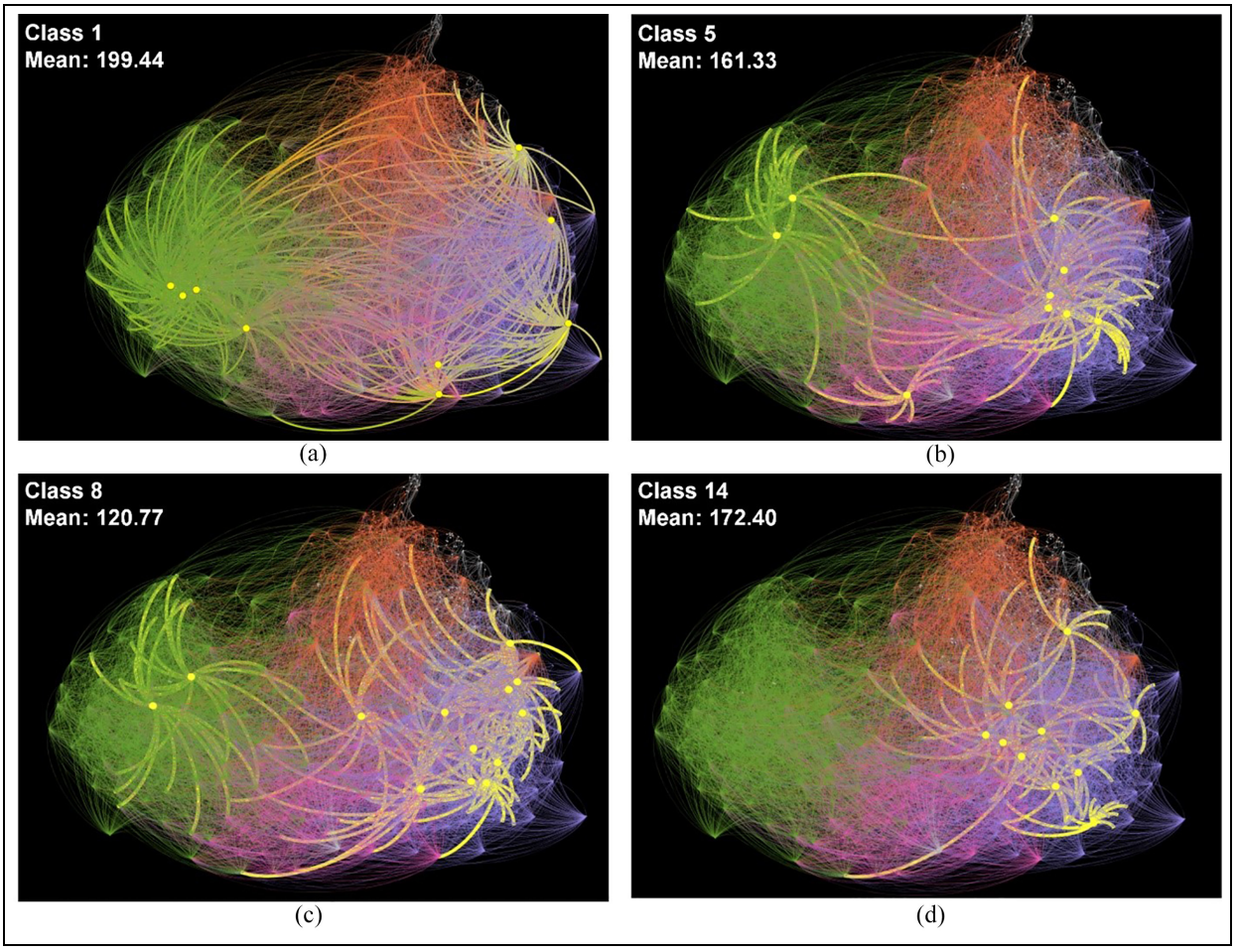

Figure 8. Direct citation networks of medium-citation clusters. (a)–(d) Visualisations of four representative clusters: 1, 5, 8 and 14.

perspective, how autophagy papers published in 2005 were embedded as new content into the pre-2005 citation network. First, we import 1343 relevant papers published from 1973 to 2005 into Citespace software and perform network relationship computation using the Kamada and Kawa layout algorithm; then the Spectral Clustering algorithm was used for association clustering and division to obtain 16 communities, shown in Figure 10(a). New knowledge connections made by a certain paper are shown in bold in Figure 10(b)–(d).

Next, we constructed the SVA visualisation diagram of the paper of each cluster and selected clusters 19, 15 and 1 as high-, medium- and low-citation examples for presentation and analysis. For each cluster, we selected one representative paper to analyse the new knowledge created by the paper that is characterised by new combinations made by the paper. Figure 10(b) shows the linkages of the paper (WOS id: 000238461400001) in the high-citation cluster 19, this article brings many new connections between communities and these links are present as a dense community structure. Figure 10(c) shows the paper (WOS id: 000233992700002) in cluster 1 in which most of the papers are mediumly cited. We can see from the new connections in Figure 10(c) that it brings some new connections to the co-citation network but does not perform as significantly as the paper in cluster 19 in terms of the number and community span of these new linkages. Figure 10(d) shows the paper (WOS id: 000226927100005) in the low-citation cluster 15 and the paper has few new connections in the co-citation network – in some cases none – with little change to the network structure.

