Clustering citation histories in the Physical Review

Clustering citation histories in the Physical Review

Giovanni Colavizza*
Digital Humanities Laboratory
École Polytechnique Fédérale de Lausanne – Switzerland
giovanni.colavizza@epfl.ch

Massimo Franceschet
Department of Physics, Mathematics and Computer Science
University of Udine – Italy
massimo.franceschet@uniud.it

June 30, 2016

Abstract

We investigate publications trough their citation histories – the history events are the citations given to the article by younger publications and the time of the event is the date of publication of the citing article. We propose a methodology, based on spectral clustering, to group citation histories, and the corresponding publications, into communities and apply multinomial logistic regression to provide the revealed communities with semantics in terms of publication features. We study the case of publications from the full Physical Review archive, covering 120 years of physics in all its domains. We discover two clear archetypes of publications – marathoners and sprinters – that deviate from the average middle-of-the-roads behaviour, and discuss some publication features, like age of references and type of publication, that are correlated with the membership of a publication into a certain community.

Keywords: Citation histories, Clustering, Regression analysis, Physical Review.

1 Introduction

In bibliometrics, the number of citations received by a publication is a rough indicator of the impact of the work among its peers. Several more elaborated

*Corresponding author. Please address all correspondence to Giovanni Colavizza, EPFL, Digital Humanities Laboratory, Station 14, INN 116, CH-1015 Lausanne.

citation measures have been proposed. All of them, regardless of the complexity of their defining formulas, assign a publication with a single rating, so that a total ranking among a set of publications can be compiled.

In this paper, we take a different perspective: *citation temporalization*, by considering citation histories [Redner, 2004]. There are two main approaches to study the citation history of a publication: synchronous and diachronous. The former approach focuses on the distribution of the publication years of cited publications, the latter on the distribution of received citations over time [Nakamoto, 1988]. We mainly focus on the the latter: instead of the single number of citations received by a publication at a given time, we consider the full citation history of the publication since its origin. More precisely, the *citation history* of publication i is a vector

$$h_{i,*} = (h_{i,1}, h_{i,2}, \ldots, h_{i,m}),$$

where $h_{i,j}$ is the number of citations received by i during period j and $T = (1, 2, \ldots, m)$ is a series of consecutive temporal periods in some time granularity (e.g., months or years), where we assume 1 to be the period of publication of i. Citation histories extend citation counts by adding a temporal dimension, providing a more informative and less immediate indication of the impact of a publication. While citation counts are snapshots of publication impact at a given time, citation histories move publication impact over time and map a publication's ageing process.

The study of patterns of ageing of scientific publications has been very active since decades, its main focus being understanding scientific discourse in different fields and times, and the determinants of the success of a publication. As an example, this problem was posed by Garfield [1980] as one of trying to individuate publications subject to delayed recognition or premature discovery. He framed the task in the following steps: understanding (i) what is a typical citation pattern for every scientific field; (ii) what is a deviation from this typical citation pattern; and (iii) what really qualifies as a premature discovery. Naturally, delayed recognition is but one of the citation patterns which deviate from the typical one. The ageing of scientific literature has also been compared more generally to the process of obsolescence of any kind of phenomena [Pollmann, 2000]. The average or typical citation history is linked with information diffusion processes, where several effects interact in causing a considerable amount of information items to go unnoticed, others to be considered for a short amount if time and then fade out (causing the attention peak), others still to remain relevant for longer, even indefinitely (causing the long tail). The speed of recognition, if any, is also driven by intrinsic as well as extrinsic factors (cf. e.g. Line and Sandison [1974]). For example, curves similar to archetypal citation histories are to be found in the proportion of re-shares of Facebook photos during the first hours since upload: even identical photos were found to be associated with very different diffusion "histories" [Dow et al., 2013].

Citation histories are not meant to rank publications in a compilation. Nevertheless, citation histories associated with different publications can be compared in a more involved way with respect to a total ordering relation. In this

paper, we use *clustering techniques* to group citation histories, and hence their corresponding publications, into a set of clusters or communities. Each cluster corresponds to a set of publications with similar citation histories. Hence, the total, non-symmetric ordering relation used to rank publications with citation counts is substituted with a symmetric similarity relation that prescribes which publication belongs to which community. Each community can be represented with its average citation history – we call this aggregated history the *citation macrohistory* of the cluster. Different clusters correspond to different citation macrohistories, and the flexibility of hierarchical clustering methods allows us to tune the granularity of clusters and hence to calibrate the degree of difference of the corresponding macrohistories. Furthermore, we identify a set of determinants, that is independent variables such as the number of received citations, the number of references, the age of references, the number of authors, the length of a publication, and the publication year and type. We use these determinants to elucidate the membership of a publication to a given cluster, in order to provide each cluster with semantics in terms of publication characteristics. We apply the described methodology to the full Physical Review archive, containing more than half a million publications spanning all domains of physics during the last 120 years.

The layout of the paper is as follows. We describe the methodology proposed in this work in Section 2. In particular, we define histories and macrohistories in Section 2.1, we describe the clustering methods adopted to group citation histories in Section 2.2, we discuss the many choices of our experimental setting in Section 2.3, and briefly present the Physical Review dataset in Section 2.4. Section 3 contains the main results of the application of the method to the dataset. In particular, Section 3.1 is devoted to the results of clustering and Section 3.2 identifies the determinants for the detected clusters. Section 4 compares the present work with related literature and Section 5 concludes and outlines further directions of research.

2 Methodology

In this section we formally discuss citation histories as a way to recover the temporalization in received citations. We also introduce and motivate the choice of spectral clustering in order to cluster publications according to their citation histories, present our experimental setup and briefly describe the Physical Review dataset, which will be used as a case study.

The matrix notation we use in this paper is the following: given a matrix A and two valid indices i and j, the entry of A in row i and column j is denoted by $a_{i,j}$. The i-th row of A is denoted by $a_{i,*}$ and the j-th column of A is denoted by $a_{*,j}$.

2.1 Citation histories

The *citation history* of a publication P tracks the citations that P received since its origin (the date of publication). The events composing this special history are the citations given by younger publications Q towards P, the time of the event being the date of publication of the citing article Q. Suppose, for instance, that P is published in year 2011 and now is end of 2015. If P received 5 citations in 2011, 10 citations in 2012, 3 citations in 2013, 2 citations in 2014, and no citations in 2015, then the citation history of P, using a yearly temporal granularity, is the vector $h_P = (5, 10, 3, 2, 0)$. Notice that sum of the citation history vector components corresponds to the total number of citations accrued by P at the present moment (20 in the example). A publication brought out before P has a longer history, while a publication issued after P has a shorter history.

Formally, let i be a publication, $m \geq 1$ be an integer and $T = (1, 2, \ldots, m)$ be a series of consecutive temporal periods in some time granularity (e.g., month or year), where we assume 1 to be the period of publication of i. For every $j \in T$, let $h_{i,j}$ be the (non-negative integer) number of citations received by i during period j. The citation history of publication i over T is the following vector:

$$h_{i,*} = (h_{i,1}, h_{i,2}, \ldots, h_{i,m})$$

In the following, we discuss relevant choices for the definition of a proper citation history. First of all, what is the *minimum length* of a history? And, related to history length, what is the *minimum number of events* (citations) that define a history? If we want to evaluate an object (publication) according to its history, a minimum length has to be imposed. If, otherwise, histories are used only for aggregated histories, then any length might be acceptable, with the caveat of using histories of equal length for comparison. It seems reasonable, moreover, that a history, to be considered as such, has a minimum number of citations. This threshold can be fixed in advance (eg., 20), or can be a function of the history length, for instance, a history of length t must contain at least t citations (on average, one per temporal period, which is the global average number of citations per paper suggested by De Solla Price [1965]).

