Technological Forecasting & Social Change 112 (2016) 121–133

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Research infrastructures in the LHC era: A scientometric approach☆

Stefano Carrazza [a,*], Alfio Ferrara [b], Silvia Salini [c]

[a] Theoretical Physics Department, CERN, Geneva, Switzerland
[b] Department of Computer Science, Università degli Studi di Milano, Italy
[c] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 25 January 2016
Accepted 5 February 2016
Available online 26 February 2016

Keywords:
Research infrastructures
Scientometrics
Clustering
Citation distribution
Publication distribution
Knowledge propagation

ABSTRACT

When a research infrastructure is funded and implemented, new information and new publications are created. This new information is the measurable output of discovery process. In this paper, we describe the impact of infrastructure for physics experiments in terms of publications and citations. In particular, we consider the Large Hadron Collider (LHC) experiments (ATLAS, CMS, ALICE, LHCb) and compare them to the Large Electron Positron Collider (LEP) experiments (ALEPH, DELPHI, L3, OPAL) and the Tevatron experiments (CDF, D0). We provide an overview of the scientific output of these projects over time and highlight the role played by remarkable project results in the publication–citation distribution trends. The methodological and technical contributions of this work provide a starting point for the development of a theoretical model of modern scientific knowledge propagation over time.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

The main purpose of this study is to investigate whether there is a pattern of propagation of knowledge related to research infrastructures and, if it exists, what it depends on and how to measure it. The time and manner of dissemination of knowledge are hard to measure and to predict. The processes of dissemination are diverse and often not observable, but the number of publications associated to a project and the citations that it receives are the most immediate information that we are able to measure. Scientometric techniques (de Solla Price, 1986) are the most used approaches to evaluate knowledge propagation. These methods are based on the analysis of scientific publications and their citations over time. The creation of knowledge is certainly one of the benefits that can justify the high costs for the construction of research infrastructures. We are also motivated by the idea of providing a first understanding of knowledge growth derived from the funding of research infrastructures (Martin & Irvine, 1984; Martin, 1996; Florio et al.,).

In particular, in this paper, we focus our study on the most modern accelerator project in High Energy Physics, the Large Hadron Collider (LHC), completed at the European Organization for Nuclear Research (CERN) in 2008. The LHC's primary function is to search for the Higgs boson and, more generally, for new physics discoveries involving high collision energies. The LHC accelerator is utilized in seven experiments that use detectors to analyze the particles produced by the collisions. In this work, we will focus on the four biggest experimental collaborations: ATLAS, CMS, ALICE and LHCb. ATLAS and CMS are two general purpose experiments composed by a large number of collaborators worldwide, they are specialized in the search for signs of new physics and the hunt for the Higgs boson. ALICE and LHCb are specific experiments looking at heavy-ion collisions and antimatter respectively, their community is smaller than the general purpose experiments.

The data from LHC are complemented with data collected from the Large Electron-Positron Collider (LEP) and the Tevatron experiments, in order to compare results at different times and using different technologies and infrastructures. Our work is focused on a period starting with the first publication of Tevatron, that is, 1982 to 2012. We describe the knowledge output of the projects considered here by considering the following variables that bring out interesting regularities and make data from different projects comparable:

- the different evolution of the reference scientific community as reflected by different rates of publications and interrelations among scientists and infrastructures;
- the lifetime cycle of each specific project and its community; and
- the eventual remarkable project results that can enhance or modify the distribution of citations.

To this end, we describe the *activity* (number of publications) and the *impact* (number of citations) of scientific output by comparing the

☆ Preprints: CERN-PH-TH-2015-246, TIF-UNIMI-2015-17.
* Corresponding author.
E-mail addresses: stefano.carrazza@cern.ch (S. Carrazza), alfio.ferrara@unimi.it (A. Ferrara), silvia.salini@unimi.it (S. Salini).

http://dx.doi.org/10.1016/j.techfore.2016.02.005
0040-1625/© 2016 Elsevier Inc. All rights reserved.

results with the rate of overall publications in physics, as reported by Web Of Science.[1]

Moreover, we note that not all papers are equal in terms of citation trajectory; for each experiment there are papers with different weights. The weight classifies the behavior from excellent to mediocre papers in terms of propagation impact.

As a first step, we group the papers according to the the shape of their distribution of citations over time. We also study if the citation patterns depend on the semantic dimension and on the temporal dimension.

The cluster of papers could depend on some covariates, such as the characteristics of the scientific community that produced them, the number of authors involved, the reputation of them, etc.

Beyond this first description of the knowledge growth due to the analyzed projects, the data collected and the methodological and technological tools used in this paper will be the starting point for the definition of a statistical model predicting the outcome of a project, given the human and financial resources available and its timing.