The structural changes brought by a paper can be explained by structural variation metrics. Table 3 shows the values of the structural variation of the three papers mentioned above. For the MCR, the higher values indicate that the addition of these papers changes the communities of the original network to a greater extent. The higher the CL value, the larger the change in the degree of linkage of the original community due to the addition of the paper. The CD scale is used to measure the magnitude of the mediated change in centrality induced by a paper's addition, the bigger the value means the larger impact on the network structure. It can be seen that the three values of most cited papers are the largest, and

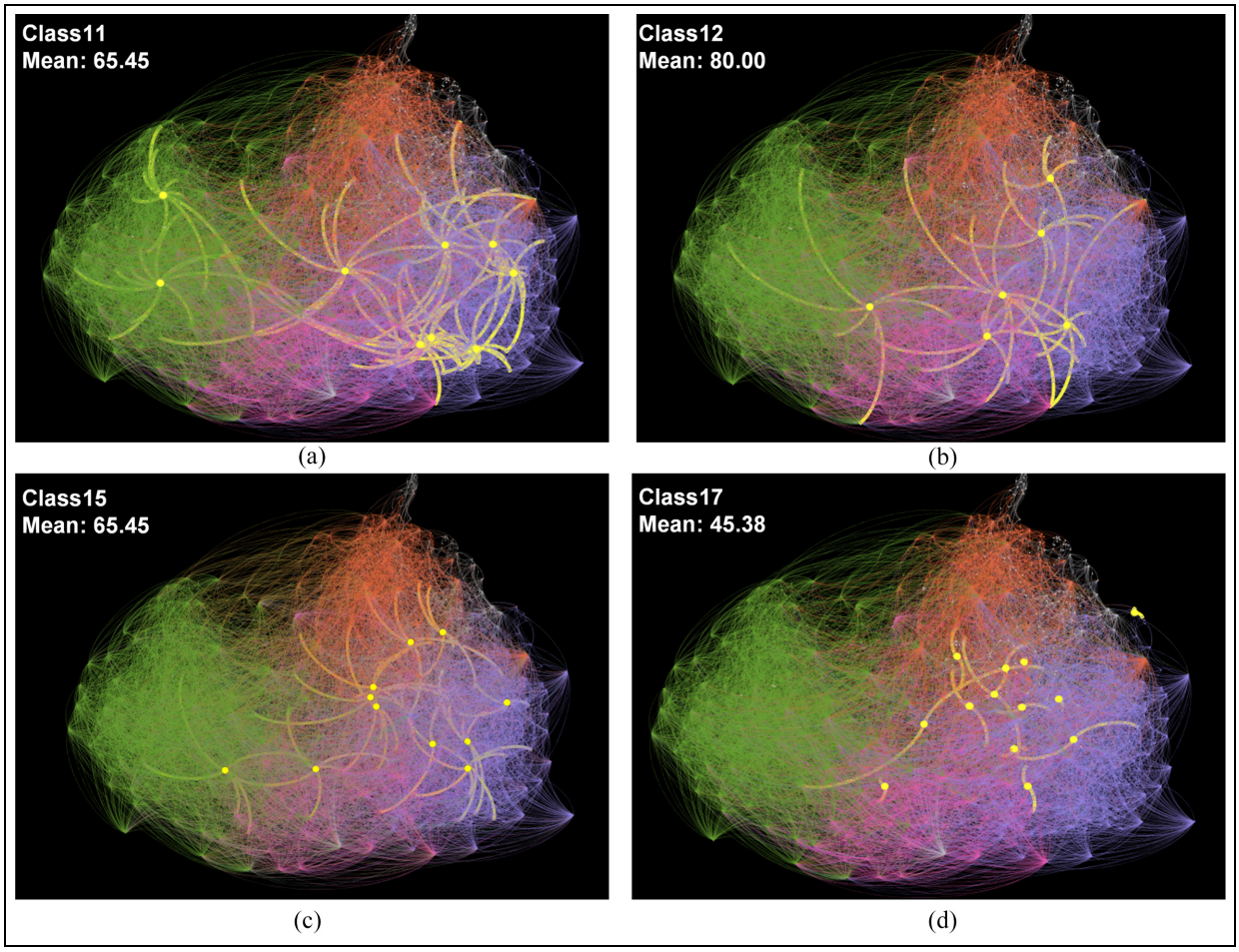

Figure 9. Direct citation networks of four low-citation clusters. (a)–(d) Visualisations of four representative clusters: 11, 12, 15 and 17.

the fewer cited, the smaller each value is. The above analysis tentatively demonstrates that these three values have a relationship with the future citations of the papers, which is consistent with Chen's [57] study. Later, these three metrics will be used as predictor variables of statistical models to predict future citations of papers.

4. Statistical testing

Since it is hard to understand the regression results of 128-dimensional embedding values that characterise the local network structure, we randomly choose 10 embedding vectors from the 128 vectors and set them as the independent variable of our regression model. Furthermore, we introduce two Nobel Prize-winning topics – *Helicobacter pylori* and Graphene for statistical testing, in addition to the first topic – Autophagy.