Another relevant choice for the definition of citation histories is *temporal granularity*. Granularities, which are intrinsic to temporal data, provide a mechanism to hide details that are not known or not pertinent for an application [Bettini et al., 2009]. Day, month, and year are examples of temporal granularities related to the Gregorian calendar. Usually, we know the year of publication, and sometimes the month. Hence year or month might be suitable temporal granularities for citation histories. Working with a finer granularity is preferable since it enhances precision of the analysis; nevertheless, the dataset is larger and hence more computationally intensive. Hence, a compromise between precision and complexity is necessary.

A final issue about citation histories is *normalization*: do we use raw citation counts as elements of the history vector of do we normalize them by dividing by the total number of citations contained in the history? Given a publication

i, let us define

$$c_i = \sum_{t=1}^{m} h_{i,t}$$

as the total number of citations accrued by publication i. Normalization entails defining probabilities $p_{i,t} = h_{i,t}/c_i$ that publication i receives a citation during time t as well as the normalized history as follows:

$$p_{i,*} = (p_{i,1}, p_{i,2}, \ldots, p_{i,m})$$

Notice that for every $t \in T$ we have $0 \leq p_{i,t} \leq 1$ and $\sum_{t=1}^{m} p_{i,t} = 1$, hence the normalized history $p_{i,*}$ is indeed a probability distribution. The advantage of working with normalized histories is that we can compare two or more histories on the same playground. For example, consider two citation histories $h_{i,*} = (5, 10, 3, 2, 0)$ and $h_{j,*} = (15, 30, 9, 6, 0)$ of publications i and j. The raw citation values are quite different (citation counts of j are exactly 3 times those of i). Nevertheless, the normalized citation histories are the same $p_{i,*} = p_{j,*} = (0.25, 0.50, 0.15, 0.10, 0)$; for instance, there is the same probability of 1/2 that both i and j receive a citation during the second temporal period. The disadvantage of normalized histories is that the magnitude of citations is lost. For instance, considering only the two normalized histories, it is not clear anymore that j received much more citations than i.

Given a set of n publications with citation histories defined over m temporal periods, we can collect all citation histories in a *citation history matrix* H of size $n \times m$ such that $h_{i,j}$ is the number of citations received by publication i in period j. Normalization can be easily defined as a transformation of matrix H. Let $r = (r_1, \ldots, r_n)$ be the row sum vector of H, that is, r_i is the sum of the i-th row of H. Let R be a diagonal matrix with vector r on the diagonal. Then, the normalized citation history matrix is $R^{-1}H$, where R^{-1} is a diagonal matrix with elements $1/r_i$ on the diagonal.

After normalization has been applied to the history matrix, let H be the resulting $n \times m$ history matrix. We can finally compute the aggregated citation history for all publications contained in H, that is, what we call the *citation macrohistory*. The citation macrohistory of H is a vector $\Omega_H = (\mu_1, \ldots, \mu_m)$ such that

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} h_{i,j}$$

that is, μ_j is the average number of citations received by publications in H during period j. Notice that, if H is a normalized history matrix, then for every j we have $0 \leq \mu_j \leq 1$ and

$$\sum_{j=1}^{m} \mu_j = \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} h_{i,j} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} h_{i,j} = \frac{1}{n} \cdot n = 1$$

hence Ω_H is a probability distribution like the rows of H. Examples of macrohistories of different lengths are given in Figure 1.

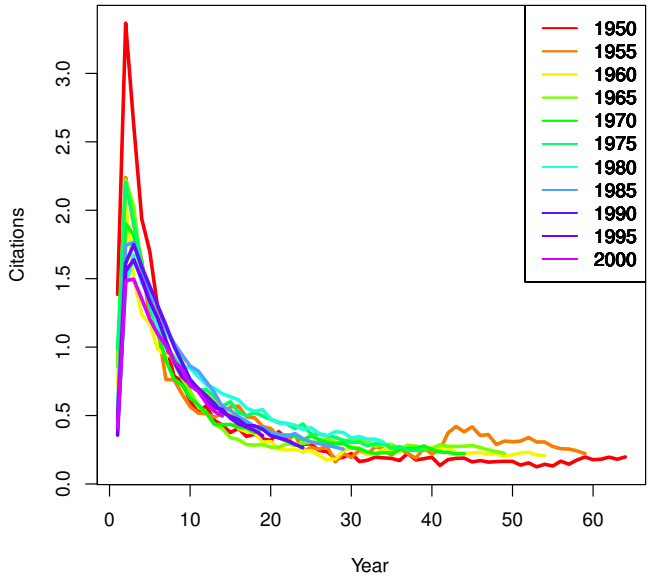


Figure 1: Macrohistories of all publications published in the given year in Physical Reviews that received at least 20 citations using year granularity.

2.2 Clustering method

Given the citation macrohistory of a set of articles, representative for example of a journal or an academic field, the following question is meaningful: if we split the whole macrohistory in a set of macrohistories, in order to group articles with a citation history of similar shape, what would be the outcome? A pulverization into small groups or a set of few large clusters? Previous work has mostly relied on heuristics or external variables (e.g. venue or field) in order to group articles into similarly cited groups. We propose to automatize this procedure by using clustering techniques and systematically analyze the outcome in order to see what patterns determine increasingly refined and smaller cluster of citation histories.

We begin by defining our task as one of finding if two empirical probability distributions are the same or not. In our case the probability distributions to compare are the macrohistories of increasingly smaller clusters of articles: given a suitable clustering procedure, we keep increasing the number of clusters until the most similar pair of cluster macrohistories is statistically indistinguishable. Several clustering methods exist in the literature, for what follows we adopt spectral clustering, specifically Normalized Cuts (Ncut) [Shi and Malik, 2000,

Yu and Shi, 2003]. Ncut is often found to be superior to traditional methods such as k-means, and to other spectral clustering approaches, such as unnormalized spectral clustering. For more details on spectral clustering and its experimental tuning, refer to von Luxburg [2007].

Define $H \in \mathbb{R}^{n \times m}$ as the row-stochastic matrix of filtered and normalized citation histories $h_{i,*}$ to analyze, according to the definitions given in Section 2.1. In what follows we assume all individual histories to be of equal length m, therefore filtering out (at least) articles published after $end - m$, where end is the last year represented in the dataset ($m \times 12$ if we consider months and not years). Intuitively, the goal of spectral clustering is to divide datapoints according to their pairwise similarity, and do this through the definition of a weighted graph and its partition into communities. The first step of spectral clustering is therefore the definition of a distance matrix $V \in \mathbb{R}^{n \times n}$ among data points (citation histories in our case), in order to reformulate clustering as a graph partition problem: finding groups such that edges among them have low weights and edges within the group have high weights. Each element $v_{i,j}$ in V is a distance between data points (histories) $h_{i,*}$ and $h_{j,*}$. The construction of a distance matrix requires the choice of a metric: common options are Euclidean, radial and cosine distances. In order to have a graph, represented as a (possibly sparse) weighted adjacency matrix $W \in \mathbb{R}^{n \times n}$, we also need a construction method in order to establish edges between similar datapoints: common choices are the ϵ-neighborhood (keep all the edges between a node and its neighbors at a distance $\leq \epsilon$), the r-nearest neighborhood (keep all edges between a node and its r nearest neighbors), and a fully connected graph.

