Journal of Informetrics 18 (2024) 101507

Contents lists available at ScienceDirect

Journal of Informetrics

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

Research Paper

A multiple k-means cluster ensemble framework for clustering citation trajectories

Joyita Chakraborty [a], Dinesh K. Pradhan [b,*], Subrata Nandi [a]

[a] Department of CSE, National Institute of Technology, Durgapur, 713209, India
[b] Department of CSE/IT, Dr. B.C. Roy Engineering College, Durgapur, 713206, India

ARTICLE INFO

Keywords:
Clustering citation trajectories
Time-series clustering
Unsupervised machine learning
k-Means
Cluster ensemble

ABSTRACT

Citation maturity time varies for different articles. However, the impact of all articles is measured in a fixed window (2-5 years). Clustering their citation trajectories helps understand the knowledge diffusion process and reveals that not all articles gain immediate success after publication. Moreover, clustering trajectories is necessary for paper impact recommendation algorithms. It is a challenging problem because citation time series exhibit significant variability due to non-linear and non-stationary characteristics. Prior works propose a set of arbitrary thresholds and a fixed rule-based approach. All methods are primarily parameter-dependent. Consequently, it leads to inconsistencies while defining similar trajectories and ambiguities regarding their specific number. Most studies only capture extreme trajectories. Thus, a generalized clustering framework is required. This paper proposes a *feature-based multiple k-means cluster ensemble framework*. Multiple learners are trained for evaluating the credibility of class labels, unlike single clustering algorithms. 195,783 and 41,732 well-cited articles from the Microsoft Academic Graph data are considered for clustering short-term (10-year) and long-term (30-year) trajectories, respectively. It has linear run-time. Four distinct trajectories are obtained – *Early Rise-Rapid Decline (ER-RD)* (2.2%), *Early Rise-Slow Decline (ER-SD)* (45%), *Delayed Rise-Not yet Declined (DR-ND)* (53%), and *Delayed Rise-Slow Decline (DR-SD)* (0.8%). Individual trajectory differences for two different spans are studied. Most papers exhibit *ER-SD* and *DR-ND* patterns. The growth and decay times, cumulative citation distribution, and peak characteristics of individual trajectories' are re-defined empirically. A detailed comparative study reveals our proposed methodology can detect all distinct trajectory classes.

1. Introduction

A citation trajectory represents the time-series distribution of annual citations received by a paper Min et al. (2021). The other terms used are *'citation curve'*, *'citation pattern'*, *'citation histories'*, and *'citation time-series'*. Clustering them refers to grouping papers with similar shapes or identical patterns in their citation life cycle (Pradhan et al., 2019). Thus, a cluster represents a trajectory pattern.

* Corresponding author.
E-mail address: pdkrin@yahoo.co.in (D.K. Pradhan).

https://doi.org/10.1016/j.joi.2024.101507
Received 22 May 2023; Received in revised form 28 December 2023; Accepted 30 January 2024

Available online 13 February 2024
1751-1577/

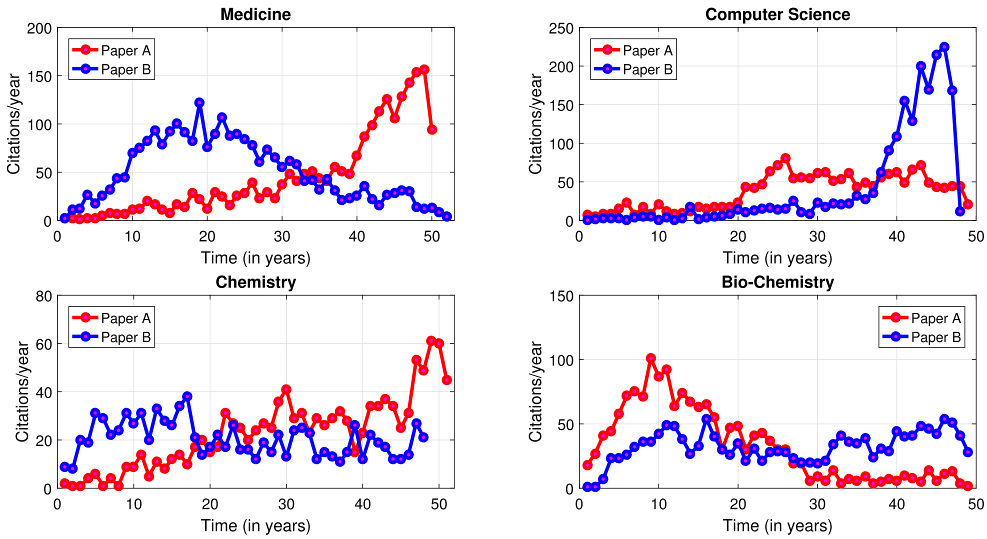


Fig. 1. Citation trajectories are plotted for two randomly chosen papers, each from four fields – Medicine, Computer Science, Chemistry, and Bio-Chemistry.

Different articles differ significantly based on their citation growth times. We present an example to illustrate this. The citation trajectories of two randomly chosen papers from four fields are plotted in Fig. 1. Both articles from each field were published around the same time and got equal citations. Medicine, Computer Science, Chemistry, and Biochemistry papers received 2593, 1959, 1166, and 1607 citations in around 50 years, respectively. Comparing early citations of two articles in the medicine field, we find that paper B is more likely to achieve a higher impact in the future than paper A. However, paper A suddenly jumps to receive greater than 50 citations annually after 40 years of publication. Although both papers receive the exact final and average citations, there is considerable variability in their trajectories.

There are diverse applications of studying citation trajectories. It is a fundamental source of time and topic-correlated information in scholarly networks. Thus, it can capture the knowledge diffusion process and track the evolution of new emerging fields (Chi et al., 2022). How some information gets early attention and fades out soon, how some information gets attention and remains relevant indefinitely, how some information goes unnoticed and gets delayed attention (Min et al., 2021), etc. Besides, popular research evaluation metrics (Hirsch, 2005, Chakraborty et al., 2021) use the final or average citations received by an article in a specific window. However, it cannot truly capture the time-varying changes in a publication's impact. Clustering helps empirically understand the entire citation lifecycle of different trajectory patterns. It can evaluate cross-disciplines, where some fields may require more time to grow than others (Clermont et al., 2021). Besides, the specific growth and decay times can be used as an input to paper recommendation algorithms for predicting breakthrough discoveries early (Bai et al., 2019). The same algorithms can retract articles based on the estimated decay times of individual clusters.

There are inherent challenges of clustering citation trajectory data. Citation trajectories exhibit non-linear, non-stationary, and long-ranged correlations (Pradhan et al., 2019, Golosovsky & Solomon, 2017). Recently, Zamani et al. (2021) mathematically prove that their variance varies $\propto t^{2H}$ where H $\neq 1/2$ after analyzing trajectories of more than 300,000 articles. Baumgartner and Leydesdorff (2014) reports that a fifth-order polynomial fits citation trajectories analyzed for over 16 years. Besides, these data sets are flooded by mediocre papers, self-cited papers, citation cartels, and lengthy reference lists Chakraborty et al. (2022), Chakraborty and Pradhan (2022), Pradhan et al. (2020). This adds to the variance. Besides, in general, time-series data are usually high dimensional and noisy. Thus, clustering them is computationally expensive. Moreover, it is not easy to interpret cluster results from raw trajectories.

Despite the above challenges, one of the first classical works on clustering trajectories was done by Aversa (1985) in 1985. She clustered the raw citation trajectories of 400 papers, each cited for nine years. She proposed an unsupervised k-means algorithm with k = 2. After this, a major volume of literature has first determined the possible types of citation trajectory patterns, that is, the number of clusters (k) based on intuition. Then, considering individual clusters, they defined various trajectory features and hard-code thresholds. The features primarily include the characteristics of growth and decay of citations over time. Then, they define the clusters. Recently, two works have mapped it again as an unsupervised clustering problem. Colavizza and Franceschet (2016) proposed a non-linear spectral clustering method with raw time series as input. Zhang et al. (2017) proposed a simple k-means algorithm with regression model coefficients as input. Thus, researchers, based on their understanding, have fixed the k, feature set, thresholds of different features, and other model parameters. Consequently, different methods lead to different groups of trajectory patterns being identified.

The primary motivation draws from the variance across studies in the value of k. As a result, identical clusters comprising articles with similar temporal citation behavior are studied as different clusters. Further, the choice of trajectory features and their thresholds for defining such clusters varies across studies. It leads to ambiguous cluster definitions. Moreover, with rapid changes in trajectory patterns, the thresholds chosen need to be quickly changed over time. Thus, paper recommendation algorithms cannot make use of such thresholds. Due to the above pitfalls, existing methods can only cluster a portion of all articles in a data set. For instance, Chakraborty et al. (2015) could not define trajectory patterns for 45% articles.

Our main objective is to propose a robust clustering framework that can capture all distinct *generalized trajectory patterns*. Thus, it should reveal the ideal value of k algorithmically based on data. For this, the challenge of evaluating the credibility of cluster labels in unsupervised learning algorithms needs to be addressed. Further, we aim to empirically verify the accuracy of thresholds chosen to define a particular cluster based on its growth and decay characteristics. We further aim to analyze whether there are any redundant interpretations in identified clusters.

The present study attempts to address multiple gaps and, in doing so, makes important contributions. We investigate the same problem following a more generic approach. *First,* we map the problem as an *unsupervised clustering problem*. The study proposes a *feature-based multiple k-means cluster ensemble (MKMCE)* framework. The cluster ensemble technique can algorithmically find the ideal value of k based on data by iteratively checking the credibility of cluster labels. Thus, there is no prior need to fix the typical number of clusters. Motivated by prior studies, we propose nine comprehensive features that can broadly capture time, citation, and peak count-related characteristics of a trajectory. Out of them, previous studies mostly consider a combination of either two or three features. *Second,* once the clusters are obtained, an accurate feature analysis guides us to the choice of thresholds. Thus, there is no prior need to manually define thresholds. We also draw a threshold comparison with similar clusters identified in prior literature. The precise thresholds are needed to calculate article relevance scores in paper recommendation algorithms. *Third,* we analyze redundant interpretations of similar trajectory patterns identified as different clusters and named differently in the literature. *Fourth,* we analyze varying lengths of trajectories, as some patterns can only be captured when analyzed for a longer window. We consider both short-term (10-year) and long-term (30-year) trajectories.