Poisson regression results for the first topic are shown in Table 4; each column reports one Poisson regression model. There are 216 observations and the p-value of seven models is 0.0000 for each model, indicating that those models are suitable for this topic's data and are statistically significant. Column 1 estimates the effect of having 10 randomly chosen structural embeddings as our independent variable, and the result shows that all those 10 embeddings are significant with $p < 0.001$. Columns $2-6$ estimate the effects of the references, authors, MCR, CL and CD as control variables and use the structural embeddings as the focal independent variable. In terms of the model fitting effect, we use the indicator pseudo-R^2 to analyse the overall fitting effect of the model. When a statistically significant variable enters the regression model, the indicator increases, and vice versa. It can be seen from the model in Table 4 that the control variables we selected have a predictive effect on the model, as indicated by the increase in pseudo-R^2 and AIC values. However, the degree to which those control variables contribute to model explainability varies and is slight. Among them, the CD has

Table 2. Four patterns of connecting new papers to existing knowledge.

Pattern	Cluster	Position	Connections	Citations
Pattern 1	19	Do not belong to any community	Extensively connect to various communities	Very likely to be highly-cited
Pattern 2	1, 5, 8, 18	Belong to a certain community	Mainly connect to the community they belong to but also connect to other communities	Likely to be highly-cited or mediumly-cited
Pattern 3	7, 12	Do not belong to a certain community; lie in boundary zones between communities	Almost equally connect to various communities	Highly-cited or lowly-cited
Pattern 4	9, 11, 14, 15, 17	Belong to a certain community	Mostly connect to the community they belong to, rarely to other communities	Likely to be mediumly-cited and lowly-cited

Table 3. Structural variation values of three demo papers.

WOS: ID	Modularity change rate	Cluster linkage	Centrality divergence	Global citations
000238461400001	8.1129	1.4495	0.0953	412
000233992700002	3.4837	0.4855	0.0285	115
000226927100005	0.1157	0.0163	0.0013	13

WOS: Web of Science.

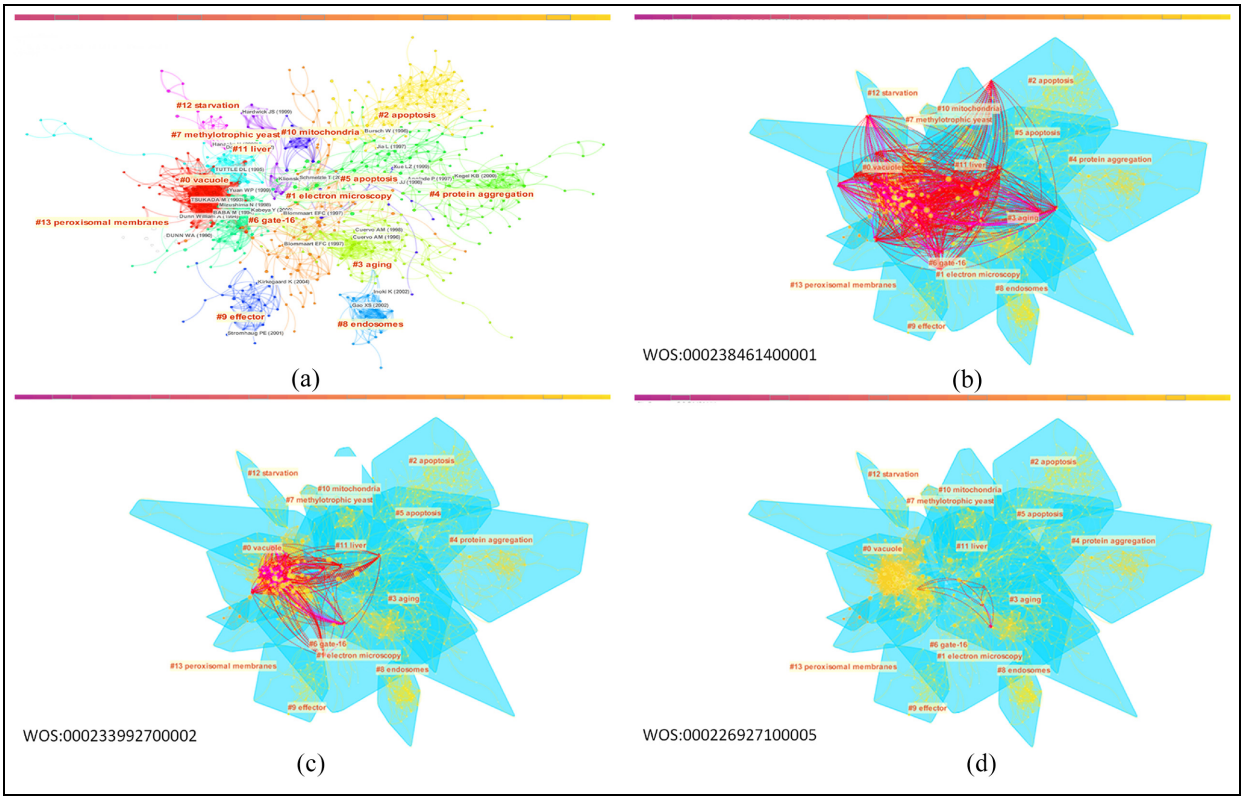

Figure 10. Visualisation of the structural variation of three demo papers (a) indicates the 16 communities, (b) shows a paper in the high-citation cluster, (c) shows a paper in the middle-citation cluster and (d) shows a paper in the low-citation cluster.

Table 4. Regression results for the topic of Autophagy.