Ncut approximates a solution to the optimization problem of finding a number k of partitions of W by balancing two divergent objectives: the *cut operator*, defined as the sum of the weights of the edges between the vertices of any cluster and the rest of the vertices in W, which we want to minimize, and the *volume operator* of each cluster, which is the sum of the weighted degrees of its vertices, which we want to maximize. The motivation behind Ncut is to find cohesive clusters which are not too unbalanced in the number of vertices. The relaxed solution to the Ncut problem is equivalent to finding the first k (smallest) eigenvalues, and corresponding eigenvectors $x_1, \ldots, x_k$, of the random walk normalized Laplacian of W, defined as $L = I - D^{-1}W$, where I is the identity matrix and D a diagonal matrix with $d_{i,i}$ equal to the sum of the i-th row of W (that is, the weighted degree of node i in the graph of W). Let $Z \in \mathbb{R}^{n \times k}$ be the matrix whose j-th column $z_{*,j} = x_j$. The method uses each row $z_{i,*}$ of Z as a representant of row $h_{i,*}$ of H and then takes advantage of k-means to cluster the rows $z_{i,*}$ of Z into k clusters. Finally, it assigns citation history $h_{i,*}$ to the cluster of its representant $z_{i,*}$.

In order to find the appropriate number of clusters for a given dataset, we propose an iterative method starting with a number of clusters $k = 2$, and increasing their amount until the citation macrohistories of the two most similar clusters are found to be not significantly different by a Kolmogorov–Smirnov two-sample test (KS test). The desired level of significance is a parameter of the algorithm. We settle for the value of k just before this event happens. This

iterative method is especially amenable to clustering approaches which require the number of clusters k to be specified, but it can nevertheless be applied to approaches which do not require k to be specified, by helping tuning their parameters instead. Such is for example the case of the Louvain community detection method by Blondel et al. [2008], using its resolution parameter. Other methods to chose a good number of clusters k are possible, for example the eigengap heuristic [von Luxburg, 2007], which nevertheless gave ambiguous and unhelpful results in our setting.

2.3 Experimental setup

We summarize here our experimental setup ad parameter choices. To begin with, we filtered the articles to be considered for analysis using a method suggested by Radicchi and Castellano [2011]. Define the relative success s_i of an article i using the following ratio:

$$s_i = \frac{c_i}{max(\mu, 5)} \tag{1}$$

where c_i is the number of citations accrued to article i during the time window of interest, and μ is the mean number of citations accrued to articles from the same year of publication during the same time window. We use $max(\mu, 5)$ at denominator of the success formula in order to account for years with low average citation rates. We consider for analysis only articles with a value $s_i \geq 1$, which means an above average relative performance. This technique can be adopted in order to select articles for analysis before calculating their history, and so doing guarantee a minimum amount of received citations from which to analyze the citation history. Other methods could also be used in order to determine an above norm performance, such as the median [Lin et al., 2016]. In this way we can account for the varying amount of citations papers receive due to the time of publication, while at the same time maintain a minimum threshold to produce meaningful citation histories. Crucially, we always consider histories of equal length for clustering, and normalize them as described in Section 2.

Ncut entails several design choices, which we detail here along a summary of its steps. First we construct a symmetric distance matrix V, where $v_{i,j}$ is the distance between histories $h_{i,*}$ and $h_{j,*}$. We use the Euclidean metric, which anyway yielded similar results to the cosine and radial alternatives. Secondly, we build the symmetric weighted adjacency matrix W. To be sure, we would like edges to be weighted proportionally to the similarity of the two datapoints: the lower the distance between data points, the more similar two data points are, the higher the weight of the corresponding edge. This is why W is sometimes called affinity matrix. In order to build our affinity matrix, we follow and extend a technique proposed by Zelnik-Manor and Perona [2004]. First of all, we keep only the r-nearest neighbors j for every datapoint i in V. We settle for a number of neighbors $r = 100$, after experimenting r from 25 to 500. Then, we define:

$$w_{i,j} = exp\left(- \frac{v_{i,j}^2}{\delta^2} \right) \tag{2}$$

for pairs i, j that are neighbors, and $w_{i,j} = 0$ otherwise.

The quantity δ is a scale parameter, normally taken to be global for every datapoint, e.g. using the mean of distances in V. Zelnik-Manor and Perona [2004] proposed to use instead a local scaling parameter δ_i, which has the benefit of considering the local statistics of the neighborhood of every datapoint i. The local scaling, which we adopt, entails changing the above definition to

$$w_{i,j} = exp\left(- \frac{v_{i,j}^2}{\delta_i \delta_j} \right) \tag{3}$$

for pairs i, j that are neighbors, with δ_i defined as the maximum distance between i and any of its r neighbors. Intuitively, local scaling allows neighborhoods at different relative distances to be weighted similarly in W, therefore improving the result of clustering.

Given the weighted adjacency matrix W, we apply the iterative process of increasing the number of clusters k from 2 to the maximum significant value, according to the procedure described in Section 2.2. We use a significance level of 0.1 to be rather tolerant with the desired number of clusters. The KS test's p-value is monotonically increasing in all experiments we did, as the number of clusters is raised until reaching non-significance.

To summarize, the steps we follow to cluster the citation history matrix H are:

1. compute the distance matrix V, where $v_{i,j}$ is the Euclidean distance among histories $h_{i,*}$ and $h_{j,*}$;
2. set the number of neighbors $r = 100$ and compute the weighted adjacency matrix W using Eq. 3;
3. construct the random walk normalized Laplacian $L = I - D^{-1}W$;
4. set the initial number of clusters $k = 2$;
5. find the first k eigenvectors $x_1, \ldots, x_k$ of L and put them as columns of a matrix Z;
6. cluster the rows $z_{i,*}$ of Z with k-means, to find the required k clusters, and put the history $h_{i,*}$ in the cluster of $z_{i,*}$;
7. evaluate the quality of the clustering by comparing the macrohistories of every cluster with a KS test at significance 0.1;
8. stop if any two cluster macrohistories do not significantly differ using the KS test and set $k = k - 1$. Otherwise, set $k = k + 1$ and iterate steps 5-8.

2.4 The Physical Review dataset

We consider the full Physical Review dataset—from now on APS dataset, to distinguish it from one of its journals, the Physical Review (PR)—from 1893

to 2013 for the purpose of demonstrating and testing the proposed approach.[1] The APS dataset contains articles from several journals which stemmed from the original Physical Review, and is currently an important venue for publications in all domains of physics. The history of the Physical Review is in a sense that of physics, as a growing theoretical maturity has being paralleled by an increased structuring of scientific discourse and practices, as this over 100-year old dataset highlights [Bazerman, 1988, Chapter 6]. We specifically considered the following article typologies: normal articles, letters, rapid and brief communications, excluding non-standard (from a citations point of view) article typologies such as editorials, comments and errata. A total number of 510137 publications is considered, divided in 31769 rapid, 34572 brief and 8247 letter communications, and 435549 articles.

The following journals are part of the APS dataset:

- **PR** (Physical Review, 1893-1969, articles: 47940): All of physics.
- **RMP** (Reviews of Modern Physics, 1929-today, articles: 3139): All of physics.
- **PRL** (Physical Review Letters, 1958-today, articles: 110080): All of physics.
- **PRA** (Physical Review A, 1970-today, articles: 65170): Atomic, molecular, and optical physics and quantum information.
- **PRB** (Physical Review B, 1970-today, articles: 161257): Condensed matter and materials physics.
- **PRC** (Physical Review C, 1970-today, articles: 34443): Nuclear physics.
- **PRD** (Physical Review D, 1970-today, articles: 69481): Particles, fields, gravitation and cosmology.
- **PRE** (Physical Review E, 1993-today, articles: 46009): Statistical, nonlinear and soft matter physics.