The paper is organized as follows. In section 2, we elaborately discuss related literature. In section 3, we define the feature set, the clustering methodology, experimental settings, and the MAG data set in the brief. Section 4 contains the main clustering results, cluster characteristics, and comparative study for validation. Section 5 concludes the research and discusses limitations and future implications.

2. Literature survey

Two separate lines of work investigate the problem of clustering citation trajectories. The first group of works proposes methods to detect a single cluster and examines *specialized trajectories*. Only a handful exhibit them. Sleeping beauties are one such example. The second group proposes methods to detect multiple clusters and analyzes *generalized trajectories*. They cluster the citation trajectories of millions of articles in a data set. Besides, the single clusters are sub-groups of the generalized cluster set. In this paper, we primarily focus on identifying generalized patterns.

First, we briefly summarize single cluster studies identifying specialized trajectories. Garfield (1989) first identified a specialized trajectory exhibiting the *Delayed Recognition (DR)* phenomenon. Van Raan (2004) in 2004, first coined the term '*Sleeping Beauty (SB)*' for the *DR* phenomenon. SBs are articles that do not receive citations for an extended period after publication, followed by a sudden spike in popularity. In the past decade, many of the works Li et al. (2014), Li and Shi (2016), Ke et al. (2015), He et al. (2018), van Raan (2021), Yang et al. (2022) had only studied SBs analyzing them from multiple dimensions. He et al. (2018) sub-categorized SBs into *single-peak SBs*, *second-act SBs*, and *second-act non-SBs*. Contrary to it, there exists another specialized trajectory described as – *Flashes-in-the-Pan (FP)* (Van Dalen & Henkens, 2005), *shooting stars* (Mingers, 2007), or *breakthrough or discovery papers* (Wang et al., 2023, Wei et al., 2023, Xu et al., 2022). Such articles receive high citations immediately after publication but without lasting impact. Moreover, Li and Ye (2012) identified a sub-group of SB's – *All-Element-Sleeping-Beauties (ASB)*. They observe that ASBs first appear as FP and then follow the trajectory of SB. Unlike standard SBs, ASBs receive high annual citations initially until spindles occur, and they make them fall into deep sleep. Finally, the prince re-awakens them, and they regain popularity long after publication.

Next, we present diverse literature on multiple cluster studies identifying generalized trajectories. Please refer to Table 1 for a summarized survey. Based on the methods proposed in prior works, we can broadly categorize them into – *unsupervised clustering-based* and *threshold-based* methods.

1. Unsupervised clustering-based methods: The problem of grouping articles with similar citation trajectories is mapped as an unsupervised clustering problem. It requires the number of clusters (k) to be specified. Aversa (1985) conducted the first classical study in 1985. A raw time series of 400 highly-cited papers was considered with nine years of citation histories. She proposed a simple k-means algorithm with k = 2. Two clusters were identified – *Early Rise-Rapid Decline (ER-RD)* and *Delayed Rise-Slow Decline (DR-SD)*. After this, none of the works solved it as an unsupervised clustering problem except for two recent studies.

Colavizza and Franceschet (2016), in 2016, input raw time series and proposed a spectral clustering method. It is a non-linear clustering algorithm that recognizes clusters with any shape. However, it is computationally expensive due to pairwise similarity matrix calculations. Three clusters were identified – *Sprinters*, *Middle-of-the-Roads*, and *Marathoners*. In 2017, Zhang et al. (2017) proposed a Poisson regression model to fit trajectories and then used coefficients as input to a simple k-means algorithm. Four clusters were identified – *Normal Low (NL), Normal High (NH), Delayed Documents (DD),* and *Evergreens (EG)*. As raw time series or model coefficients are considered input, the trajectory patterns based on their growth and decay cannot be precisely defined.

2. Threshold-based methods: A major volume of literature has first intuitively determined all possible citation trajectory patterns (number of clusters (k)). Then, based on the proposed patterns, they hard-code thresholds on trajectory features. After this, the features define the proposed trajectory patterns. The feature set primarily includes – the properties defining the rise and fall of a trajectory. We can further sub-categorize them into – *intuitively selected based on the understanding of individual researchers* and *thresholds selected by analytical models*.

Table 1
Summary of existing literature on clustering citation trajectories.

Citation trajectory clustering method		Previous literature	Number of identified clusters (k)	Cluster names	Features/method specifications	Remarks
[1] Threshold-based	Intuitively-selected thresholds	Aksnes (2003)	3	ER-RD, MR-SD, and DR-ND	% of citations received in 3 years and after 7-12 years	Number of clusters intuitively defined and features manually chosen
		Redner (2004)	3	DP, HP, and SB	Final citation count and the ratio of mean citation age to publication age	
		Lange (2005)	2	Hits and Missed Signals	Number of citations and their time of occurrence	
		Costas et al. (2010)	3	FP, ND, and DD	Time taken by an article to receive 50% or half of its final citation count	
		Li (2014)	3	FP, ASB, and DR	Number of citations received in the sleeping period and awakening period	
		Chakraborty et al. (2015)	6	PeakInit, MonInc, PeakMult, MonDec, PeakLate, and Others	Number of peaks and their time of occurrence in a trajectory	
		Bornmann et al. (2018)	2	HP and DR	Field and time normalized impact of citations calculated as peak and the location of peaks in the early and later half of the trajectory	
		Ye and Bornmann (2018)	2	SG and SB	Time taken by an article to receive 50% of its total citations, number of citations in that period, and citation angle	
	Analytically-selected by models	Baumgartner and Leydesdorff (2014)	2	TKC and SKC	Group-Based Trajectory Modeling (GBTM)	Number of clusters and features defined by GBTM model
		Min et al. (2018), Bjork et al. (2014)	4	(small p, small q), (large p, large q), (small p, large q), and (large p, small q)	BASS model (innovation (p) and imitation (q) co-efficients)	Number of clusters and features defined by Bass model
[2] Unsupervised **clustering-based**		Aversa (1985)	2	ER-RD and DR-SD	Raw time-series to k-means clustering algorithm	Number of clusters pre-defined. The interpretation of obtained clusters is unclear as raw time-series and model co-efficients are analyzed after clustering
		Colavizza and Franceschet (2016)	3	Sprinters, Middle-of-the-Roads, and Marathoners	Raw time-series to non-linear spectral clustering algorithm	
		Zhang et al. (2017)	4	NL, NH, DD, and EG	Co-efficients of Poisson regression model to k-means clustering algorithm	

Intuitively selected threshold-based studies include the following. Aksnes (2003) clustered 297 articles cited over 16 years. For each article, he calculated citations received in 3 and 7-12 years of time windows. The rise of a trajectory was categorized into *Early*, *Medium*, and *Delayed rise*, if a publication got > 30%, between 15%-30%, and < 15% of its final citations in the first three years, respectively. Further, the decline was categorized into *Rapid*, *Slow*, and *No Decline* if a publication received < 30%, between 30%-50%, and > 50% of its final citations in the later period. Three clusters were identified – *ER-RD*, *MR-SD*, and *DR-ND*. Redner (2004) identified three clusters by arbitrarily defining thresholds – *Sleeping Beauties (SB), Discovery Papers (DP),* and *Hot Papers (HP)*. SBs received > 250 citations with the ratio of the mean citation age to publication age (r) > 0.7. DP's received > 500 citations with r < 0.4. HP's received > 350 citations with r > 2/3. Lange (2005) considered the number of citations and their time of occurrence in psychology journal articles. They divided them into two clusters – *Hits* and *Missed Signals*.

Costas et al. (2010) initially divided the trajectory into two halves based on the time taken to attain 50% of its total citations (Y50). Next, they determine the peak location by comparing the time of receiving 25% and 75% of its total citations with Y50. They considered articles cited over 29 years. They identified three clusters – '*Flashes-in-the-Pan (FP)*,' '*Normal Documents (ND)*,' and '*Delayed Documents (DD)*. Chakraborty et al. (2015) considered the number of peaks (n_{CP}) and their location (t_{CP}) for defining thresholds. They defined six clusters – *PeakInit* (n_{CP} in t_{CP} <= 5 years followed by an exponential decline), *MonInc* (monotonic increase in n_{CP} till 20 years after publication), *PeakMult* (multiple n_{CP}), *MonDec* (n_{CP} in the first year after publication followed by a monotonic decrease in citations), *PeakLate* (few initial citations and single n_{CP} in t_{CP} > 5 years but not in the last year), and *Others* (undefined trajectory). Recently, Gou et al. (2022) defined a *literature revival* phenomenon. They are similar to *PeakMult* group of papers receiving multiple citation peaks even after decay. In the revival phase, they chose the threshold as the number of citations received by a paper should be greater than 20% of the peak of its annual citations.

Bornmann et al. (2018) measured field and time normalized citation impact scores and identified two clusters – *Hot Papers (HP)* and *Delayed Recognition (DR)*. The thresholds were defined based on peaks in the early or later half period, similar to Costas et al. (2010). Ye and Bornmann (2018) categorized into two clusters – *Smart Girls (SG)* and *Sleeping Beauties (SB)*. They used beauty co-efficient (Ke et al., 2015) and proposed the concept of citation angles. The citation angle for *SGs* was > 60° and < 30° for *SBs* as compared to the zero citation line.

Studies that select thresholds based on analytical model-based approaches include the following. Few studies Baumgartner and Leydesdorff (2014), Comins and Leydesdorff (2017) grouped them into two clusters-*transient* and *sticky* knowledge claim. Bjork et al. (2014) and Min et al. (2018) used the BASS model from management studies. Min et al. (2018) considered two parameters – *innovation (p)* and *imitation (q)* coefficients and defined four clusters – papers with *(small p, small q)*, *(large p, small q)*, *(small p, large q)*, and *(large p, large q)* values.