Paper citations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EM1	0.57*** (42.43)	0.38*** (27.21)	0.58*** (43.13)	0.48*** (34.62)	0.43*** (31.10)	0.38*** (27.59)	0.41*** (27.99)
EM2	1.12*** (63.10)	1.04*** (57.97)	1.03*** (58.85)	0.86*** (43.53)	0.77*** (38.95)	0.83*** (45.41)	0.88*** (43.82)
EM3	− 0.098*** (− 6.82)	0.045** (2.98)	− 0.088*** (− 6.23)	− 0.075*** (− 5.07)	− 0.048** (− 3.24)	− 0.017 (− 1.14)	0.0057 (0.36)
EM4	0.11*** (8.24)	0.00097 (0.07)	0.11*** (8.38)	− 0.016 (− 1.20)	− 0.068*** (− 5.03)	− 0.21*** (− 16.23)	− 0.20*** (− 15.21)
EM5	− 0.22*** (− 14.05)	− 0.14*** (− 8.85)	− 0.29*** (− 18.94)	− 0.13*** (− 7.85)	− 0.083*** (− 5.08)	0.18*** (10.39)	0.12*** (7.15)
EM6	0.64*** (36.83)	0.58*** (31.64)	0.49*** (28.36)	0.72*** (39.79)	0.75*** (40.88)	0.91*** (46.45)	0.66*** (33.92)
EM7	− 0.22*** (− 16.24)	− 0.26*** (− 18.62)	− 0.23*** (− 16.89)	− 0.20*** (− 14.00)	− 0.17*** (− 12.37)	− 0.15*** (− 10.49)	− 0.22*** (− 14.77)
EM8	0.48*** (34.93)	0.50*** (35.51)	0.46*** (33.39)	0.43*** (30.59)	0.42*** (30.40)	0.26*** (18.42)	0.16*** (10.61)
EM9	0.18*** (12.65)	0.15*** (10.30)	0.071*** (5.06)	0.18*** (12.28)	0.16*** (11.02)	0.073*** (4.97)	− 0.071*** (− 4.84)
EM10	0.66*** (43.86)	0.45*** (28.34)	0.67*** (44.55)	0.58*** (37.75)	0.55*** (35.29)	0.33*** (20.59)	0.31*** (18.60)
References		0.0056*** (55.95)					0.0018*** (14.67)
Authors			0.063*** (46.54)				0.085*** (62.45)
MCR				0.038*** (32.62)			− 0.026** (− 2.93)
CL					0.96*** (44.61)		− 0.091 (− 0.49)
CD						11.4*** (96.41)	13.9*** (52.57)
Pseudo-R^2	0.171	0.204	0.198	0.183	0.194	0.279	0.338
AIC	62,955.9	60,411.5	60,885.7	61,987.3	61,166.1	54,766.0	50,299.5
Prob > chi^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	216	216	216	216	216	216	216

MCR: modularity change rate; CL: cluster linkage; CD: centrality divergence; AIC: Akaike information criterion.

Unit of analysis: paper. The dependent variable is the number of paper's citation count; t statistics in parentheses; *N*, the number of observations; Em1 − Em10 represent embedding values.

*$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$.

the stronger ability to predict a paper's citation counts. For a more complete representation of the predictive power of our structural embeddings, we report the prediction results of the model with 128 embeddings as independent variables in Annex 1. When the 128 embeddings together represent the structural features of the network, our model with only independent variables (column 1) shows stronger predictability. As shown in Annex 1, the model in column 1 gains the pseudo-R^2 0.767, when introducing the other control variables, the prediction effect of our model has little or no growth. This indicates that the addition of other control variables does not contribute much to the predictability of the model, and our independent variables explain most of the predictability of the model.

Tables 5 and 6 report the results of Poisson regression models for the second topic Helicobacter and the third topic Graphene; 245 and 786 papers were observed on these two topics, respectively. The p-values for all models in both topics are less than 0.001, indicating that our models also fit those two topics. The pseudo-R^2 values increased with the incorporation of the control variables, and the data for both topics simultaneously illustrated the positive predictive effect of our embedding on citation counts. The decrease in the AIC values for both topics suggests that the addition of our control variables slightly improved the predictive ability of the model.

In summary, the regression results for the three different topics all show that our model is statistically significant, although the predictions are not entirely consistent due to differences in publication volume and citation bias in different fields. Furthermore, we observed that adding control variables after the embeddings enhanced the model's fit, indicating that these measurements also explain the dependent variables in our experiments. Among them, the two bibliometric indicators, the number of references and the number of authors are not as effective as the three structural variation indicators, in which MCR has the worst explanatory power and CL has the strongest explanatory power. The experimental results show that the network structure embedding introduced in this article is effective in predicting the number of citations. The framework proposed in this article is also reasonable to explain the relationship between citation network structure and paper citations.

5. Discussion and conclusion

5.1. Potential influence of publications revealed by citation network structure

This article presents a novel framework to examine the relationship between citation network embeddings and the future citation counts of papers. We propose the visualisation and analysis of two distinct types of networks to better illustrate network embeddings. The first type is a direct citation network, from which we identify four patterns of connections between newly published papers and existing knowledge. The second type is a co-citation network, where we quantify three structural variation indicators for new papers based on previous research findings.