Three caveats of the APS dataset were identified by Redner [2004]: first, the dataset is fully self-contained, with citations from and to articles within the APS dataset itself, meaning that only an estimated 20% of APS articles total citations are accounted for. Secondly, approximately 5 to 10% of citations are incorrect. In order to mitigate the impact, we discarded all citations which were patently erroneous—such as made by articles published before the cited article. Thirdly, as it is well known, raw citation counts vary widely according to time and academic field. The same author stressed how skewed the APS citation distribution is, given a very few articles are cited more than 10 times during their lifespan (less than 20% over the largest analyzed dataset) [Redner, 1998]. For these lucky few, citation lifespans are nevertheless long, suggesting

[1] The dataset was kindly provided by the American Physical Society (APS), and is available for request here: http://journals.aps.org/datasets.

how different citation regimes apply once a publication becomes popular. This consideration supports at the same time a focus on above-average cited articles, and the selection of fixed time windows for the analysis of citation histories, in order to set a boundary to the long tail of the accumulation procedure and compare histories of equal length.

3 Results

In this section we apply spectral clustering to the APS dataset, discuss the resulting archetypical macrohistories and investigate their determinants through multinomial logistic regression.

3.1 Clustering citation histories

We analyzed several sets of data using histories of different time windows, and settled for a comparison over citation histories of four different spans: 6, 12, 24 and 48 years, with a filter on the minimum number of received citations during the period as described in Section 2.3. The normalized macrohistory of the 12 years dataset, in Figure 2, shows the typical rapid peaking and slow decay of highly cited scientific literature. A decrease in the ageing time of articles is also apparent if we consider (not normalized) macrohistories from different publication years in Figure 1. These time windows have been chosen because 6 years is just more than what impact factors normally consider (2-5 years), and is often suggested to be the maximum average delay for citation peaks in most fields of science [Amin and Mabe, 2003], albeit some disciplines, such as social sciences, might take even longer to peak [Glänzel and Schoepflin, 1995]. Moreover, 12, 24 and 48 years are multiples of 6, which should allow us to investigate the long-term behaviors of citation histories and their determinants.

Remarkably, the number of clusters obtained with the KS test was always low and significant, in the sense that it provided with the maximum number of qualitatively different curves/clusters, in terms of the position and magnitude of the peak, and speed and shape of the descending curve. Results are summarized in Table 1. For the largest datasets (windows of 6 and 12 years), we provide in what follows results from samples for computational reasons. Multiple samplings have been taken and tested from the same dataset in order to assure the coherence of our results.

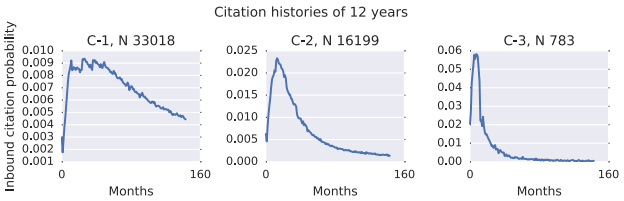
We find a low number of significantly different clusters, which are essentially variations over a continuous space between two extremes: citation histories of *sprinters* and *marathoners*, with a relevant number of articles close to the average, "normal" citation history. We define three types of citation histories as follows: (i) *Marathoners* which present fast or slow-rise, moderately peaked histories, followed by a slow decline, or absence of decline, or even a constant rise in received citations over time. (ii) *Sprinters* instead have an early and high peak and a fast decline. (iii) *Middle-of-the-roads* articles are and in-between average. See for example the results of clustering from 50000 randomly selected articles of

Figure 2: The macrohistory for the 12 years dataset, with a yearly and monthly granularity.

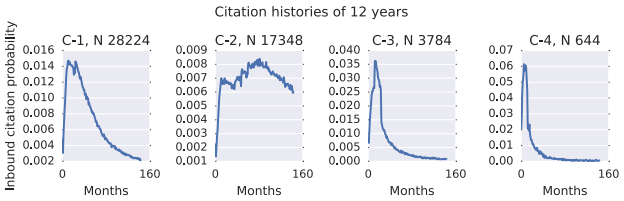
the 12 years dataset, in Figure 3a: marathoners are on average more represented in cluster 1, average articles in cluster 2, and sprinters in cluster 3. Increasing the number of clusters does not give qualitatively different curves, as shown in Figure 3b, where the two last clusters are very small in the number of datapoints, and similar to each other. Different time-spans yield similar results as well: macrohistories can slightly change in peak and shape, and the proportion of articles for any given typology as well, but the overall interpretation does not. This can be seen in Figure 4a for the time windows of 48 years, were marathoners are proportionally fewer in numbers but more markedly so, and the same is true if we increase to 4 clusters as in Figure 4b. We will therefore mostly use a time

Window	**Valid Data**	**Filtered data**	**KS clusters**
6 years	399617	117243*	2
12 years	302810	85861*	3
24 years	158819	42976	4
48 years	40044	10089	3

Table 1: Summary of datasets used for different time windows, and results of clustering. The valid data represents all articles with a sufficiently long history, filtered data is the amount of articles which received an above average number of citations compared to articles published in the same year. KS clusters indicates the number of clusters found using the KS test. * indicates time windows for which the provided results come from samples of data.

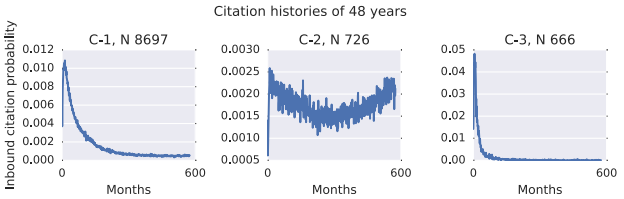


(a) 3 clusters: marathoners (C-1), middle-of-the-roads (C-2), and sprinters (C-3).

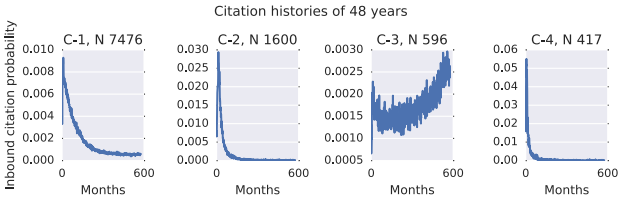


(b) 4 clusters: middle-of-the-roads (C-1), marathoners (C-2), and two clusters of sprinters (C-3 and C-4).

Figure 3: Clustering results for the dataset of length 12 years over a 50000 sample. The number of articles in each cluster is indicated in the title of each figure, which are given in decreasing order.



(a) 3 clusters: middle-of-the-roads (C-1), marathoners (C-2) and sprinters (C-3).



(b) 4 clusters: middle-of-the-roads (C-1), marathoners (C-3), and two clusters of sprinters (C-2 and C-4).

Figure 4: Clustering results for the dataset of length 48 years. The number of articles in each cluster is indicated in the title of each figure, which are given in decreasing order.

period of 12 years to exemplify our findings.

Marathoners represent the most rich typology of citation histories, as they tend to amalgamate quite different curves, as can be seen in Figure 5, where we kept dividing cluster 1 in Figure 3a into three further clusters. A considerable proportion of marathoners are close to their cluster average curve (cluster 1-1), but two sub-clusters are in fact made by sprinting marathoners (publications with a slow rise, relatively fast decline and a plateau on a positive asymptote, cluster 1-2), and extreme marathoners (publications with little sign to stop rising in terms of received citations, cluster 1-3). Notably, the same effect is not at all present if we keep splitting clusters 2 and 3 in Figure 3a, as we find very similar clusters. The reason for this more complex nature of marathoners might be the fact that several known relevant yet rare typologies of citation histories are contained within this category: sleeping beauties [Van Raan, 2004], all-elements-sleeping-beauties [Li, 2014], persistently relevant publications (stable positive asymptote in the number of received citations over the long-term), increasingly relevant publications (monotonic increase in the number of received citations), to name but a few.