The major drawbacks of previous studies are — (1) the distinct number of clusters (k) or possible types of trajectory patterns varies across studies. This is because the k-value is pre-defined and mostly hard-coded based on the intuition of a researcher (refer to Table 1 for variance in k-value). As a result of which, redundant clusters are identified across studies. For an example, clusters identified as – early rise-rapid decline (Aversa, 1985, Aksnes, 2003), flashes-in-the-pan (Van Dalen & Henkens, 2005, Costas et al., 2010), sprinters (Colavizza & Franceschet, 2016), transient-knowledge-claim (Baumgartner & Leydesdorff, 2014), MonDec (Chakraborty et al., 2015), hot papers (Bornmann et al., 2018), and smart girls (Ye & Bornmann, 2018) represent similar trajectories. However, they are studied as different clusters. Further, sparsely populated clusters may be sub-clusters to a broader cluster set. Thus, different identification methods capture different groups of trajectories. Consequently, there are ambiguities regarding the exact number of distinct trajectories. (2) Besides, the choice of trajectory features and their thresholds for defining such trajectory patterns is also questionable (refer to features/method specifications in Table 1). It leads to inconsistencies in their definitions of citation behavior. Also, the trajectory patterns rapidly change over time due to the exponential addition of articles. Consequently, the value of hard-coded thresholds may change over time. (3) As most existing methods have parameter dependence, they have scalability issues and are unsuitable for large-scale datasets. Consequently, a major proportion of articles remains unclustered.

The main objective of our work lies in answering the following fundamental questions. What should be the best value of k or typical trajectory patterns? Can it be solely decided algorithmically depending on data distributions? How can we deal with the high dimensionality issues in clustering raw time series data? Can we verify whether the thresholds defining a particular trajectory pattern in prior studies are accurate? Do they need to be re-defined? Are there any redundant interpretations of clusters? Does one of them belong to the broader cluster of the other?

3. Methodology

This section initially presents the trajectory features selected for clustering. We also introduce and motivate the choice of the *Multiple K-Means Cluster Ensemble algorithm (MKMCE)* for clustering. Further, we present our experimental setup and briefly describe the Microsoft Academic Graph (MAG) data set.

3.1. Feature selection

This paper derives the trajectory features from a paper's raw time series. The feature-based approach removes the necessity of manually defining thresholds or parameters for different trajectory patterns. Besides, it helps to represent the same information in a lower-dimension space and reduces memory requirements (Baghizadeh et al., 2020). It speeds up the clustering process as the distance calculation in clustering algorithms using raw time series can be computationally expensive (Blázquez-García et al., 2021). Moreover, some distance measures are sensitive to noise. Thus, clustering using raw data may group time series similar in noises

Table 2
The generic feature set for evaluating trajectories.

Summary	No.	Features	Notation	Description
Time-related features	F1	Initial time	T_i	It is time a paper takes to attain the geometric mean of its final citations.
	F2	Growth time	T_g	It is the difference between the time a paper takes to receive its highest annual citation (maximum) and geometric mean.
	F3	Decay time	T_d	It is the difference between the time when a paper is last cited (publication age) and the time a paper takes to receive its highest annual citation.
Citation-related features	F4	Citation gain in T_i	C_{T_i}	It is the ratio of cumulative citations received by a paper in time T_i to its final citations. $C_{T_i}(c) = \frac{\sum_{t=0}^{T_i} c_t}{c}$
	F5	Citation gain in T_g	C_{T_g}	It is the ratio of cumulative citations received by a paper in time T_g to its final citations. $C_{T_g}(c) = \frac{\sum_{t=0}^{T_g} c_t}{c}$
	F6	Citation gain in T_d	C_{T_d}	It is the ratio of cumulative citations received by a paper in time T_d to its final citations. $C_{T_d}(c) = \frac{\sum_{t=0}^{T_d} c_t}{c}$
Peak count-related features	F7	Number of peaks of high-intensity	n_{I^h}	The total number of citation values in the time series greater than or equal to $\mu_i + 3\sigma_i$ is calculated separately for two periods, T_g and T_d.
	F8	Number of peaks of medium-intensity	n_{I^m}	The total number of citation values in the time series greater than or equal to $\mu_i + 2\sigma_i$ is calculated separately for two periods, T_g and T_d.
	F9	Number of peaks of low-intensity	n_{I^l}	The total number of citation values in the time series greater than or equal to $\mu_i + \sigma_i$ is calculated separately for two periods, T_g and T_d.

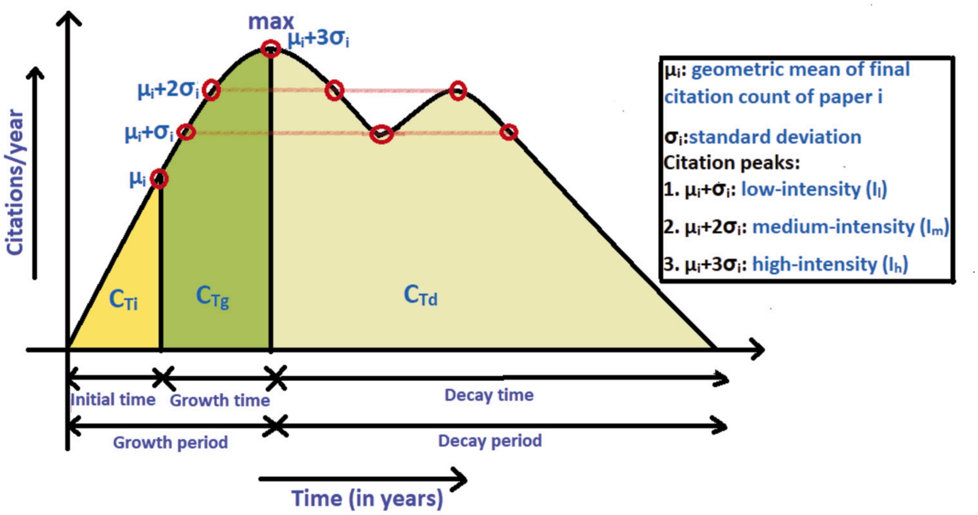


Fig. 2. A hypothetical citation trajectory of paper 'i' is plotted to understand the derived feature set quickly. Here, μ_i is the geometric mean of final citations, and σ_i is the standard deviation considering the entire time series. The location of citation peaks is shown.

than identical in their characteristic shape (Baghizadeh et al., 2020). Thus, it helps to input an even-length vector. Table 2 represents a complete description of features and their notation.

Fig. 2 represents a hypothetical trajectory. Broadly, we divide it for any given paper into two phases – *growth* and *decay* period. One of the pioneer studies modeling citation distribution is the WSB model proposed by Wang et al. (2013). They highlighted the importance of two features – *impact time* and *immediacy time* for studying a trajectory. Here, impact time is the time to achieve the geometric mean of the final citation count, and immediacy time is when a paper receives its maximum or highest annual citation. We use these two variables – impact and immediacy time to measure early growth or initial time and growth time, respectively. Both of them refer to the growth period.

Summarizing, we empirically define three measurements of time (time-related features F1, F2 and F3) – *initial time* (T_i), *growth time* (T_g), and *decay time* (T_d) (see full description in Table 2). Further, we use statistical methods to measure a given paper's *citation* *growth or decay* (citation-related features F4, F5, and F6) separately in each of the above three times. We calculate the gain variable to measure citation growth or decline.

Finally, we separately measure *the number of peaks of three different intensities* (peak count-related features F7, F8, and F9) in growth and decay times. We use an empirical outlier detection rule (Laptev et al., 2015, Blázquez-García et al., 2021) widely used in time-series analysis. First, we calculate i^{th} paper's mean citation (μ_i) and standard deviation (σ_i) considering its entire time series. Next, we calculate citation values of magnitude greater than $\mu_i + \sigma_i$ (*low-intensity*), $\mu_i + 2\sigma_i$ (*medium-intensity*), and $\mu_i + 3\sigma_i$ (*high-intensity*). The resultant number of peaks of three different intensities separately occurring in growth and decay times are considered as input features (refer to Fig. 2).

It is essential to standardize features before clustering, as the distance between different data points is measured. Normalization refers to the probability distribution of features. We scale each feature (f_i) using a z-score value given as, $z_{f_i} = \frac{x_n - \mu_{f_i}}{\sigma_{f_i}}$.

3.2. Clustering framework

This section motivates the choice of the *MKMCE algorithm* (Bai et al., 2020). We map the problem of clustering citation trajectories into an *unsupervised clustering* problem. Thus, the accuracy of labels has no clear meaning. Different algorithms or the same algorithm with different input parameters often produce different groups of clusters on the same data set. Single clustering algorithms could not identify non-linearly separable clusters. Moreover, nonlinear clustering algorithms such as spectral clustering (Colavizza & Franceschet, 2016), and density-based spatial clustering of applications with noise (DBSCAN) (Costas et al., 2010) have expensive time costs. Their pair-wise distance calculations between objects are not suitable for large data sets. Compared to them, Bai et al. (2020) proved that the MKMCE algorithm is faster and can rapidly discover non-linearly separable clusters. It also worked well with large data sets such as KDD99, which had 1,048,576 entities. Thus, we proposed to use this technique.

3.2.1. Problem definition

Let X be an original data set consisting of N objects given by $\{x_i\}_{i=1}^{N}$. F is a set of M features. Besides, X(F) is the representation of X on F. It is a matrix of dimension N x M. Further, $X(f_j)$ represents j^{th} column, $x_i(F)$ represents i^{th} row, and $x_i(f_j)$ is the value of object x_i in feature f_j.

1. *Multiple k-means clustering problem to generate a base cluster set:* The k-means algorithm is used as a base clusterer and is run multiple times to generate a base cluster set. Let, $Z = \{\zeta_h\}_{h=1}^{T}$ represent a set of T base clustering of X(F) where, $\zeta_h = \{C_{hl}\}_{l=1}^{k_h}$ is the h^{th} base clustering. $K = \{k_h\}_{h=1}^{T}$ represents the entire set of the number of clusters in each base clustering ζ_h where k_h is the number of clusters in h^{th} base clustering. Further, $\zeta_h(x_i(F))$ is the cluster label of object $x_i(F)$ in clustering ζ_h. $\zeta_h(x_i(F)) = l$ denotes that object $x_i(F)$ belongs to cluster C_{hl}. The objective function (A) of k-means can be given as,

$$A(\zeta_h, y_h) = \sum_{l=1}^{k_h} \sum_{\zeta_h(x_i(F))=l, x_i(F)\in X(F)} d(x_i(F), y_{hl})^2 \tag{1}$$

   Here, y_{hl} is the l^{th} cluster center and $y_h = \{y_{hl}\}_{l=1}^{k_h}$. Further, $d(x_i(F), y_{hl}) = \sqrt{|(x_i(F), y_{hl})|^2}$ is the Euclidean distance between object $x_i(F)$ and the center y_{hl} of the l^{th} cluster. The algorithm minimizes A by constantly updating ζ_h and y_h. The clustering results are different each time due to different initial cluster centers. The next part resolves this issue.
2. *Cluster ensemble problem to generate a final cluster set:* The cluster ensemble problem integrates the base clusters rapidly to generate a final clustering Z^* of data set X(F) based on the clustering set Z. The final cluster set can be given as C_*. It can also be represented as $C_* = X(\zeta_*)$ where ζ_* is the final clustering feature. If $x_i(F) \in C_{*l}$ then, $x_i(F)(\zeta_*) = \zeta_{*l}$. Here, ζ_{*l} is the cluster label of C_{*l} for $1 \leq i \leq N$.