Section 2 describes the data used in this work. Section 3 shows the activity and impact measures. Section 4 motivates the modeling of knowledge propagation in High Energy Physics (HEP). Section 5 introduces a methodology of clustering of papers based on citation patterns. Section 6 studies the cluster collections according to the semantic and temporal dimensions. Finally we list our conclusions and future tasks in Section 7.

2. Data description

In practice, tracking knowledge creation consists of quantifying the knowledge outputs generated by scientists' experiments (first wave knowledge), by papers written by other scientists and citing those of the first wave, by other papers citing those of the second wave and so on. In the following, we define knowledge as outputs generated by *insider* scientist papers as *level 0* papers and knowledge outputs generated by *outsiders*-scientist-literature papers as *level 1* papers. Papers by scientists outside *level 1* are called *level 2*, and so on.

Fig. 1 shows a synthetic view of the projects and relative experiments taken into account by the present analysis. The LHC was constructed after the LEP project at CERN, and operated from 1989 until 2000. The LEP project comprised four experiments: ALEPH, DELPHI, L3 and OPAL. We also include all the available information from these LEP experiments in order to compare the research output from projects organized in the same laboratory but at different time periods.

Another potential comparison involves projects from multiple infrastructures. In order to perform such a comparison, we also include the Tevatron project at the Fermi National Accelerator Laboratory (Fermilab) in the USA, which started operating in 1983 and ceased operations in 2011. The Tevatron is a synchrotron accelerator used in two experiments, CDF and D0.

The LHC, LEP and Tevatron are projects involving the same physics field, which is High Energy Physics, but the time periods of operation do not allow a comparison of the absolute values for the paper and citations produced. It should be noted that in the 1990s, when pre-prints and open access were not yet available, it was difficult to get a paper in electronic format on a home computer. In 1991, the Internet was born and the database SPIRES High Energy Physics (SPIRES-HEP), installed at the Stanford Linear Accelerator Center (SLAC) in the 1970s, became the first website in North America and the first database accessible via the World Wide Web.

The bibliographic database used in the current analysis was extracted directly from the INSPIRE website (http://inspirehep.net/) by querying the public user interface. The database was constructed during September 2013, and we include papers up to 2012 in order to avoid the inclusion of unconsolidated papers. The collection of papers obtained by this procedure contains the information needed to reconstruct the citation evolution of the most important papers in HEP. However, we are aware that several papers not published in INSPIRE were used in the technical development of large research machines, such as the LHC, and also that technical patents provide benefits which are important to the scientific community.

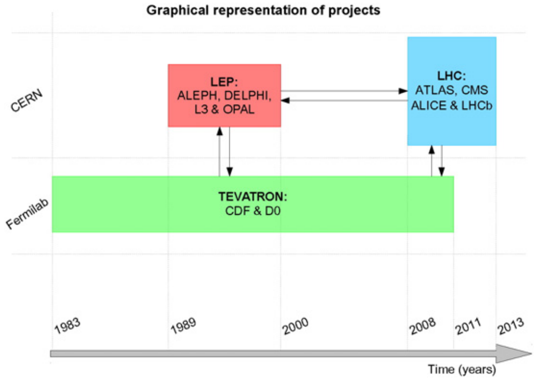


Fig. 1. Graphical representation of scientific projects included in the present work by function of time, subdivided by laboratory. The lifetime of each project is represented by the width of the respective rectangle.

Using that collection of papers we perform comparisons and studies about the respective scientific communities, infrastructures and the diffusion of scientific knowledge across time.

Technical tools have been developed in order to create the database. The procedure is summarized in the following steps: *i*) download all available information obtained by querying the name of the experimental collaboration, e.g. "collaboration: 'ATLAS'" with a custom python script able to build a catalog of records using information from papers stored in custom tags; *ii*) extract and download the respective citation and reference records from papers obtained in *i*; and *iii*) import all information to a final MySQL database. A graphical summary of such steps is shown in Fig. 2.

In the next sections, we show results obtained from this database.

3. Activity measures and impact measures

The simplest measure of activity that can be considered is the number of papers produced by authors working on an experiment. We note that the number of produced papers does not match the number of papers actually published. There are a substantial number of pre-prints loaded in arXiv that are not published in scientific journals. These papers are found in bibliometric databases, such as Scopus or Web of Science, and are considered in our analysis. In the following, we will denote experiment papers as *level 0* paper and literature papers as *level 1* papers. We denote experiment paper cited by literature papers as *1to0* and literature papers cited by experiment papers as *0to1*.

Table 1[2] shows the total number of papers for each experiment, separately for published and unpublished and for *levels 0 and 1*.