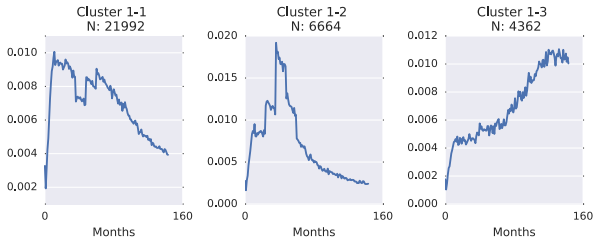


Figure 5: Three clusters obtained from cluster 1 on the 12 years dataset. Marathoners and sprinters appear again, within the marathoners cluster. The number of articles in each cluster is indicated in the title of each figure.

To summarize, we find the following typologies of citation curves in the APS dataset:

1. *Marathoners*: publications which start fast or slow, reach a moderate peak and keep improving the ratio of received citations, or at least keep being relevant over prolonged amounts of time by manifesting a slow decline or a plateau. Marathoners in effect tend to age slowly, or not at all, and are also more numerous and varied than sprinters.

2. *Sprinters*: publications with fast, even extremely fast and high peak, and equally rapid ageing. These publications are immediately relevant for their community, and rapidly forgotten thereafter, and are fewer in number in the APS dataset.

3. *Middle-of-the-roads*: publications with a citation history close to the global

average citation history, that is, a fast but moderately peaking curve with a gradual decay over time.

Citation histories can also be used to investigate the venue of articles, and verify where different journals stand with respect to the sprinters and marathoners balance. Every APS journal has a characteristic citation history, shown in Figure 6 for a time period of 12 years. Relative differences in the ageing patterns of journals from the same domain are a known phenomenon [Moed et al., 1998]. Interestingly, the original Physical Review seems to be the top sprinter: to what extent this effect is partially due its older and thus less reliable data is unknown. The rest of the journals seem to be marathoners instead, with a more or less rapid peak but slow decline. At the opposite sides of the spectrum we find the PRL (slightly more of a sprinter, likely due to the format of letters, which are meant to be quickly digested by the community) and the RMP (more of a marathoner as a review journal, publishing fewer articles which are relevant for a longer time). To be sure, the average filtered APS article has a clear long-tail with slow ageing.

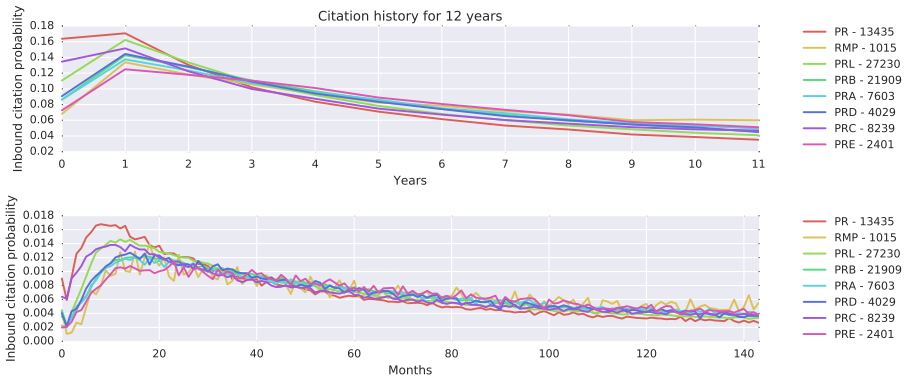
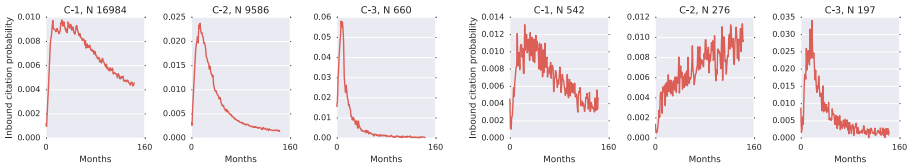


Figure 6: The macrohistories of the journals in the APS dataset, for a time period of 12 years. The number of articles in each journal is indicated in the legend.

Our clustering procedure can be applied to articles from specific journals as well. If we cluster the articles of the PRL and RML, as shown in Figure 7, we find different macrohistories. The PRL, also due to the amount of articles from this journal in the dataset, has all macrohistories which are similar to the global ones in Figure 3a. Conversely the RMP, which has fewer articles in the dataset, emerges as a journal distinct for its focus on reviews. As a consequence, sprinters are almost absent and we can find a monotonically increasing cluster (number 2 in Figure 7b), representing articles that continue to be increasingly more relevant over several years after publications. Nevertheless even for the RMP, the first cluster macrohistory is similar to the first global cluster in Figure



(a) Physical Review Letters: marathoners and sprinters reappear.

(b) Review of Modern Physics: cluster 2 is made of monotonically increasing articles.

Figure 7: Comparing journals in the APS dataset: 3 clusters for histories of 12 years.

3a.

3.2 Determinants of citation histories

In order to further investigate some determinants of different citation histories, we run a series of multinomial logistic regressions with the results of clustering as dependent variable, and a set of metadata on articles as independent, as summarized in Table 2. Multinomial logistic regression works on nominal (not ordered) categories, taking one as reference category, and evaluating the relative risk of being in one of the remaining categories against the reference using a linear combination of predictors. The resulting MLE-estimated coefficients represent the effects of every predictor variable in the log-odds of being in any other cluster versus the reference cluster. In other words, given any coefficient β, the probability of being in a certain cluster against the reference cluster increases $exp(\beta)$ times for each unit increase in the corresponding independent variable, all else kept constant. For example, given a hypothetical coefficient β of -0.1823 on the number of authors for, say, cluster 2 against cluster 1, we have that a unit increase in the number of authors increases the probability of being in cluster 2 relative to cluster 1 by a multiplicative factor of $exp(-0.1823) = 0.833$, thus actually decreases it by -17.7%. The effect of a single feature unit increase on the probability of a datapoint being in a given cluster, all else being equal, is instead calculated via marginal effects, that is the derivative of the probability of being in a given category with respect to a predictor. We provide both regression and marginal effects coefficients in what follows. See e.g. Greene [2012] for a thorough introduction to multinomial logistic regression.

We estimated two models for the 12 years and one for the 48 years datasets: (i) a baseline for both windows using filtered articles, whose results are reported in the Appendix; (ii) a model only for articles published from 1970 with a window of 12 years, thus excluding the original Physical Review and the oldest articles from PL and RMP. Both the full dataset and the post-1970 models were also verified for history lengths of 6 and 24 years, providing coherent results. The reference cluster is always cluster 1 (the first one in Figure 3a for the 12-year window and the first one in Figure 4a for the 48-year window. Both are

y (dependent)	cluster number
cits	total number of received citations during time window
citsNSL	total number of received citations from journals different from publication venue during time window
refs	total number of cited references
refsAge3P	3rd percentile of the distribution of the age of cited references
refsAgeMean	mean of the distribution of the age of cited references
numAuthors	number of authors
numPages	number of pages
pubYear	publication year
isArticle	if article (as opposed to rapid, brief or letter communication)

Table 2: Independent variables used for regression on cluster number. All references, made and received, are to and from publications within the APS journals.

the most representative in the number of articles).