3.2.2. MKMCE algorithm

It is an unsupervised ensemble learning method where several base clusters integrate to form final clusters with improved stability and accuracy. A cluster ensemble method is more efficient than single clustering algorithms as it assesses the credibility of class labels. For instance, simple k-means is known for its low computational cost. However, it is susceptible to data distributions (Xiong et al., 2006). Roughly, the proposed algorithm includes four main steps: initially generate multiple k-means base clusters, evaluate the local credibility of each label, build the relationship between clusters, and generate the final clustering set. It improves the robustness and quality of k-means. It can rapidly recognize non-linearly separable clusters.

Broadly, the MKMCE algorithm can be divided into two parts – (1) initially, it produces a base cluster set using k-means for multiple clusterings, (2) finally, integrate them into a final cluster set using cluster ensemble algorithm. The key steps are as follows:

1. *Local credibility function:* Unlike supervised learning algorithms, we do not know the exact cluster label of data points in unsupervised learning methods. Besides, a cluster center is used to represent a cluster in k-means. However, the cluster center is not convenient for representing a non-linear cluster. Therefore, a local credibility function is introduced to check whether the cluster label of objects falls into the local ψ neighborhood of its cluster center. It is defined as,

$$\psi_h(x_i(F)) = \begin{cases} 1 & \text{if } x_i(F) \in \mathbb{N}(y_{hl}), \\ 0 & otherwise, \end{cases} \tag{2}$$

Here, $l = \zeta_h(x_i(F))$ and $\mathbb{N}(y_{hl})$ is the ψ neighborhood of the cluster center y_{hl}. It is also called as the local credible space of cluster C_{hl}, for $1 \leq i \leq N$ and $1 \leq h \leq T$.

2. *Multiple k-means clustering algorithm:* The objective function (U) to produce multiple base clusters using k-means can be given as,

$$\min_{Z} \left[U(Z) = \sum_{h=1}^{T} \sum_{i=1}^{N} \theta_h(x_i(F))\ \psi_h(x_i(F))\ d(x_i(F), v_{h\zeta_h(x_i(F))}) \right] \tag{3}$$

Here, $\theta_h(x_i(F))$ is a boolean variable whose value is 1, if $x_i(F)$ takes part in h^{th} base clustering and 0 otherwise. Also, once an object forms a base cluster, it does not participate in further k-means clustering. Thus, the aim is to minimize the value of objective function (U) with the constraint

$$\sum_{h=1}^{T} \theta_h(x_i(F))\ \psi_h(x_i(F)) = 1, 1 \leq i \leq N \tag{4}$$

Initially, we set h = 1, S = X(F), and $\theta_h(x_i(F)) = 1$ for $1 \leq i \leq N$. Next, we randomly select k_h points as initial cluster centers from S. We apply the k-means algorithm multiple times with constraint equation (4). Only those objects are included in the cluster in the ϵ neighborhood of randomly initialized cluster centers. Finally, we assign each object in X(F)-S to the cluster with the nearest cluster centers. Next, we update $S = S - S'$ where S' contains objects already clustered in h^{th} base clustering. Also, update h = h+1 and $\theta_h(x_i(F)) = 1$, if $x_i(F) \in S$ else 0. This incremental clustering method runs until the desired number of base clusterings (T_{max}) is obtained, or the number of objects in S' ($|S'|$) is lesser than k_h^2. T_{max} can be set depending on user requirements. We initially set $T_{max} = 100$, and based upon running it for a few iterations, we find $T_{max} = 10$ as the ideal choice. Also, considering all prior works Chakraborty et al. (2015), we observed that up to a maximum of six clusters had been reported. The output of this part of the algorithm are base cluster set $Z = z_h, 1 \leq h \leq T$ and a cluster center set $P = p_h, 1 \leq h \leq T$.

3. *Cluster ensemble algorithm:* The final clusters should have high concurrence with the features of the base and original clusters. The overlap of credible local space between any two base clusters is used to measure their similarity. However, the credible local space between two base clusters is naturally small. It is due to the generation mechanism of base clusters using multiple k-means clustering. Thus, a latent cluster is introduced to measure the indirect overlap between base clusters.
Let, C_{hl} and C_{gj} be any two base clusters and C_q be a latent cluster as per assumption. Let y_{hl} and y_{gj} represent the cluster centers of two base clusters. The center of the latent cluster can measure the similarity between any two base clusters. The mid-point of two cluster centers of base clusters is used to calculate the center of the latent cluster. It determines whether there is an indirect overlap between two base clusters. Further, it is calculated as, $d(y_{hl}, y_{gj}) = \frac{y_{hl}+y_{gj}}{2}$. If $d(y_{hl}, y_{gj}) \leq 4\epsilon$, the clusters are indirectly overlapped. Thus, the similarity measure is inversely proportional to $d(y_{hl}, y_{gj})$. Mathematically, it can be represented as,

$$\lambda(C_{hl}, C_{gj}) = \begin{cases} \frac{|P(\frac{y_{hl}, y_{gj}}{2})|}{d(y_{hl}, y_{gj})} & \text{if } d(y_{hl}, y_{gj}) \leq 4\epsilon, \\ 0 & otherwise, \end{cases} \tag{5}$$

Finally, based on similarity measures, an undirected weighted graph is constructed G = (B, W). Here, B is a set of vertices representing a cluster label from Z, and W is the weight set. The similarity between any two clusters say, C_x and C_y, is the weight of the edges between them. After constructing a weighted graph, a normalized graph cuts problem is proposed to derive cluster relations (Shi & Malik, 2000). The larger similarity value between vertices represents that they belong to the same cluster with a higher probability. Also, they are dissimilar from the vertices of other clusters. Finally, the cluster labels are re-labeled to form the final clusters. Let $L(C_x)$ be the label of the subset which C_x belongs to,

$$L(C_x) = l,$$

if $C_x \in A_l$, for $1 \leq l \leq k$ and $x \in A$. After re-labeling the base cluster set, Z can be transformed into a re-labeled set V as follows:

$$V_{x_i} = L(C_{hZ_h(x_i)})$$

for $1 \leq i \leq N$ and $1 \leq h \leq T$.

3.3. Experimental setup

Let 'x' be a publication, and T = (1, 2, 3, ..., n) represent a series of consecutive years when 'x' is cited where $n \geq 1$ is an integer. For every $y \in T$, let $k_{x,y}$ represent the number of citations received by a paper x during a period y. A vector can define the citation trajectory of a publication x over T as:

$$k_{x,T} = (k_{x,1}, k_{x,2}, k_{x,3}, ...k_{x,n}) \tag{6}$$

The first fundamental question is, how many minimum citations should a paper receive to produce meaningful patterns in its trajectory? We use the metric *relative success ratio (s_x) of a paper x* as defined by Radicchi and Castellano (2011). It is given as,

$$s_x = \frac{c_x}{max(\mu_x, 5)} \tag{7}$$

Here, c_x and μ_x are the total number of citations and mean citation rate received by a publication x over a period T. They can be represented as,

$$c_x = \sum_{T=1}^{n} k_{x,T} \tag{8}$$

$$\mu_x = \frac{c_x}{T} \tag{9}$$

Colavizza and Franceschet (2016) use $max(\mu_x, 5)$ parameter in the denominator to account for years that receive a low mean citation rate. We only consider papers in our study with a value of $s_x \geq 1$. This method is used in the beginning to filter out well-cited articles with above-average citation impact so that its time-series trajectory can help us derive meaningful trends. It helps to address the problem of citation variance that occurs with various factors added due to the time of publication.

The following question is, what should be the exact length of a citation trajectory to be considered for clustering? Citation histories can only be compared if their lengths are equal (Colavizza & Franceschet, 2016). Consequently, trajectories of three different time windows – 10 years, 20 years, and 30 years are considered separately. Above it, a minimum citation threshold, as discussed earlier, is set to filter out articles. The 10-year time window is chosen because impact factors (Garfield, 2006) usually consider 2 to 5 years as it is the average time for attaining citation growth in most disciplines. However, some fields, such as social sciences, take longer to peak. Besides, we examine another 5-year window for capturing the decay pattern. Next, 20 and 30 years are multiples of 10, and it should allow us to investigate the long-term behavior of citation histories. Besides, the internet era began around 1985, and citation-based metrics such as the h-index came into practical use in 2005. Consequently, papers published in 1985, 1995, and 2005 and cited till 2015 are considered for studying 10-year (Chakraborty et al., 2015, Colavizza & Franceschet, 2016), 20-year (Chakraborty et al., 2015), and 30-year trajectories (Zhang et al., 2017), respectively.

3.4. Data

The Microsoft Academic Graph (MAG) dataset is used for empirical study. The first version was published on 5^{th} June 2015.[1] However, we have downloaded the version created on 5^{th} February 2016 for this study. It was made accessible via API. It is published as a set of tab-separated files of a total size of 28 GB (compressed to a ZIP format). Sinha et al. (2015) have vividly described the entire dataset. Papers in MAG can be divided into 'primary documents' with complete metadata information (including authors, venue of publication, date of publication, references, and URL) and 'secondary documents', existing only as IDs. The latter type is removed from the retrieved set before analysis. For this study, we have specifically used the following files—"Papers" and "Paper References" to calculate temporal citation trajectories of individual papers.

We get 195,783 papers published in 2005, 56,380 papers published in 1995, and 41,732 papers published in 1985 after filtering using methods described in section 3.3. All sets of papers receive citations till 2015. The cumulative citation distribution of papers is right-skewed. For example, while studying the 30-year distribution of cumulative citation count, we find that only 30 papers receive citations greater than 10,000, 106 papers receive citations greater than 5000, 1810 papers receive citations greater than 500, and 73.6% receive citations fewer than 100. Similarly, while studying the 10-year distribution of cumulative citation count, we find that only 28 papers receive citations greater than 10,000, 154 papers receive citations greater than 5000, 5362 papers receive citations greater than 500, and approximately 75% receive citations fewer than 100.