Furthermore, we introduce an innovative combination of three types of factors – embeddings, bibliometric data and network structural variation values – into the regression model to validate our findings. Descriptive statistics on various data demonstrate that our model is statistically significant, and embeddings are predictive of the impact of paper citations. We address the question of whether the similarity in network structure can be utilised to predict the citations of papers. The representations of newly published papers, as learned by the graph representation learning model, are determined by how these papers are connected with existing knowledge. We also find that the graph-embedding algorithm functions like a well-designed black box. Although graph representation learning techniques can provide rich information about network structure, the hidden features learned by these techniques still lack interpretability.

In addition, we qualitatively explore the structural features of papers in citation networks through visual analysis of both the direct citation network and the co-citation network. For papers published within the same time frame on a specific topic, we discovered that the greater the change induced in the original citation network, the more innovative and influential the paper potentially becomes, and the more citations it may receive in the future. Through a visual analysis of direct citation networks, we identify four patterns in how papers cite prior works, which can, in turn, explain their future citation counts. For instance, highly-cited papers are more likely to appear at the edge of a region in the network structure and have more global associations with the entire region. Intriguingly, for collections of high-, medium- or low-citation papers, different patterns were found to explain the citations for each group.

The co-citation network analysis yields results that are consistent with those of the direct citation network analysis. This indicates that the way a new paper connects with existing knowledge elements is related to how papers establish new links between existing knowledge elements. For example, for a high-citation cluster, specific traits can be seen in its original citation network, for example highly-cited papers [64,65] usually tend to assemble a larger base of previous work, which may represent the discovery of more unusual combinations that render the paper's contribution more valuable and innovative.

Table 5. Regression results for the topic of Helicobacter.

Paper citations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EM1	−0.82*** (−24.71)	−0.35*** (−10.06)	−0.95*** (−28.55)	−0.17*** (−4.83)	−0.11** (−3.22)	−0.83*** (−25.09)	−0.34*** (−9.18)
EM2	0.49*** (15.87)	0.37*** (11.75)	0.54*** (17.44)	0.39*** (12.48)	0.42*** (13.26)	0.48*** (15.48)	0.54*** (16.49)
EM3	1.29*** (30.54)	0.84*** (18.49)	1.13*** (26.26)	0.63*** (13.37)	0.56*** (11.83)	1.16*** (27.15)	0.35*** (7.34)
EM4	0.33*** (10.73)	−0.20*** (−6.02)	0.41*** (13.12)	−0.26*** (−7.59)	−0.27*** (−7.79)	0.20*** (6.23)	−0.16*** (−4.39)
EM5	0.69*** (16.89)	0.27*** (6.30)	0.75*** (17.91)	0.073 (1.66)	0.15*** (3.54)	0.42*** (9.81)	0.40*** (8.23)
EM6	0.034 (0.83)	−0.27*** (−6.31)	0.31*** (7.35)	−0.24*** (−5.38)	−0.29*** (−6.55)	−0.021 (−0.50)	−0.027 (−0.60)
EM7	−0.11*** (−3.43)	0.024 (0.71)	−0.049 (−1.56)	0.054 (1.56)	0.029 (0.82)	−0.12*** (−3.90)	−0.014 (−0.37)
EM8	0.99*** (28.46)	1.12*** (29.42)	0.91*** (26.09)	0.99*** (25.66)	1.04*** (26.79)	0.85*** (23.68)	0.92*** (23.21)
EM9	0.071 (1.86)	−0.29*** (−7.49)	0.26*** (6.63)	−0.38*** (−9.41)	−0.39*** (−9.89)	0.049 (1.27)	−0.11** (−2.72)
EM10	0.55*** (17.55)	0.026 (0.73)	0.52*** (16.30)	−0.045 (−1.28)	−0.099** (−2.82)	0.39*** (11.88)	−0.22*** (−6.04)
References		0.0077*** (40.68)					−0.0000088 (−0.02)
Authors			0.043*** (24.92)				0.043*** (24.85)
MCR				0.10*** (45.01)			−0.33*** (−17.67)
CL					4.66*** (49.52)		18.3*** (22.64)
CD						3.60*** (22.42)	3.37*** (15.39)
Pseudo-R^2	0.213	0.290	0.238	0.312	0.330	0.235	0.385
AIC	16,150.3	14,564.7	15,648.5	14,113.1	13,745.6	15,694.6	12,637.5
Prob > chi^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	245	245	245	245	245	245	245

MCR: modularity change rate; CL: cluster linkage; CD: centrality divergence; AIC: Akaike information criterion.

Unit of analysis: paper. The dependent variable is the number of paper's citation count; t statistics in parentheses; *N*, the number of observations; Em1 − Em10 represent embedding values.

*$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$.

Table 6. Regression results for the topic of Graphene.