Regressions consistently point to specific features for clusters with different degrees of marathoners/sprinters. Some features are stable to all time spans, others change role predicting sprinters over short windows and marathoners over long windows. It is useful to consider marginal effects and regression coefficients together for an absolute and relative comparison. More specifically different degrees of the two profiles compare in the following way:

- *Marathoners* are cited less on average over short windows (up to 24 years included), but more over long windows (48 years). They are cited more from other journals than the original venue over short windows but less over long windows. They have fewer authors, are longer in number of pages, and have older and (slightly) shorter bibliographies, despite review articles being commonly marathoners. They are also younger on average, and more likely to be articles than a rapid, brief or letter communication.

- *Sprinters* are cited more on average over short windows, but less over long windows. They are cited less from other journals than the original venue over short windows. They have more authors, are shorter, with (slightly) longer and younger bibliographies. Finally, they are on average older and more likely to be a rapid, brief or letter communication.

Several effects are weak even if significant. For example, the list of references of sprinters and marathoners is found to be correlated in the sense of a longer but younger list for sprinters. This effect is somewhat counter intuitive, especially as review articles are typically marathoners (see the RMP macrohistory in Figure 6). To be sure, the length of bibliographies is correlated by a very weak coefficient, while the age of cited articles is stronger as a signal, and intuitively points to older reference lists for marathoners than sprinters.

From our results we can identify some clear determinants of a citation history, broadly divided into extrinsic and intrinsic. Extrinsic determinants depend on the context, timing and ensuing history of a publication, and can be diachronic; intrinsic determinants are instead synchronic, related to the materiality and content of a publication. The main extrinsic determinant is the year of publication, which very clearly points to a slowing down in the ageing process of articles over recent times, as already discussed in the literature (e.g. [Bouabid and Larivière, 2013]). Received citations are also an extrinsic variable to some extent, one that change in importance with different time windows. The shorter the time window, the more significant is the impact of having more citations for the likelihood of being a sprinter: indeed sprinters accumulate most of their citations over short amounts of time after publication. This effect gets diluted as the window is enlarged, with the number of citations becoming a (weak) predictor of being a marathoner for the 48-year window. Interestingly, the opposite trend applies for citations coming from other venues than the publication journal. This might signify that sprinters are short and rapidly digested publications which less likely surpass the boundaries of the original community of interest.

Some intrinsic effects are also relevant, having to do with the contents and structure of the article. The most important determinant is the type of publication—a full article being more likely a marathoner. An important signal also comes from the age of the reference list, as discussed before. Weak but significant coefficients also come from the number of pages and authors of the article. In general a cautionary word must be spent to highlight how the estimated explanatory power of the model is rather low (pseudo squared R of circa 0.15 in for all models).

4 Related literature

The first result showing two groups of citation histories with a similar interpretation to ours was Aversa [1985], where sprinters where identified as early rise and rapid decline publications—also mentioned as "flashes in the pans" in the literature (e.g. [Van Dalen and Henkens, 2005]); and marathoners as delayed rise and slow decline ones. An important point is that our marathoners can in fact early rise too, the only difference is in the decaying behavior (or lack of). A further typology has been used by several authors, that of "normal science" (e.g. [Van Dalen and Henkens, 2005]). Indeed a lot of articles close to the global average citation curve exist, as shown in the second cluster in Figure 3a for the 12 years period: the average article is not an artifact of aggregation. We avoid explicitly considering "sleeping beauties" (cf. e.g. [Van Raan, 2004]), that is articles which go unnoticed for a long time before attracting a lot of attention, due to the relative rarity of the phenomenon in the APS dataset [Redner, 2004].

In order to group articles according to their citation histories, most of previous literature relied on heuristics. As an example, Costas et al. [2010] used the rules originally proposed by Price in a private communication to Aversa [1985]:

a publication is considered a flash in the pans if it received at least 50% of its citations before the 75% of the papers in its domain still did not; a delayed-type if it received 50% after 75% of papers in its domain already received at least 50%, and normal if in between. Exceptions are Aversa [1985], who experimented with k-means on a limited set of articles; Baumgartner and Leydesdorff [2014], who used group-based trajectory modeling to distinguish different developmental trajectories for citation histories; Sun et al. [2016] who proposed a two-parameter method looking at the shape and drastic fluctuations in the citation curve in order to distinguish different patterns of obsolescence.

Considering the determinants of citation histories, Van Dalen and Henkens [2005] investigated some predictors of citation curve typology for a set of articles in demography using multinomial logit. The reference category were 1) not cited or barely cited papers, vs three other categories: 2) sleeping beauties, 3) flash in the pans and 4) so called normal science. The venue, as represented by the journal impact factor, was found to be strongly positively correlated with typologies 2, 3 and 4, the author's reputation and the length of the paper (in number of pages) were found to be only weakly correlated. We also elaborate on the results of Costas et al. [2010] in the comparison of flash in the pans and delayed documents. Delayed papers were found by them to be more cited and having higher field-specific impact, in agreement with previous literature [Aversa, 1985, Aksnes, 2003, Levitt and Thelwall, 2009]. We show how marathoners in the APS dataset are more cited in absolute only at a relatively late stage of their history, being less cited than sprinters at the beginning. Marathoners are nevertheless more likely to be cited by articles from different journals from their early stages, which from the 1970 represents a coarse field indicator with the publication of Physical Reviews A to E. They therefore seem to be more relevant to a domain than just a community, on average. Delayed papers were also found by Costas et al. [2010] to be less collaborative on average, with several possible explanations such as fewer discovery options compared to flash in the pans given by having fewer authors and outbound citations. Flash in the pans represent immediacy, are shorter and typically not archival papers, tending to be published in higher impact journals. This immediacy might be also due to self-citations and self-promotion effects. The authors advance the hypothesis that a research is in fact a series of papers linked among each-other, with flash in the pans, normal and delayed types all playing a role in the diffusion of a research effort.

We further complement these findings by evidencing other effects as well. First of all, the space between full marathoners (delayed recognition) and sprinters (flash in the pans) is in a sense continuous, as average articles tend to be in-between in terms of regression coefficients. The same regression coefficients also increase or decrease according to the more or less pronounced history of the cluster in terms of degree of marathoner/sprinter. Secondly, results seems to be consistent over time, as they do not differ substantially for papers published from 1970 onwards. To be sure, the exponential growth of PR journals and therefore of produced citations, shown in Redner [2004], already makes the signal of older articles less significant. Our results seem also to confirm what

suggested by Bouabid and Larivière [2013], anticipated in Larivière et al. [2009] and further expanded in Parolo et al. [2015], on the increasing average age of articles published in recent years. They find an overall steady increase in the expected life of papers, as measured by the residual citations after 12 years from publication. Sub-fields and journals differ considerably in this respect. The Physical Review (1980-2000), specifically, is found to be strongly aged due to it publishing main advances in physics which remain of interest for an indefinite time. It goes from an over 30% of residual citations in 1980 to an over 60% of residuals in 2000. Marathoners in the APS dataset are on average published more recently. This was already evident from journals citation histories, as for example the original Physical Review is a clear sprinter. This result might also be partially determined by disciplinary differences, where for example theoretical physics articles are on average marathoners, and citing marathoners in turn (i.e. other theoretical physics articles). Disciplinary differences could entail different referencing practices. This is in agreement with Glänzel and Schoepflin [1995], who proposed that differences in the speed of ageing and reception are more due to field than venue (journal) specific effects, and for example mathematics and other theoretical-oriented disciplines have a slower reception speed. Others nevertheless found a link between the role of a journal within a community, namely supporting the research front (fast ageing) or being an archival or reference journal (slow ageing) [Griffith et al., 1979]. Sinatra et al. [2015] also linked differences in the speed of ageing of articles, as mapped by length of impact, to the more or less insular nature of the respective sub-field. In our case journals could proxy fields to some extent (Physics Review A to E).