4. Results

This section extracts features and applies the multiple k-means cluster ensemble algorithm (MKMCE). The flowchart in Fig. 3 shows the number of papers obtained in each base and final cluster set. We do not find any significantly different clusters while considering data of the first two windows – 10 years and 20 years. Thus, the resulting distinct clusters are discussed considering two lengths – short-term (10 years) and long-term (30 years).

Table 3 represents the final cluster set. Further, the characteristics of each cluster are discussed, that is, citation trajectories growth and decay cycles. Finally, a comparative study is performed to validate obtained clusters with identical trajectories detected in prior literature.

4.1. Clustering short-term trajectories

The length of a citation trajectory considered for this study is 10 years. We consider papers published in the year 2005 and cited till 2015. The final set of papers considered is 195,783. We obtain three distinct clusters – cluster S1, S2, and S3 (see Fig. 3). The

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

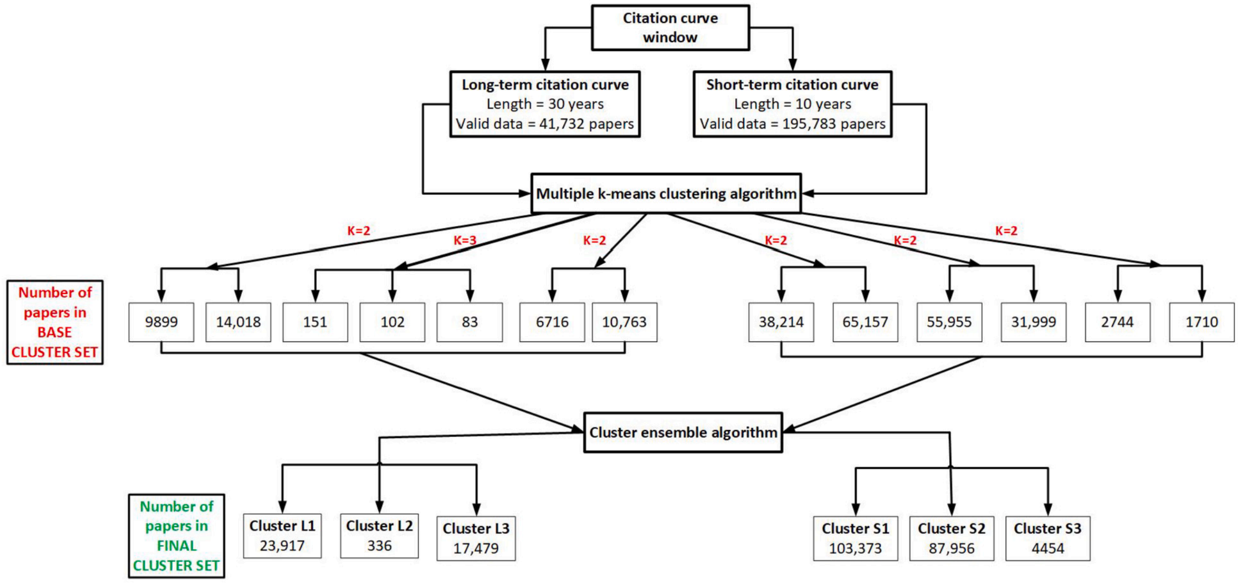


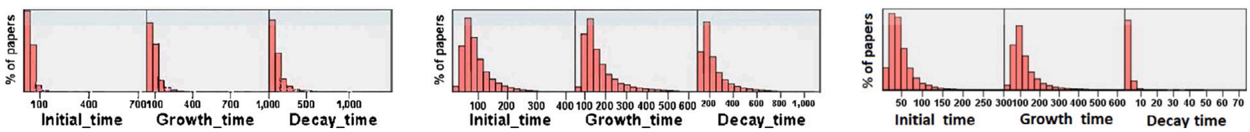
Fig. 3. The number of papers in each base and final cluster set after applying multiple k-means cluster ensemble algorithm.

Table 3
Universal final cluster set.

	Not yet Declined (ND)	Rapid Decline (RD)	Slow Decline (SD)
Early Rise (ER)	-	Cluster S3	Cluster L3, Cluster S2
Delayed Rise (DR)	Cluster L1, Cluster S1	-	Cluster L2

Table 4
(Short-term trajectory study) The descriptive statistics of initial time (T_i), growth time (T_g), and decay time (T_d) is shown for final clusters.

Short-term citation trajectory	Early Rise-Rapid Decline			Early Rise-Slow Decline			Delayed Rise-Not yet Declined		
	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)
Mean	1.51	2.16	2.11	2.31	3.15	3.88	3.05	5.72	0.5
Standard deviation	1	1	1	0.7	0.5	1	0.5	2	0.5
Quartile 1	0	1	1	1	2	2	2	3	0
Quartile 2	1	2	2	2	3	3	3	5	0
Quartile 3	2	3	3	2	3	4	3	6	0



(a) Early Rise-Rapid Decline (b) Early Rise-Slow Decline (c) Delayed Rise-Not yet Declined

Fig. 4. The histogram plot shows the cumulative citation distribution of each cluster for three consecutive times – initial time, growth time, and decay time. Figures (a), (b), and (c) represent ER-RD, ER-SD, and DR-ND clusters for short-term trajectories, respectively.

citation patterns for clusters S3, S2, and S1 are '*Early Rise-Rapid Decline (ER-RD)*', '*Early Rise-Slow Decline (ER-SD)*', and '*Delayed Rise-Not yet Declined (DR-ND)*', respectively (see Table 3). Moreover, the percentage of papers in *ER-RD*, *ER-SD*, and *DR-ND* clusters are 2.2%, 45%, and 53%, respectively.

4.1.1. Cluster analysis

Table 4 shows each cluster's descriptive statistics of initial time, growth time, and decay time. Besides, the histogram plot in Fig. 4 shows cluster-wise cumulative citation distribution separately for three consecutive times. Finally, in Fig. 5, the box plot shows the distribution of citation peaks of different intensities.

1. *ER-RD:* They have an average initial time of 1 to 1.5 years and a growth time of ∼ 2 years. Thus, the total growth period for papers in this cluster is 3 to 3.5 years, followed by a quick decay in 2 years (see Table 4). Cumulative citation distribution

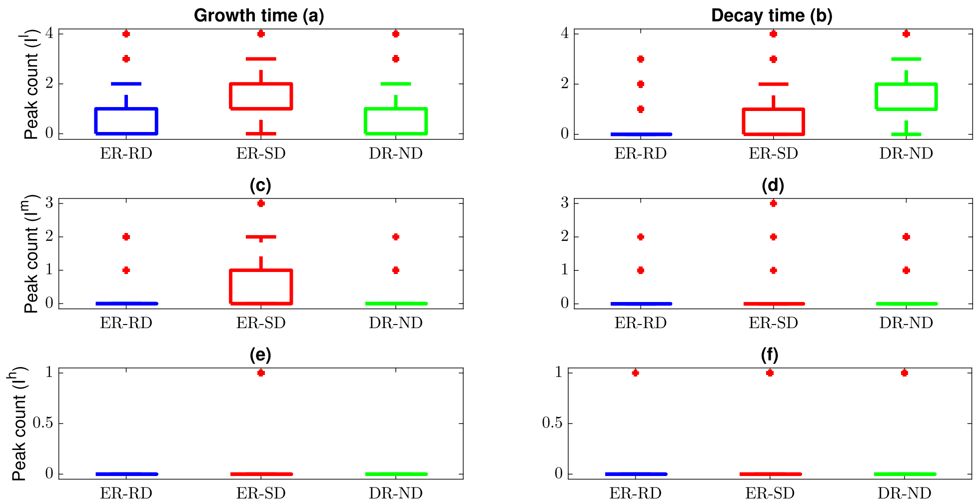


Fig. 5. The box plot represents the distribution of the number of citation peaks of three different intensities – n_{I^l} (figures (a), (b)), n_{I^m} (figures (c), (d)), and n_{I^h} (figures (e), (f)) separately for growth and decay times. Blue, red, and green indicate ER-RD, ER-SD, and DR-ND clusters.

in Fig. 4 (a) reveals that they receive the least citations compared to other clusters as they receive citations only for a short window. However, a handful of papers also receive as high as 1000 citations. Haghighat and Hayatdavoudi (2021) points out self-citation stacking as one of the reasons behind the citation patterns of such hot papers. The number of I^l intensity peaks are seen chiefly during their growth period (see blue colored box-plot in Fig. 5 (a)). Table 3 reveals it is only a characteristic of short-term citation trajectories.

2. *ER-SD:* They have an average initial time of 2 years and a growth time of 3 years. Consequently, the total growth time for papers in this cluster is 5 years, followed by a slow decay in the next 3 to 4 years (see Table 4). The histogram plot depicting cumulative citation distribution in Fig. 4 (b) reveals a right-skewed distribution where papers receive more citations as it shifts from initial to growth to decay time. They receive the highest number of peaks of I^l and I^m intensity mainly in the growth time (see red colored box plot in Fig. 5 (a), (c)). Also, considering three consecutive times, they receive the maximum number of peaks compared to the other two clusters.
3. *DR-ND:* They have an average initial time of 3 years and an average growth time of ∼ 6 years. We observed no citation decay for the period analyzed. Thus, they are defined only for the considered study period. The histogram plot depicting cumulative citation distribution in Fig. 4 (c) shows that a significant proportion of citations is received in the growth time. They attain the highest number of citation peaks of I^l intensity during their delayed growth time (see green colored box plot in Fig. 5 (b)).

4.2. Clustering long-term trajectories

The length of a citation trajectory considered for this study is 30 years. We consider papers published in the year 1985 and cited till 2015. The final set of papers considered is 41,732. We obtain three distinct clusters – cluster L1, L2, and L3 (see Fig. 3). The citation patterns for clusters L3, L1, and L2 are '*Early Rise-Slow Decline (ER-SD)*', '*Delayed Rise-Not yet Declined (DR-ND)*' and '*Delayed Rise-Slow Decline (DR-SD)*', respectively (see Table 3). The percentage of papers in *ER-SD*, *DR-ND*, and *DR-SD* clusters are 42%, 57%, and 0.8%, respectively.