Paper citations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EM1	−0.52*** (−50.84)	−0.52*** (−50.98)	−0.38*** (−35.91)	−0.53*** (−51.49)	−0.50*** (−48.90)	−0.52*** (−50.83)	−0.40*** (−37.83)
EM2	−1.36*** (−112.31)	−1.37*** (−113.64)	−1.47*** (−109.40)	−1.43*** (−118.98)	−1.45*** (−119.64)	−1.36*** (−112.23)	−1.69*** (−126.11)
EM3	−0.38*** (−32.13)	−0.39*** (−32.91)	−0.34*** (−28.57)	−0.38*** (−32.02)	−0.38*** (−32.12)	−0.39*** (−32.12)	−0.39*** (−33.23)
EM4	2.69*** (164.73)	2.67*** (163.12)	1.91*** (114.62)	2.61*** (159.74)	2.63*** (160.39)	2.69*** (164.52)	1.70*** (101.89)
EM5	2.29*** (172.26)	2.30*** (173.29)	2.37*** (181.49)	2.33*** (176.87)	2.33*** (177.10)	2.29*** (171.68)	2.50*** (194.89)
EM6	0.35*** (24.89)	0.36*** (25.94)	0.34*** (23.09)	0.33*** (24.16)	0.34*** (24.60)	0.35*** (24.90)	0.33*** (23.93)
EM7	0.80*** (63.84)	0.81*** (64.36)	0.41*** (30.99)	0.81*** (65.22)	0.79*** (63.39)	0.80*** (63.45)	0.33*** (25.52)
EM8	0.12*** (11.25)	0.13*** (11.50)	−0.014 (−1.28)	0.13*** (11.43)	0.11*** (10.47)	0.12*** (11.26)	0.022* (2.11)
EM9	−1.06*** (−75.37)	−1.04*** (−74.22)	−1.25*** (−84.77)	−1.03*** (−74.49)	−1.00*** (−72.09)	−1.06*** (−75.21)	−1.12*** (−76.82)
EM10	−1.79*** (−135.62)	−1.79*** (−135.45)	−1.30*** (−96.95)	−1.85*** (−140.03)	−1.84*** (−139.24)	−1.79*** (−135.60)	−1.40*** (−106.25)
References		0.0018*** (15.82)					0.0028*** (17.09)
Authors			0.23*** (251.35)				0.25*** (265.35)
MCR				0.072*** (46.38)			0.13*** (34.82)
CL					4.81*** (44.68)		3.83*** (16.05)
CD						−0.097 (−0.87)	−5.55*** (−33.64)
Pseudo-R^2	0.198	0.198	0.294	0.201	0.201	0.198	0.310
AIC	474,745.4	474,513.5	417,704.3	472,895.9	473,010.5	474,746.6	408,292.5
Prob > chi^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	786	786	786	786	786	786	786

MCR: modularity change rate; CL: cluster linkage; CD: centrality divergence; AIC: Akaike information criterion.

Unit of analysis: paper. The dependent variable is the number of paper's citation count; t statistics in parentheses; *N*, the number of observations; Em1 − Em10 represent embedding values.

*$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$.

The visualisation results clarify why papers in a certain cluster with similar structural features have comparable citation impacts in the future and assist in identifying the structural features in question. Furthermore, the analysis results reveal that newly published papers with similar structural features in the citation network exhibit similar citation impacts in the future. Descriptive statistics demonstrate that our regression model is predictive, suggesting that citation network embedding captures the features associated with papers' citation counts. The result also implies that the connection of a new paper to the existing body of knowledge exerts a predictive influence on the citation impact of that paper.

5.2. Theoretical implications and practical implications

From both theoretical and practical perspectives, our framework offers significant insights into the research field by seamlessly integrating cluster analysis, visual analysis and statistical analysis. Theoretically, our framework advances the understanding of citation networks by identifying groups with similar characteristics through the clustering process. The visual analysis further enhances this understanding by providing an intuitive perception and comprehension of these shared features. This combination of methods contributes to the development of a more comprehensive understanding of the underlying patterns and structures in citation networks. Practically, our framework has the potential to improve the evaluation and prediction of citation impacts. By employing descriptive statistical analysis, we can validate our assumptions and assess the effectiveness of the features identified through representation learning. This validation process ensures that the identified features are indeed relevant and useful for predicting citation impacts. In conclusion, the integration of these cutting-edge data analysis techniques and statistical methods holds great promise for enhancing our understanding of the scientific development process from both theoretical and practical perspectives.

5.3. Limitations and future work

In addition to the contributions mentioned above, this study also identifies areas with significant opportunities for improvement and future work. On the one hand, to investigate the structural paradigm discussed here, we selected the Nobel Prize-winning themes as the subject. It remains to be seen whether different patterns in citation networks exist for more general themes. On the other hand, it is shown that struc2vec models can characterise the network consisting of more than a thousand nodes in this article; however, the applicability of such models to characterise larger networks (e.g. citation networks consisting of millions of nodes) is a question to be further investigated. In a follow-up study, we plan to conduct our research on a broader range of topics and examine how these network structural features change over time. Our study may offer a method for applying the graph representation learning approach to citation network analysis and the larger process of scientific development.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This work is supported by the Key Projects of the National Natural Science Foundation of China(Grant No. 72234005).

ORCID iDs

Zhuoran Luo https://orcid.org/0000-0003-0677-8350
Jiangen He https://orcid.org/0000-0002-3950-6098
Yuqi Wang https://orcid.org/0000-0001-9704-4672
Wei Lu https://orcid.org/0000-0002-0929-7416

References

[1] Fortunato S, Bergstrom CT, Borner K et al. Science of science. *Science* 2018; 359: eaao0185.
[2] Weis JW and Jacobson JM. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nat Biotechnol*. Epub ahead of print 17 May 2021. DOI: 10.1038/s41587-021-00907-6.
[3] Uzzi B, Mukherjee S, Stringer MJ et al. Atypical combinations and scientific impact. *Science* 2013; 342: 468–472.
[4] Kim D, Cerigo DB, Jeong H et al. Technological novelty profile and invention's future impact. *EPJ Data Sci* 2016; 5: 8.