We nevertheless elucidate another possible motivation for the speed of ageing, which could be the typology of articles. Previous research has found article typology to be correlated to the amount of received citations, but not to different ageing patterns [Baumgartner and Leydesdorff, 2014]. We instead show how a standard research article is more likely to be a marathoner than a rapid, brief or letter communication, which is indeed meant to have a rapid diffusion and equally rapidly be superseded by further results. This effect applies from the early stages of the history of APS articles, and grows in time.

5 Conclusions and future work

We proposed a methodology that uses spectral clustering to group citation histories of publications into communities and then applies multinomial logistic regression to provide the detected communities with semantics in terms of publication attributes. We applied the methodology to the full Physical Review archive.

Not surprisingly, the typical publication in physics has a citation history with a fast but moderately peaking curve and a gradual decay over time. Nevertheless, we found that two opposite archetypes of publication neatly deviate from this pattern: marathoners, that are publications that after the initial moderate success keep improving the share of received endorsements, or at least manifest

a plateau or a slow decline, and sprinters, that are publications with fast and impetuous initial success and equally rapid decrease. Notably, these behaviors are typical in information diffusion processes, where an information item can be totally unnoticed, or intensely considered for a short amount of time and then suddenly neglected, or remain relevant for a longer amount of time, even indefinitely. Marathoners, compared to sprinters, are determined by receiving less early-citations (particularly from the same journal), but increasing late-citations over time. They receive more citations from journals different than the publication venue, signaling a comparatively higher relevance beyond the community of interest. Marathoners also tend to have older bibliographies, fewer authors, and be longer and younger articles. The outlined methodology can be directly applied to journals, scholars or other bibliometric units in order to verify where different bibliometric units stand with respect to the sprinter and marathoner metaphors.

We plan to apply a the methodology to more disciplines and domains, in particular comparing the differences between humanities and sciences. Another direction of future work entails considering the macrohistory of publications cited by a given publication, and the relation to the macrohistory of the cluster of the citing publication.

Acknowledgements

Giovanni Colavizza is supported by the Swiss National Fund under Division II, project number 205121_159961. The authors would also like to thank Xavier Bresson (EPFL) for helpful discussions.

References

Dag W. Aksnes. Characteristics of highly cited papers. *Research Evaluation*, 12(3):159–170, 2003.

Mayur Amin and Michael Mabe. Impact factors: use and abuse. *MEDICINA (Buenos Aires)*, 63:347–354, 2003.

Elizabeth Aversa. Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics*, 7(3-6): 383–389, 1985.

Susanne E. Baumgartner and Loet Leydesdorff. Group-based trajectory modeling of citations in scholarly literature: dynamic qualities of transient and sticky knowledge claims. *Journal of the Association for Information Science and Technology*, 65(4):797–811, 2014.

Charles Bazerman. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. University of Wisconsin Press, Madison, Wisconsin, 1988.

Claudio Bettini, Sean X. Wang, and Sushil Jajodia. *Encyclopedia of Database Systems*, chapter Temporal Granularity, pages 2968–2973. Springer US, 2009.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Hamid Bouabid and Vincent Larivière. The lengthening of papers life expectancy: a diachronous analysis. *Scientometrics*, 97(3):695–717, 2013.

Rodrigo Costas, Thed N. van Leeuwen, and Anthony F. J. van Raan. 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, 2010.

Derek De Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

Alex P. Dow, Lada A. Adamic, and Adrien Friggeri. The anatomy of large Facebook cascades. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 145–154, 2013.

Eugene Garfield. Premature Discovery or Delayed Recognition — Why? *Essays of an Information Scientist*, 4:488–493, 1980.

Wolfgang Glänzel and Urs Schoepflin. A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, 21 (1):37–53, 1995.

William H. Greene. *Econometric Analysis*. Pearson, 2012.

Belver C. Griffith, Patricia N. Servi, Anita L. Anker, and Carl Drott. The Aging of Scientific Literature: a Citation Analysis. *Journal of Documentation*, 35 (3):179–196, 1979.

Vincent Larivière, Yves Gingras, and Éric Archambault. The decline in the concentration of citations, 1900-2007. *Journal of the American Society for Information Science and Technology*, 60(4):858–862, 2009.

Jonathan M. Levitt and Mike Thelwall. The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, 78(1):45–67, 2009.

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

Zhenquan Lin, Shanci Hou, and Jinshan Wu. The correlation between editorial delay and the ratio of highly cited papers in Nature, Science and Physical Review Letters. *Scientometrics*, 107(3):1457–1464, 2016.

Maurice B. Line and A. Sandison. Progress in Documentation: obsolescence and changes in the use of literature with time. *Journal of Documentation*, 30 (3):283–350, 1974.

Henk F. Moed, Theodorus N. Van Leeuwen, and Jan Reedijk. A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, 54(4):387–419, 1998.

Hideshiro Nakamoto. Synchronous and diachronous citation distribution. *Infometrics*, 87/88:157–163, 1988.

Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Santo Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734–745, 2015.

Thijs Pollmann. Forgetting and the Ageing of Scientific Publications. *Scientometrics*, 47(1):43–54, 2000.

Filippo Radicchi and Claudio Castellano. Rescaling citations of publications in physics. *Physical Review E*, 83(4), 2011.

Sidney Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

Sidney Redner. Citation statistics from more than a century of physical review. *Physics Today*, 58:49–54, 2004.

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

Roberta Sinatra, Pierre Deville, Michael Szell, Dashun Wang, and Albert-László Barabási. A century of physics. *Nature Physics*, 11(10):791–796, 2015.

Jianjun Sun, Chao Min, and Jiang Li. A vector for measuring obsolescence of scientific articles. *Scientometrics*, 107:745–757, 2016.

Hendrik P. Van Dalen and Kène Henkens. Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics*, 64(2):209–233, 2005.

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

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 313–319. IEEE, 2003.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, volume 17, pages 1601–1608, 2004.

Appendix

Multinomial logit marginal effects. Results for 3 clusters over the 12 years dataset.

Dep. Variable:	y				
Method:	dydx				
y=1	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
cits	-0.0587	0.001	-51.873	0.000	-0.061 -0.057
citsNSL	0.0186	0.002	9.704	0.000	0.015 0.022
numAuthors	-0.0007	9.31e-05	-7.236	0.000	-0.001 -0.000
numPages	0.0083	0.000	19.057	0.000	0.007 0.009
pubYear	0.0046	0.000	35.138	0.000	0.004 0.005
refs	-0.0017	0.000	-6.848	0.000	-0.002 -0.001
refsAge3P	-0.0011	0.001	-1.206	0.228	-0.003 0.001
refsAgeMean	0.0138	0.001	9.665	0.000	0.011 0.017
isArticle	0.0357	0.007	5.061	0.000	0.022 0.050
y=2	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
cits	0.0544	0.001	48.467	0.000	0.052 0.057
citsNSL	-0.0163	0.002	-8.653	0.000	-0.020 -0.013
numAuthors	0.0007	9.34e-05	7.048	0.000	0.000 0.001
numPages	-0.0068	0.000	-15.310	0.000	-0.008 -0.006
pubYear	-0.0040	0.000	-30.354	0.000	-0.004 -0.004
refs	0.0017	0.000	6.469	0.000	0.001 0.002
refsAge3P	0.0016	0.001	1.655	0.098	-0.000 0.003
refsAgeMean	-0.0132	0.001	-8.982	0.000	-0.016 -0.010
isArticle	-0.0341	0.007	-4.813	0.000	-0.048 -0.020
y=3	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
cits	0.0043	0.000	25.467	0.000	0.004 0.005
citsNSL	-0.0023	0.000	-11.858	0.000	-0.003 -0.002
numAuthors	1.577e-05	2.63e-05	0.600	0.549	-3.58e-05 6.73e-05
numPages	-0.0015	0.000	-9.324	0.000	-0.002 -0.001
pubYear	-0.0006	3.41e-05	-16.552	0.000	-0.001 -0.000
refs	6.805e-05	8.62e-05	0.789	0.430	-0.000 0.000
refsAge3P	-0.0005	0.000	-1.297	0.195	-0.001 0.000
refsAgeMean	-0.0007	0.001	-1.271	0.204	-0.002 0.000
isArticle	-0.0016	0.001	-1.130	0.258	-0.004 0.001