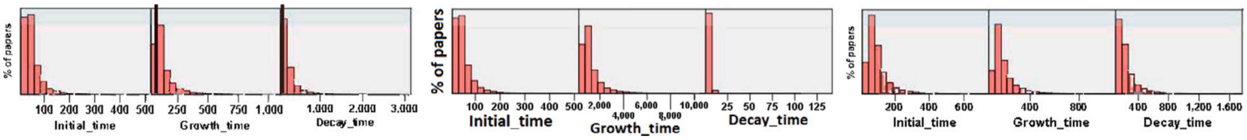
4.2.1. Cluster analysis

Table 5 shows each cluster's descriptive statistics of initial time, growth time, and decay time. Besides, the histogram bar plot in Fig. 6 shows cluster-wise cumulative citation distribution separately for three consecutive times. Finally, in Fig. 7, the box plot shows the distribution of citation peaks of different intensities.

1. *ER-SD:* They have an average initial time of 2 years and a growth time of 4 to 4.5 years. Consequently, the total growth time for papers in this cluster is 6 years, followed by an average decay time of 20 years (see Table 5). Cumulative citation distribution in Fig. 6 (a) reveals ER-SD papers receive a significant proportion of their final citations during growth time. They receive the highest number of I^l and I^m intensity peaks in growth time (see blue colored box plot in Fig. 7 (a) and (c)). Specifically, it gets a median of two peaks of I^l intensity in growth time (see blue colored box plot in Fig. 7 (a)).
2. *DR-ND:* They have an average initial time of 4 years and an average growth time of 25 years. No significant decay is seen for the period analyzed (see Table 5). Thus, they are defined only for the considered study period. Cumulative citation distribution in Fig. 6 (b) shows that they receive significant citations during their delayed growth period. Further, we find several I_l intensity

Table 5
(Long-term trajectory study) The descriptive statistics of initial time (T_i), growth time (T_g), and decay time (T_d) is shown for final clusters.

Long-term citation trajectory	Early Rise-Slow Decline			Delayed Rise-Not yet Declined			Delayed Rise-Slow Decline		
	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)	T_i (in yrs.)	T_g (in yrs.)	T_d (in yrs.)
Mean	2.14	4.41	20.82	4.06	25.46	0	4.73	16.06	7.84
Standard deviation	1.22	2.85	3.72	2.00	2.93	0	2.42	2.63	2.09
Quartile 1	1	3	18	2	22	0	2	6	5
Quartile 2	2	4	20	4	25	0	4	16	7
Quartile 3	3	6	22	4	26	0	5	17	7



(a) Early Rise-Slow Decline (b) Delayed Rise-Not yet Declined (c) Delayed Rise-Slow Decline

Fig. 6. The histogram plot shows the cumulative citation distribution of each cluster for three consecutive times – initial growth time, growth time, and decay time. Figures (a), (b), and (c) represent ER-SD, DR-ND, and DR-SD clusters for long-term trajectories, respectively.

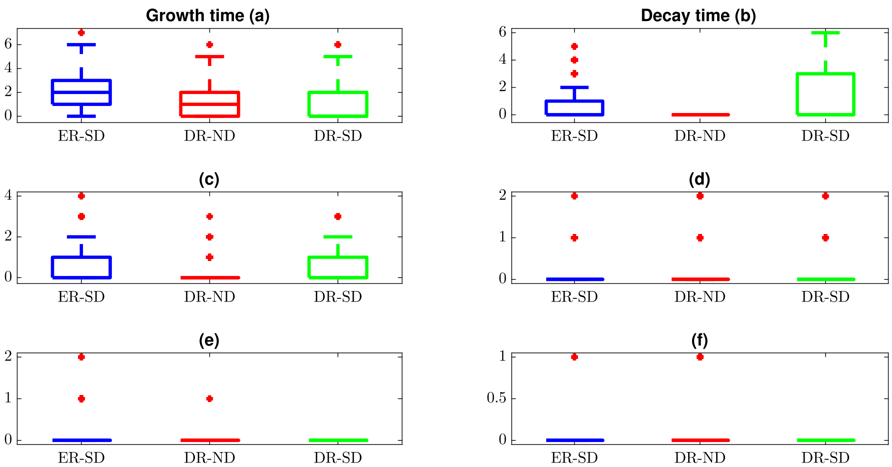


Fig. 7. The box plot represents the distribution of the number of citation peaks of three different intensities – n_{I^l} (figures (a), (b)), n_{I^m} (figures (c), (d)), and n_{I^h} (figures (e), (f)) separately for growth and decay times. Blue, red, and green indicate ER-SD, DR-ND, and DR-SD clusters.

peaks during growth time (see red colored box plot in Fig. 7 (a)). It receives a single median peak and up to 5 peaks in the growth period.

3. *DR-SD:* They have an average initial time of 5 years and an average growth time of 16 years. Consequently, the total growth time for papers in this cluster is 16 to 20 years, followed by an average slow decay in the next 14 years (see Table 5). Cumulative citation distribution in Fig. 6 (c) reveals that they receive negligible citations in the initial time, moderate citations in growth time, and a maximum proportion of citations in decay time. Low-intensity I^l citation peak values are prominently visible in decay time (see green colored box plot in Fig. 7 (b)). Table 3 reveals it is only a characteristic of long-term citation trajectories.

4.3. Statistical cluster validation

The analysis of variance (ANOVA) test is conducted to validate the final clusters statistically. (Tables 1 and 2) [2] represents the ANOVA test results for short and long-term trajectories, respectively. We compare the mean values of a feature from k-groups and check whether the difference is statistically significant. It separates the variance into two components due to – mean differences and random influences (Riffenburgh, 2012). We find that the final clusters in both studies are statistically significant (p-value$<$0.05), considering each feature separately. The largest F-values depicting feature importance are obtained for time-related features – *decay* and *growth time*.

[2] https://github.com/decodejoyita/Clustering-citation-trajectories/tree/main.

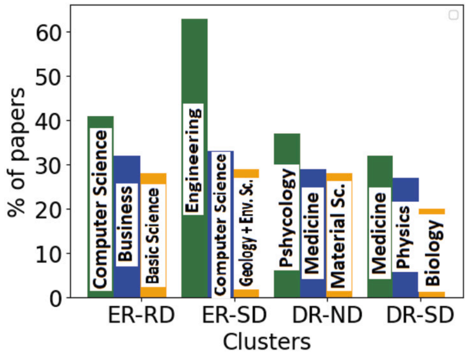


Fig. 8. The bar plot depicts the percentage of papers from top 3 disciplines in each cluster.

4.4. Cluster differences across research domains

The MAG data has four level categories for classifying domains (Sinha et al., 2015). Due to interdisciplinary research, multiple parent fields are tagged for a single article. At the top level, an article is tagged to 19 distinct disciplines. We consider this tagging for our analysis. To understand the disciplinary differences, we analyze the top 10 percentile papers filtered using the mean citation rate from each cluster. Papers from two data sets comprising short-term and long-term trajectories are combined. Fig. 8 depicts a bar plot showing the percentage of papers belonging to the top 3 disciplines in each cluster.

Some of the interesting findings are as follows. Broadly, we find that Computer Science (CS) papers mostly fall into either the ER-RD (41%) or ER-SD cluster (33%). Besides, the ER-RD class has 32% and 28% of papers in the Business and Basic Science fields (Physics, Chemistry, Biology, and Mathematics), respectively. In contrast, a majority of Medicine field papers belong to the DR-SD and DR-ND clusters. For instance, 29% of papers from the Genetics sub-field and 32% of papers from the Neuroscience sub-field belong to the DR-SD and DR-ND clusters, respectively. Moreover, the ER-SD class has 63% of papers from Engineering and 29% of papers from the Geology and Environmental Science domain. From the results, we can infer that applied domains such as CS have a majority of early-rise papers due to quick visibility, whereas papers from the medical field mostly experience delayed rise. This might be due to the time taken to obtain the clinical trial outcomes before a novel finding is recognized by the community.

4.5. Comparative study

This sub-section presents a qualitative comparison study between clusters to validate the proposed methodology. Table 6 examines how final cluster sets obtained in this study get mapped to identical clusters defined in prior literature. Here, a quantitative comparison is not feasible as different methods have different thresholds and parameter dependencies.

Aversa (1985) identified two clusters – 'Early Rise-Rapid Decline (ER-RD)' and 'Delayed Rise-Slow Decline (DR-SD).' ER-RD are defined as papers with a growth time of 3 years followed by a rapid decay. The exact decay time is not mentioned. DR-SD are defined as papers with a growth time of 6 years followed by a decline in the next 4 years. The ER-RD and DR-SD clusters are identical and align with *ER-RD* and *ER-SD* clusters obtained in this study.

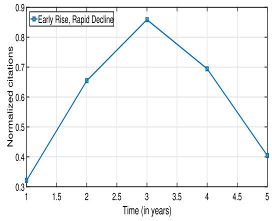
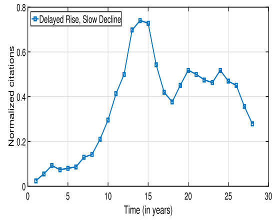
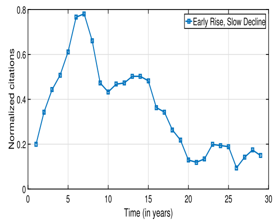
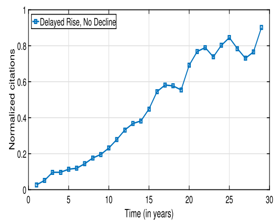
Aksnes (2003) identified three clusters – 'Early Rise-Rapid Decline (ER-RD),' 'Medium Rise-Slow Decline (MR-SD),' and 'Delayed Rise-No Decline (DR-ND).' The ER-RD and MR-SD clusters have a growth time of 2 to 3 years and 4 to 5 years, respectively. The exact decay time is not mentioned. DR-ND are defined as papers with a slow citation rise initially and receiving a significant proportion of citations only after 12 years. No decay is observed. The ER-RD, MR-SD, and DR-ND clusters are identical to this study's *ER-RD*, *ER-SD*, and *DR-ND* clusters.