[5] Bornmann L. How much does the expected number of citations for a publication change if it contains the address of a specific scientific institute? A new approach for the analysis of citation data on the institutional level based on regression models. *J Assoc Inf Sci Tech* 2016; 67: 2274–2282.
[6] Merton RK. The Matthew effect in science. The reward and communication systems of science are considered. *Science* 1968; 159: 56–63.
[7] Lu C, Ding Y and Zhang C. Understanding the impact change of a highly cited article: a content-based citation analysis. *Scientometrics* 2017; 112: 927–945.
[8] Penner O, Pan RK, Petersen AM et al. On the predictability of future impact in science. *Sci Rep* 2013; 3: 3052.
[9] Abramo G, D'Angelo CA and Felici G. Predicting publication long-term impact through a combination of early citations and journal impact factor. *J Informetr* 2019; 13: 32–49.
[10] Acuna DE, Allesina S and Kording KP. Predicting scientific success. *Nature* 2012; 489: 201–202.
[11] Yu T and Yu G. Features of scientific papers and the relationships with their citation impact. *Malays J Libr Inf Sc* 2017; 19: 37–50.
[12] Peng TQ and Zhu JJ. Where you publish matters most: a multilevel analysis of factors affecting citations of internet studies. *J Am Soc Inf Sci Tec* 2012; 63: 1789–1803.
[13] Klimek P, Jovanovic AS, Egloff R et al. Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks. *Scientometrics* 2016; 107: 1265–1282.
[14] Akella AP, Alhoori H, Kondamudi PR et al. Early indicators of scientific impact: predicting citations with altmetrics. *J Informetr* 2021; 15: 101128.
[15] Ajiferuke I and Famoye F. Modelling count response variables in informetric studies: comparison among count, linear, and log-normal regression models. *J Informetr* 2015; 9: 499–513.
[16] Onodera N and Yoshikane F. Factors affecting citation rates of research articles. *J Assoc Inf Sci Tech* 2015; 66: 739–764.
[17] Glänzel W and Schubert A. Predictive aspects of a stochastic model for citation processes. *Inform Process Manag* 1995; 31: 69–80.
[18] Bornmann L and Daniel HD. What do citation counts measure? A review of studies on citing behavior. *J Doc* 2008; 64: 45–80.
[19] Wang D, Song C and Barabási A-L. Quantifying long-term scientific impact. *Science* 2013; 342: 127–132.
[20] Bai X, Zhang F and Lee I. Predicting the citations of scholarly paper. *J Informetr* 2019; 13: 407–418.
[21] Fu L and Aliferis C. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics* 2010; 85: 257–270.
[22] Wang J, Zhang F, Li Y et al. Attention-based multi-fusion method for citation prediction. In: Pan JS, Li J, Tsai PW et al. (eds) *Advances in intelligent information hiding and multimedia signal processing*. Singapore: Springer, 2020, pp. 315–322.
[23] Zhu XP and Ban Z. Citation count prediction based on academic network features. In: *Proceedings of the 2018 IEEE 32nd international conference on advanced information networking and applications (AINA)*, Kraków, 16–18 May 2018, pp. 534–541. New York: IEEE.
[24] Abrishami A and Aliakbary S. Predicting citation counts based on deep neural network learning techniques. *J Informetr* 2019; 13: 485–499.
[25] Ruan X, Zhu Y, Li J et al. Predicting the citation counts of individual papers via a BP neural network. *J Informetr* 2020; 14: 101039.
[26] Dalle Lucca Tosi M and Dos Reis JC. Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs. *J Inf Sci* 2020; 48: 71–89.
[27] Yan R, Tang J, Liu X et al. Citation count prediction: learning to estimate future citations for literature. In: *Proceedings of the 20th ACM international conference on information and knowledge management*, Glasgow, 24–28 October 2011, pp. 1247–1252. New York: ACM.
[28] Yan R, Huang C, Tang J et al. To better stand on the shoulder of giants. In: *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries*, Washington, DC, 10–14 June 2012, pp. 51–60. New York: ACM.
[29] Chakraborty T, Kumar S, Goyal P et al. Towards a stratified learning approach to predict future citation counts. In: *Proceedings of the IEEE/ACM joint conference on digital libraries*, London, 8–12 September 2014, pp. 351–360. New York: IEEE.
[30] Huang S, Huang Y, Bu Y et al. Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Inform Process Manag* 2022; 59: 102799.
[31] Davletov F, Aydin AS and Cakmak A. High impact academic paper prediction using temporal and topological features. In: *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, Shanghai, China, 3–7 November 2014, pp. 491–498. New York: ACM.
[32] Du W, Li Z and Xie Z. A modified LSTM network to predict the citation counts of papers. *J Inf Sci*. Epub ahead of print 8 August 2022. DOI: 10.1177/01655515221111000.
[33] Shi F, Foster JG and Evans JA. Weaving the fabric of science: dynamic network models of science's unfolding structure. *Soc Networks* 2015; 43: 73–85.
[34] Shibata N, Kajikawa Y and Matsushima K. Topological analysis of citation networks to discover the future core articles. *J Assoc Inf Sci Tech* 2007; 58: 872–882.