Multinomial logit regression. Results for 3 clusters over the 12 years dataset. The reference cluster is the first one in Figure 3a.

Dep. Variable:	y	**No. Observations:**	47321
Model:	MNLogit	**Df Residuals:**	47301
Method:	MLE	**Df Model:**	18
Date:	Tue, 21 Jun 2016	**Pseudo R-squ.:**	0.1458
Time:	21:51:26	**Log-Likelihood:**	-28402.
converged:	True	**LL-Null:**	-33249.

y=2	**coef**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.3098	0.007	46.416	0.000	0.297	0.323
citsNSL	-0.0964	0.010	-9.335	0.000	-0.117	-0.076
numAuthors	0.0036	0.001	7.223	0.000	0.003	0.005
numPages	-0.0424	0.002	-17.842	0.000	-0.047	-0.038
pubYear	-0.0236	0.001	-32.200	0.000	-0.025	-0.022
refs	0.0093	0.001	6.796	0.000	0.007	0.012
refsAge3P	0.0069	0.005	1.379	0.168	-0.003	0.017
refsAgeMean	-0.0737	0.008	-9.542	0.000	-0.089	-0.059
isArticle	-0.1902	0.038	-4.996	0.000	-0.265	-0.116
const	46.1689	1.451	31.820	0.000	43.325	49.013

y=3	**coef**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.5671	0.012	45.576	0.000	0.543	0.592
citsNSL	-0.2533	0.018	-14.107	0.000	-0.289	-0.218
numAuthors	0.0037	0.002	1.679	0.093	-0.001	0.008
numPages	-0.1533	0.013	-11.706	0.000	-0.179	-0.128
pubYear	-0.0626	0.003	-24.479	0.000	-0.068	-0.058
refs	0.0119	0.007	1.643	0.100	-0.002	0.026
refsAge3P	-0.0335	0.030	-1.135	0.256	-0.091	0.024
refsAgeMean	-0.1035	0.043	-2.415	0.016	-0.187	-0.019
isArticle	-0.2612	0.122	-2.148	0.032	-0.499	-0.023
const	119.7593	4.990	23.999	0.000	109.979	129.540

Multinomial logit marginal effects. Results for 3 clusters over the 48 years dataset.

	Dep. Variable:			y		
	Method:			dydx		
y=1	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	-0.0059	0.000	-14.067	0.000	-0.007	-0.005
citsNSL	0.0019	0.001	2.802	0.005	0.001	0.003
numAuthors	0.0191	0.003	5.634	0.000	0.012	0.026
numPages	0.0031	0.001	3.483	0.000	0.001	0.005
pubYear	-0.0011	0.000	-2.277	0.023	-0.002	-0.000
refs	0.0015	0.000	3.616	0.000	0.001	0.002
refsAge3P	-0.0003	0.003	-0.097	0.923	-0.006	0.006
refsAgeMean	0.0061	0.004	1.446	0.148	-0.002	0.014
isArticle	-0.0370	0.019	-1.934	0.053	-0.075	0.000
y=2	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.0028	0.000	10.180	0.000	0.002	0.003
citsNSL	-0.0020	0.000	-4.240	0.000	-0.003	-0.001
numAuthors	-0.0285	0.003	-8.722	0.000	-0.035	-0.022
numPages	0.0035	0.000	10.400	0.000	0.003	0.004
pubYear	0.0044	0.000	11.521	0.000	0.004	0.005
refs	-0.0019	0.000	-5.375	0.000	-0.003	-0.001
refsAge3P	-0.0025	0.001	-1.660	0.097	-0.005	0.000
refsAgeMean	0.0073	0.002	3.620	0.000	0.003	0.011
isArticle	0.0764	0.019	4.119	0.000	0.040	0.113
y=3	**dy/dx**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.0032	0.000	11.783	0.000	0.003	0.004
citsNSL	0.0001	0.000	0.257	0.798	-0.001	0.001
numAuthors	0.0093	0.001	7.031	0.000	0.007	0.012
numPages	-0.0066	0.001	-7.736	0.000	-0.008	-0.005
pubYear	-0.0033	0.000	-11.467	0.000	-0.004	-0.003
refs	0.0004	0.000	1.699	0.089	-6.25e-05	0.001
refsAge3P	0.0028	0.003	1.023	0.306	-0.003	0.008
refsAgeMean	-0.0134	0.004	-3.436	0.001	-0.021	-0.006
isArticle	-0.0394	0.007	-5.636	0.000	-0.053	-0.026

Multinomial logit regression. Results for 3 clusters over the 48 years dataset. The reference cluster is the first one in Figure 4a.

Dep. Variable:	y	**No. Observations:**	8483
Model:	MNLogit	**Df Residuals:**	8463
Method:	MLE	**Df Model:**	18
Date:	Tue, 21 Jun 2016	**Pseudo R-squ.:**	0.1495
Time:	22:08:30	**Log-Likelihood:**	-3645.7
converged:	True	**LL-Null:**	-4286.6

y=2	**coef**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.0476	0.004	10.651	0.000	0.039	0.056
citsNSL	-0.0331	0.008	-4.208	0.000	-0.048	-0.018
numAuthors	-0.4552	0.052	-8.747	0.000	-0.557	-0.353
numPages	0.0514	0.005	9.456	0.000	0.041	0.062
pubYear	0.0688	0.006	11.397	0.000	0.057	0.081
refs	-0.0304	0.006	-5.344	0.000	-0.042	-0.019
refsAge3P	-0.0380	0.024	-1.565	0.117	-0.086	0.010
refsAgeMean	0.1069	0.033	3.279	0.001	0.043	0.171
isArticle	1.2100	0.300	4.037	0.000	0.623	1.797
const	-138.0097	11.755	-11.741	0.000	-161.049	-114.971

y=3	**coef**	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**	
cits	0.0595	0.005	12.244	0.000	0.050	0.069
citsNSL	0.0002	0.008	0.026	0.979	-0.015	0.016
numAuthors	0.1424	0.024	5.971	0.000	0.096	0.189
numPages	-0.1145	0.015	-7.655	0.000	-0.144	-0.085
pubYear	-0.0558	0.005	-10.910	0.000	-0.066	-0.046
refs	0.0056	0.004	1.316	0.188	-0.003	0.014
refsAge3P	0.0476	0.049	0.980	0.327	-0.048	0.143
refsAgeMean	-0.2336	0.069	-3.362	0.001	-0.370	-0.097
isArticle	-0.6391	0.125	-5.124	0.000	-0.884	-0.395
const	106.8124	9.928	10.759	0.000	87.354	126.271