Costas et al. (2010) identified 3 clusters – 'Flashes-in-the-Pan (FP),' 'Normal Documents (ND),' and 'Delayed Documents (DD).' FP is defined as papers with a growth time of 3 years; however, they are not cited in the long term. ND is defined as papers with a 4 to 5 years growth time followed by an exponential decay. The exact decay times are not mentioned. DD is defined as papers with a growth time of 10 to 11 years and receiving a significant proportion of their citations later than normal documents. They receive citations even after 20 years. The FP, ND, and DD clusters are identical to this study's *ER-RD*, *ER-SD*, and *DR-SD* clusters. Moreover, Redner (2004) identified three clusters – Sleeping Beauties (SB), Discovery Papers (DP), and Hot Papers (HP), respectively. The behavior of DP and HP aligns with the *ER-SD* cluster. Their growth time is 4 to 6 years. Besides, SBs align with the *DR-ND* and receive citations 40 years after publication.

Baumgartner and Leydesdorff (2014) identified two clusters – Transient-Knowledge-Claim (TKC) and Sticky-Knowledge-Claim (SKC). Papers belonging to TKC show a typical early peak in citations followed by a steep decline. Papers belonging to SKC have a growth time of 3 to 4 years, and they continue to be cited even after more than 10 years. The TKC and SKC are identical to this paper's *ER-RD* and *ER-SD* clusters, respectively.

Chakraborty et al. (2015) defined six clusters – PeakInit, MonDec, MonInc, PeakLate, PeakMult, and Others. PeakInit papers have a growth time of exactly 5 years. MonDec papers have a growth time of 1 year followed by a monotonic decrease. MonInc papers have a growth time of 20 years and no decay. PeakLate papers have a growth time of > 5 years. The exact decay times are not determined. 45% of the papers fell into the Others category, whose trajectory characteristics are not defined. The PeakInit, MonDec,

Table 6
The ER-RD and DR-SD clusters are aligned with identical trajectories in literature.

Final clusters	Brief description	Identical clusters in prior literature
Early Rise-Rapid Decline 	**Short-term trajectory** • *Growth time:* 3 years • *Decay time:* 2 years • *No. of papers:* 4,454	• Early rise, rapid decline (Aversa, 1985, Aksnes, 2003) • Flashes-in-the-pan (Costas et al., 2010) • Sprinters (Colavizza & Franceschet, 2016) • Transient-knowledge-claims (Baumgartner & Leydesdorff, 2014) • MonDec (Chakraborty et al., 2015) • Hot papers (Bornmann et al., 2018) • Smart girls (Ye & Bornmann, 2018)
Delayed Rise-Slow Decline 	**Long-term trajectory** • *Growth time:* 16 years • *Decay time:* 14 years • *No. of papers:* 336	• Revived classics (Redner, 2004) • PeakLate (Chakraborty et al., 2015) • Sleeping beauties (Van Raan, 2004, Ke et al., 2015, van Raan, 2021, Ye & Bornmann, 2018) • Delayed documents (Zhang et al., 2017)
Early Rise-Slow Decline 	**Long-term trajectory** • *Growth time:* 6 years • *Decay time:* 20 years • *No. of papers:* 17,479 **Short-term trajectory** • *Growth time:* 5 years • *Decay time:* 3 to 4 years • *No. of papers:* 87,956	• Delayed rise, slow decline (Aversa, 1985) • Discovery papers & hot papers (Redner, 2004) • Normal documents (Costas et al., 2010) • Medium rise, slow decline (Aksnes, 2003) • Middle-of-the-roads (Colavizza & Franceschet, 2016) • Sticky-knowledge-claims (Baumgartner & Leydesdorff, 2014) • PeakInit (Chakraborty et al., 2015) • Normal-low and Normal-high (Zhang et al., 2017)
Delayed Rise-Not yet Declined 	**Long-term trajectory** • *Growth time:* 29 years • *Decay time:* no decay • *No. of papers:* 23,917 **Short-term trajectory** • *Growth time:* 9 years • *Decay time:* no decay • *No. of papers:* 1,03,373	• Delayed rise (Costas et al., 2010) • Delayed rise, no decline (Aksnes, 2003) • Marathoners (Colavizza & Franceschet, 2016) • MonInc (Chakraborty et al., 2015) • Delayed recognition papers (Bornmann et al., 2018) • Evergreens (Zhang et al., 2017)

MonInc, and PeakLate clusters are identical to *ER-SD*, *ER-RD*, *DR-ND*, and *DR-SD* from this study. Gou et al. (2022) also identified the PeakMult cluster and studied it as a 'literature revival' phenomenon. All-Element-Sleeping-Beauties, a sub-category of SBs, are also a PeakMult cluster. However, we observed that papers of all trajectories receive multiple peaks. The only difference is the time of occurrence of peaks and their varying intensities in individual trajectories. For instance, *ER-SD* received a maximum number of peaks during the growth period, and DR-ND received them during the decay period. Thus, it is one of the inherent properties of a trajectory.

Colavizza and Franceschet (2016) identified 3 clusters – sprinters, middle-of-the-roads, and marathoners. The exact growth and decay times are not explicitly mentioned. Sprinters are defined as papers with fast and high peak values followed by equally rapid aging. Middle-of-the-roads are defined as papers that attain fast but moderate peaks with a gradual decay over time. Marathoners are defined as papers that start slow, peak moderately, keep receiving a higher proportion of citations over a prolonged time, and, finally, citations decline slowly. The sprinters, middle-of-the-roads, and marathoners are identical with *ER-RD, ER-SD,* and *DR-SD* clusters of this study.

Zhang et al. (2017) identified four clusters – normal low, normal high, delayed documents, and evergreens. Papers belonging to normal low and normal high clusters receive an early peak followed by a slow decline. Delayed documents are papers with a slow citation rise followed by a slow decay. Evergreens are papers with a continual increase in citations and no decay in the 30 years analyzed. The normal-low and normal-high clusters are similar to *ER-SD,* delayed documents are identical to *DR-SD,* and evergreens align with *DR-ND* clusters of this study. Ye and Bornmann (2018) identified two clusters – Smart Girls (SG) and Sleeping Beauties (SB). SGs are papers with a growth time of 5 years, and the citation angle is $> 60°$. SBs are papers with a > 5 years growth time, and the citation angle is $> 30°$. SBs receive a major citation proportion after 15 years. The SGs and SBs are identical with *ER-RD* and *DR-SD* of this study.

Bornmann et al. (2018) identified two clusters – Hot Papers (HP) and Delayed Recognition (DR). The HP and DR are identical to *ER-RD* and *DR-SD* clusters from this study. Besides, many works only study an extreme trajectory, that is, sleeping beauties (Van Raan, 2004, van Raan, 2021, Li & Ye, 2012, Li et al., 2014, Ke et al., 2015). Such papers receive negligible citations for a long time after publication and then, depending upon the awakening intensity, suddenly jump to receive large citations. The decay time is characterized by lower annual citations than peak (Bornmann et al., 2018). It aligns with the *DR-SD* cluster of this study.

To summarize, the qualitative comparison validates our proposed methodology. We can detect all probable classes of trajectories identified in the existing literature.h Table 6 shows that we can universally categorize trajectories into four clusters – *ER-RD, ER-SD, DR-ND,* and *DR-SD*. Papers with an *ER-RD* trajectory have a growth period of 3 years and decay in the next 2 years. Short-term trajectories mostly exhibit such a pattern. They receive n_{I^l} peaks during the growth period. Besides, both short-term and long-term trajectories exhibit *ER-SD* patterns. Compared to the other two clusters, short-term trajectories with an *ER-SD* pattern receive maximum citations. They have a growth period of 5 years and decay in the next 3 to 4 years. Further, long-term citation trajectories with an *ER-SD* pattern have a growth period of 6 years and decay in the next 20 years. Compared to all other clusters, they receive a maximum n_{I^l} and n_{I^m} peaks during the growth period.

Both short-term and long-term trajectories also exhibit *DR-ND* patterns. Short-term trajectories with a *DR-ND* pattern have a delayed growth period of 9 years. Further, long-term citation trajectories with a *DR-ND* pattern have a growth period of 29 years and receive maximum citations compared to the other two clusters. No citation decline is seen for the period analyzed. Finally, papers with a *DR-SD* trajectory have a growth period of 16 years and decay in the next 14 years. Long-term trajectories mostly exhibit such a pattern. They receive a maximum n_{I^l} peaks during decay.

5. Discussion and conclusion

Most existing literature has conceptually defined the types of trajectory patterns, corresponding features, and their threshold values. It defers largely from one study to another, even for identifying similar patterns. Consequently, it results in an inconsistent number and definition of trajectory patterns. This work proposes the MKMCE algorithm for clustering citation trajectories. Motivated by prior literature, we extract nine features from a trajectory to comprehensively capture growth and decay profile of a paper. Instead of pre-defining the k-value, the MKMCE algorithm iteratively checks the credibility of cluster labels based on the ϵ neighborhood. Finally, we algorithmically obtain the number of clusters (k). Four distinct clusters are obtained- *ER-RD, ER-SD, DR-ND,* and *DR-SD*. Thus, it removes ambiguity in the number of distinct trajectory patterns. The proposed algorithm takes linear time and is ideal for clustering large-sized data sets.

Of them, *ER-RD, ER-SD, and DR-ND,* are specific to the short-term trajectories. Besides, *ER-SD, DR-ND,* and *DR-SD* are specific to the long-term trajectories. Performing feature analysis after clustering, we empirically define citation growth and decay characteristics of individual clusters and set data-driven thresholds. Thus, our proposed framework leads to a more generalized, robust, and stable cluster set. Unlike previous approaches, it is not limited to identifying only a specific subset of papers. It eliminates inconsistencies in the choice of features and thresholds across multiple studies.

We find that most papers fall into *ER-SD* and *DR-ND* clusters. A negligible share of articles fall into *ER-RD* and *DR-SD*. Further, the *ER-RD* is a characteristic of short-term trajectories, and *DR-SD* is a characteristic of mostly long-term trajectories. Delayed-rise papers receive higher total citations than early risers as they receive citations for a more extended period. However, multiple peaks are detected highest for early risers, establishing that the citations' intensity is higher for them. A comprehensive comparative study with prior literature reveals that the four clusters can capture all random groups and sub-groups of trajectory patterns.

This study has several limitations. *First,* we could not identify extremely specialized trajectories, such as sleeping beauties. It is also because they rarely occur. Also, this study captures generalized trajectory patterns of which SBs may be sub-groups. For instance, the *DR-SD* cluster closely resembles SBs; however, we cannot empirically show it. *Second,* we cannot determine the fate of articles belonging to the *DR-ND* cluster. It is uncertain whether they will continue to be highly cited or become obsolete in the future. *Third,* self-citations are not excluded from our analyses.