It is important to note that the number of papers produced from LHC experiments has already exceeded the number of papers produced from both LEP and Tevatron, although these experiments lasted much longer. The same thing occurs with the literature papers, which, as evident when examining LEP and Tevatron experiments, have continued to grow over the years, particularly literature papers that cite experiments.

Next, we examine several impact measures. The simplest measure of impact is the number of citations generated by an experiment. Table 2 shows the citations for each experiment: *0to0* are

[1] http://wokinfo.com/.

[2] All tables refer to data collected up to November 2013.

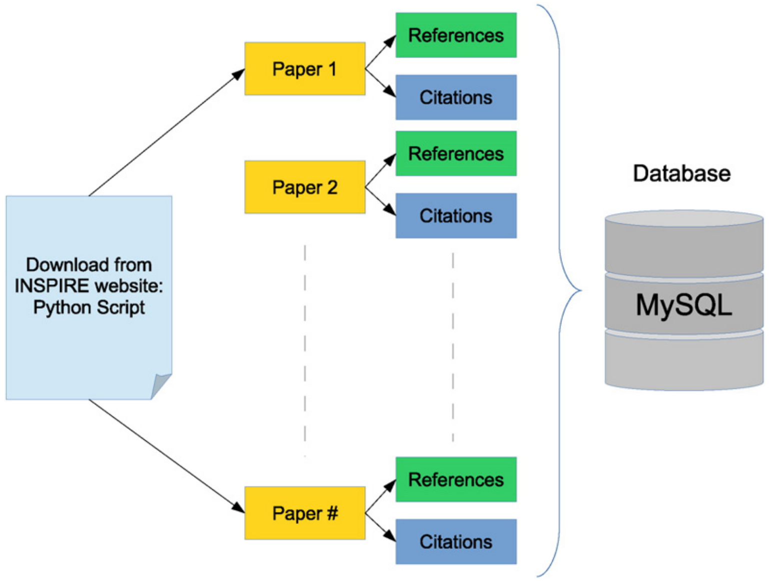


Fig. 2. Graphical representation of the database creation. The records are downloaded from the INSPIRE website by querying the project name. For each paper in the project the reference and citation papers are extracted. Finally all the records are stored in a MySQL database.

citations of experiment papers in experiment papers; *0to1* are citations of experiment papers in literature papers; *1to0* are citations of literature papers in experiment papers; and *1to1* are citations for literature papers versus literature papers that cite experiment papers. The table also shows the experiment papers' H-index and the number of papers with more than 500 citations (renowned papers). The H-index is defined as the number such that, for a general group of papers, *h* papers received at least *h* citations while the other papers received no more than *h* citations (Hirsch, 2005). The H-index measures both the productivity and citation impacts of the publications of a scientist or scholar. The index can also be applied to the productivity and impact of a scholarly journal as well as a group of scientists, such as a department or university or country.

As seen in Table 1, the number of papers in the literature citing the LEP and Tevatron is still higher than the number of papers in the literature mentioning LHC. However, this is not the case for citations. The number of citations (*0to0* and *1to1*) for LHC experiments, ATLAS and CMS in particular, are an order of magnitude higher than those of the LEP experiments. Whether this is due to the fact that the LHC operated during the era of the World Wide Web and the LEP did not or to the fact that the LHC is associated with the discovery of the Higgs boson or both together would be an interesting study to be carried out in the future.

Appendix A details the absolute value of activity and impact measures for each experiment year by year.

The LHC series (Tables A.13, A.14, A.15 and A.16) shows steady growth, with a slight increase in 2008 (when it started operations), and an explosion in 2012. On July 4, 2012, the discovery of the Higgs boson was announced. While important, this is not the only reason for the explosion; in the years 2010–2012, many important results have been obtained via experiments using LHC. In 2011, the number of literature papers citing the experiments increased rapidly, particularly for ATLAS and CMS, superseding both the number of internal papers and the literature papers cited.

Looking to the LEP project (Tables A.7, A.8, A.9 and A.10), it can be observed that the gap between produced papers and published papers is reduced. This is because, as already mentioned, there was no Internet in 1989 when the LEP experiments began. Moreover, when examining the LEP trajectories, it is evident that when the experiment began (1989), the number of literature papers citing the experiments outnumbered the number of literature papers cited. Subsequently, there was a peak in the number of experiment papers in 2000 (the year it stopped operating) and then a decline. However, this is not the case for the literature papers citing the experiments, the number of which continued to increase.