[35] Min C, Bu Y, Wu D et al. Identifying citation patterns of scientific breakthroughs: a perspective of dynamic citation process. *Inform Process Manag* 2021; 58: 102428.
[36] Min C, Bu Y and Sun J. Predicting scientific breakthroughs based on knowledge structure variations. *Technol Forecast Soc* 2021; 164: 120502.
[37] Zhang D, Yin J, Zhu X et al. Network representation learning: a survey. *IEEE T Big Data* 2020; 6: 3–28.
[38] Bhagat S, Cormode G and Muthukrishnan S. Node classification in social networks. In: Aggarwal C (ed.) *Social network data analytics*. Boston, MA: Springer, 2011, pp. 115–148.
[39] Wang Z, Chen C and Li W. Predictive network representation learning for link prediction. In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, Tokyo, Japan, 7–11 August 2017, pp. 969–972. New York: ACM.
[40] Tang J, Qu M, Wang M et al. LINE: large-scale information network embedding. In: *Proceedings of the 24th international conference on World Wide Web*, Florence, 18–22 May 2015, pp. 1067–1077. Geneva: International World Wide Web Conferences Steering Committee.
[41] Ganguly S, Gupta M, Varma V et al. Author2Vec: learning author representations by combining content and link information. *In: Proceedings of the 25th international conference companion on World Wide Web, Montréal, QC, Canada*, 11–15 April 2016, pp. 49–50. Geneva: International World Wide Web Conferences Steering Committee.
[42] Tian H and Zhuo HH. Paper2vec: citation-context based document distributed representation for scholar recommendation, 2017, https://arxiv.org/abs/1703.06587
[43] Zhang J. Research collaboration prediction and recommendation based on network embedding in co-authorship networks. *Proc Assoc Inf Sci Technol* 2017; 54: 847–849.
[44] Lu C, Zhang Y, Ahn Y-Y et al. Co-contributorship network and division of labor in individual scientific collaborations. *J Assoc Inf Sci Tech* 2020; 71: 1162–1178.
[45] Zhang Y, Zhao F and Lu J. P2V: large-scale academic paper embedding. *Scientometrics* 2019; 121: 399–432.
[46] He J and Chen C. Predictive effects of novelty measured by temporal embeddings on the growth of scientific literature. *Front Res Metr Anal* 2018; 3: 9.
[47] Hu K, Luo Q, Qi K et al. Understanding the topic evolution of scientific literatures like an evolving city: using Google Word2Vec model and spatial autocorrelation analysis. *Inform Process Manag* 2019; 56: 1185–1203.
[48] Kleminski R, Kazienko P and Kajdanowicz T. Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification. *J Inf Sci* 2020; 48: 349–373.
[49] Xu H, Luo R, Winnink J et al. A methodology for identifying breakthrough topics using structural entropy. *Inform Process Manag* 2022; 59: 102862.
[50] Shen Z, Chen F, Yang L et al. Node2vec representation for clustering journals and as a possible measure of diversity. *J Data Inf Sci* 2019; 4: 79–92.
[51] Ribeiro LFR, Saverese PHP and Figueiredo DR. *struc2vec*: learning node representations from structural identity. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, Halifax, NS, Canada, 13–17 August 2017, pp. 385–394. New York: ACM.
[52] Bianconi G and Barabási A-L. Competition and multiscaling in evolving networks. *Europhys Lett* 2001; 54: 436.
[53] Klionsky DJ, Cregg JM, Dunn WA et al. A unified nomenclature for yeast autophagy-related genes. *Dev Cell* 2003; 5: 539–545.
[54] Van der Maaten LJP and Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008; 9: 2579–2605.
[55] Shibata N, Kajikawa Y, Takeda Y et al. Comparative study on methods of detecting research fronts using different types of citation. *J Assoc Inf Sci Tech* 2009; 60: 571–580.
[56] Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J Assoc Inf Sci Tech* 1973; 24: 265–269.
[57] Chen C. Predictive effects of structural variation on citation counts. *J Assoc Inf Sci Tech* 2012; 63: 431–449.
[58] Boyack KW and Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? *J Assoc Inf Sci Tech* 2010; 61: 2389–2404.
[59] Akaike H. A new look at the statistical model identification. *IEEE T Automat Contr* 1974; 19: 716–723.
[60] Chakraborty S, Tomsett R, Raghavendra R et al. Interpretability of deep learning models: a survey of results. In: *Proceedings of the 2017 IEEE SmartWorld, ubiquitous intelligence and computing, advanced and trusted computed, scalable computing and communications, cloud and big data computing, Internet of people and Smart city innovation*, San Francisco, CA, 4–8 August 2017.
[61] Upham SP, Rosenkopf L and Ungar LH. Positioning knowledge: schools of thought and new knowledge creation. *Scientometrics* 2010; 83: 555–581.
[62] Kuhn TS. The structure of scientific revolutions. *Am J Phys* 1970; 2: 14–16.
[63] Burt RS. Structural holes: the social structure of competition. Cambridge, MA: Harvard University Press, 1995.
[64] Levine B and Yuan J. Autophagy in cell death: an innocent convict? *J Clin Invest* 2005; 115: 2679–2688.
[65] Lum JJ, Deberardinis RJ and Thompson CB. Autophagy in metazoans: cell survival in the land of plenty. *Nat Rev Mol Cell Biol* 2005; 6: 439–448.