The outcome of our study in terms of clustering citation trajectories using the MKMCE algorithm might have the following applications. *First,* one can apply our proposed methodology hierarchically to break down further the relatively larger clusters like ER-SD and DR-ND obtained in this study into sub-clusters at multiple levels. Thus, an automated taxonomy of clusters will be generated. A measure of inter-cluster similarity can correlate with various other clusters reported in all the previous studies, thereby identifying generalized (multiple-cluster) and specialized (single-cluster) patterns.

Second, growth and decay are the two fundamental features of a trajectory. An effective citation span or impact lifecycle of a publication can be well understood using those features. Our study shows that the growth and decay time varies across clusters. The thresholds defined in previous studies are inconsistent and hold only for a subset of papers studied by the authors. Unlike previous approaches, our proposed framework empirically defines feature thresholds only after obtaining clusters for different trajectory patterns. Hence, they are more generalized, consistent, and reported only after a thorough comparison with previous studies. Scientific paper recommendation algorithms can use such precise threshold information and recommend articles based on the cluster a paper belongs to. For instance, based on its growth time, a recommendation algorithm may foresee its popularity/impact and recommend it early. Similarly, based on its decay time, it may lower its weightage after a certain point. Hence, one can have more effective and consistent recommendations, saving researchers a lot of time and energy.

Third, we can easily define a range or confidence interval around citation-related (features F5 and F6) and peak count-related (features F7, F8, and F9) (refer to Table 2). It will help us to identify a paper's highest citation activity period in its entire lifecycle. Thus, by understanding the evolution of the impact of a publication over time, the idea can be further extended to identify impactful authors, journals, and research domains. For example, analysis of delayed-rise papers can help us to understand new emerging research domains in the present time. Researchers in their early career stages can benefit from such insights.

Fourth, existing popular metrics for measuring research impact, such as the h-index (Hirsch, 2005), i-10 index (Mester, 2015) for authors, and journal impact factor (Garfield, 2006) for journals, are based on citation count as a primary parameter. However, they suffer from several loopholes that researchers often exploit to inflate their quality falsely. It includes excessive self-citations added by authors, citation stacking, and cartels formed by a group of authors, editors, and publishers. Recently, the "Genetika" journal[3] was banned from indexing by the Clarivate Analytics firm for practising citation stacking. In return, the journal has retracted 31 papers because of compromised peer review and to maintain its reputation. A thorough feature analysis of the clusters identified in this study can help us to identify novel features derived from temporal citation evolution that are more robust to manipulation than single-valued citation counts. Novel metrics, free from the existing limitations might emerge from such studies which are to quantify the quality of authors, journals, and research domains.

CRediT authorship contribution statement

Joyita Chakraborty: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Dinesh K. Pradhan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. **Subrata Nandi:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – review & editing.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.
Aversa, E. (1985). Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics*, *7*(3–6), 383–389.
Baghizadeh, M., Maghooli, K., Farokhi, F., & Dabanloo, N. J. (2020). A new emotion detection algorithm using extracted features of the different time-series generated from st intervals Poincaré map. *Biomedical Signal Processing and Control*, *59*, Article 101902.
Bai, L., Liang, J., & Cao, F. (2020). A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Information Fusion*, *61*, 36–47.
Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, *7*, 9324–9339.
Baumgartner, S. E., & Leydesdorff, L. (2014). Group-based trajectory modeling (GBTM) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims". *The Journal of the Association for Information Science and Technology*, *65*(4), 797–811.
Bjork, S., Offer, A., & Söderberg, G. (2014). Time series citation data: The Nobel prize in economics. *Scientometrics*, *98*(1), 185–196.
Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, *54*(3), 1–33.
Bornmann, L., Ye, A. Y., & Ye, F. Y. (2018). Identifying "hot papers" and papers with "delayed recognition" in large-scale datasets by using dynamically normalized citation impact scores. *Scientometrics*, *116*, 655–674.
Chakraborty, J., & Pradhan, D. K. (2022). Citation biases: Detecting communities from patterns of temporal variation in journal citation networks. In *International conference on data management, analytics & innovation* (pp. 591–611). Springer.
Chakraborty, J., Pradhan, D. K., & Nandi, S. (2021). On the identification and analysis of citation pattern irregularities among journals. *Expert Systems*, *38*(4), Article e12561.
Chakraborty, J., Pradhan, D. K., & Nandi, S. (2022). Research misconduct and citation gaming: A critical review on characterization and recent trends of research manipulation. In *Data management, analytics and innovation: Proceedings of ICDMAI 2021, Vol. 2* (pp. 485–492).
Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015). On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, *58*(9), 82–90.
Chi, Y., Tang, X., & Liu, Y. (2022). Exploring the "awakening effect" in knowledge diffusion: A case study of publications in the library and information science domain. *Journal of Informetrics*, *16*(4), Article 101342.
Clermont, M., Krolak, J., & Tunger, D. (2021). Does the citation period have any effect on the informative value of selected citation indicators in research evaluations? *Scientometrics*, *126*, 1019–1047.
Colavizza, G., & Franceschet, M. (2016). Clustering citation histories in the physical review. *Journal of Informetrics*, *10*(4), 1037–1051.
Comins, J. A., & Leydesdorff, L. (2017). Identification of long-term concept-symbols among citations: Do common intellectual histories structure citation behavior? *The Journal of the Association for Information Science and Technology*, *68*(5), 1224–1233.
Costas, R., Van Leeuwen, T. N., & Van Raan, A. F. (2010). Is scientific literature subject to a 'sell-by-date'? A general methodology to analyze the 'durability' of scientific documents. *Journal of the American Society for Information Science and Technology*, *61*(2), 329–339.
Garfield, E. (1989). Delayed recognition in scientific discovery-citation frequency-analysis aids the search for case-histories. *Current Contents*, *23*, 3–9.
Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93.
Golosovsky, M., & Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E*, *95*(1), Article 012324.

[3] https://retractionwatch.com/2023/12/12/journal-retracts-31-papers-bans-authors-and-reviewers-after-losing-its-impact-factor/.

Gou, Z., Meng, F., Chinchilla-Rodríguez, Z., & Bu, Y. (2022). Encoding the citation life-cycle: The operationalization of a literature-aging conceptual model. *Scientometrics*, *127*(8), 5027–5052.

Haghighat, M., & Hayatdavoudi, J. (2021). How hot are hot papers? The issue of prolificacy and self-citation stacking. *Scientometrics*, *126*, 565–578.

He, Z., Lei, Z., & Wang, D. (2018). Modeling citation dynamics of "atypical" articles. *The Journal of the Association for Information Science and Technology*, *69*(9), 1148–1160.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431.

Lange, L. L. (2005). Sleeping beauties in psychology: Comparisons of "hits" and "missed signals" in psychological journals. *History of Psychology*, *8*(2), 194.

Laptev, N., Amizadeh, S., & Flint, I. (2015). Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1939–1947).

Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". *Scientometrics*, *100*(2), 595–601.

Li, J., & Shi, D. (2016). Sleeping beauties in genius work: When were they awakened? *The Journal of the Association for Information Science and Technology*, *67*(2), 432–440.

Li, J., & Ye, F. Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics*, *92*(3), 795–799.

Li, J., Shi, D., Zhao, S. X., & Fred, Y. Y. (2014). A study of the "heartbeat spectra" for "sleeping beauties". *Journal of Informetrics*, *8*(3), 493–502.

Mester, G. (2015). New trends in scientometrics. In *Papers of 33rd international scientific conference "science in practice"* (pp. 22–27).

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L., & Sun, J. (2018). Innovation or imitation: The diffusion of citations. *The Journal of the Association for Information Science and Technology*, *69*(10), 1271–1282.

Min, C., Bu, Y., Wu, D., Ding, Y., & Zhang, Y. (2021). Identifying citation patterns of scientific breakthroughs: A perspective of dynamic citation process. *Information Processing & Management*, *58*(1), Article 102428.

Mingers, J. (2007). Shooting stars and sleeping beauties: The secret life of citations.

Pradhan, D. K., Chakraborty, J., & Nandi, S. (2019). Applications of machine learning in analysis of citation network. In *Proceedings of the ACM India joint international conference on data science and management of data* (pp. 330–333).

Pradhan, D. K., Chakraborty, J., Choudhary, P., & Nandi, S. (2020). An automated conflict of interest based greedy approach for conference paper assignment system. *Journal of Informetrics*, *14*(2), Article 101022.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E*, *83*(4), Article 046116.

Redner, S. (2004). Citation statistics from more than a century of physical review. arXiv preprint, arXiv:physics/0407137.

Riffenburgh, R. H. (2012). *Statistics in medicine*. Academic Press.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Van Dalen, H. P., & Henkens, K. n. (2005). Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics*, *64*, 209–233.

Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

van Raan, A. F. (2021). Sleeping beauties gain impact in overdrive mode. *Scientometrics*, *126*(5), 4311–4332.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2023). Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *The Journal of the Association for Information Science and Technology*, *74*(2), 150–167.

Wei, C., Li, J., & Shi, D. (2023). Quantifying revolutionary discoveries: Evidence from Nobel prize-winning papers. *Information Processing & Management*, *60*(3), Article 103252.

Xiong, H., Wu, J., & Chen, J. (2006). K-means clustering versus validation measures: A data distribution perspective. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 779–784).

Xu, H., Luo, R., Winnink, J., Wang, C., & Elahi, E. (2022). A methodology for identifying breakthrough topics using structural entropy. *Information Processing & Management*, *59*(2), Article 102862.

Yang, J., Bu, Y., Lu, W., Huang, Y., Hu, J., Huang, S., & Zhang, L. (2022). Identifying keyword sleeping beauties: A perspective on the knowledge diffusion process. *Journal of Informetrics*, *16*(1), Article 101239.

Ye, F. Y., & Bornmann, L. (2018). "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle. *The Journal of the Association for Information Science and Technology*, *69*(3), 359–367.

Zamani, M., Aghion, E., Pollner, P., Vicsek, T., & Kantz, H. (2021). Anomalous diffusion in the citation time series of scientific publications. *Journal of Physics: Complexity*, *2*(3), Article 035024.

Zhang, R., Wang, J., & Mei, Y. (2017). Search for evergreens in science: A functional data analysis. *Journal of Informetrics*, *11*(3), 629–644.