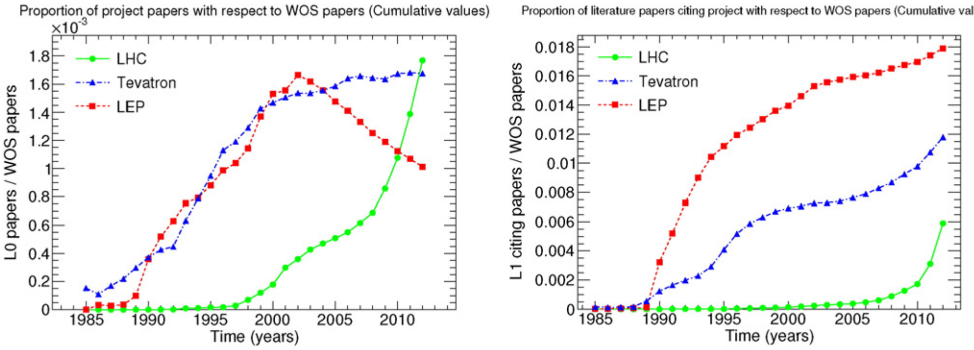
The Tevatron experiment paper trajectories (Tables A.11, and A.12), as with the LEP, show an intersection of the curves for literature papers that are cited and literature papers that cite the experiments a few years after it started. They also show a growth phase, with a small peak in

Table 1
Experiment papers (produced and published); experiment papers cited by literature papers and literature papers cited by experiment papers.

Project	Experiment	Papers *L0*	Papers *L0*_pub	Papers *1to0*	Papers *0to1*
LEP	ALEPH	636	589	383	3233
	DELPHI	736	670	417	3644
	L3	605	549	381	3563
	OPAL	694	634	475	4037
	Subtotal	2671	2442	1656	14,477
Tevatron	CDF	3077	2386	1641	6616
	D0	2383	1769	1176	4744
	Subtotal	5460	4155	2817	11,360
LHC	ALICE	1579	945	382	2963
	ATLAS	2529	1921	1195	4862
	CMS	2580	1603	1030	4640
	LHCb	735	585	248	1608
	Subtotal	7423	5054	2855	14,073

Table 2
Citations, H-index and number of renowned papers.

Project	Experiments	*0to0*	*0to1*	*1to0*	*1to1*	H-index	>500 cit
LEP	ALEPH	2244	11,075	22,475	241,877	77	4
	DELPHI	2170	12,800	18,482	206,600	66	4
	L3	2136	14,492	17,628	198,608	63	4
	OPAL	4659	18,993	25,469	243,995	79	4
	Subtotal	11,283	57,360	84,054	891,080	–	16
Tevatron	CDF	11,166	37,173	52,286	421,100	119	6
	D0	6216	25,676	29,758	280,703	85	3
	Subtotal	17,382	62,849	82,044	701,803	–	9
LHC	ALICE	1671	8169	3950	308,610	34	1
	ATLAS	7474	27,208	20,521	731,848	78	4
	CMS	5294	21,775	15,059	738,324	69	4
	LHCb	653	4117	2644	324,625	33	1
	Subtotal	15,092	61,269	42,174	2,103,407	–	10



(a) Project papers

(b) Literature papers

Fig. 3. The proportion of project papers on the left. The proportion of literature papers citing project on the right. In both cases data is normalized with respect to WOS papers. Results are presented as cumulative values.

2011 (the year in which it ceased) that decreased slightly but is not yet in the process of obsolescence. They also appear to benefit from the results of the LHC, given the extraordinary growth in literature papers that cite the experiments (more than 2000 in 2012 alone). Citations *1to1* in the tables highlight literature papers versus literature papers that cite experiment papers for LEP and Tevatron experiments, the number of which increased disproportionately as a result of diffusion of the results of LHC results. The LHC discoveries are likewise building on the scientific infrastructure of the past. Looking specifically at the trajectories of the citations, it can be seen that the quotes from outside sources about various experiments are always greater in number than those cited by the experiment papers. Regarding the LHC, citations are in the expansion phase (as the project is not finished); for Tevatron, they are at the point of maximum expansion (the project finished in 2011); and for LEP, they are in the process of obsolescence. Regarding LEP, the only research infrastructure for which all the steps have been completed, there is a peak in the number of citations immediately after the start of operations and soon after the end of the experiments.

The series of absolute values reported in the tables in Appendix A are useful to get an idea of the order of magnitude of the activity and impact measures for each experiment but cannot be used to compare projects or experiments that took place in different historical periods. Previously, Price (de Solla Price, 1986) talked about magnitudes of growth in "the size of science". To normalize the series, we used the trend of the number of physics articles published in journals found in the Web of Science each year from 1985 to 2012.[3] This series is presented in Table A.6 in Appendix A. For each experiment – for experiment papers and for literature papers that cite the experiments – we calculated cumulative values, and then we divided them by cumulative values of the series of physics papers. The next figures show the two ratios for the various projects.

The series of papers produced by the LEP and Tevatron experiments Fig. 3(a) show a concave shape, to indicate that at a certain point they will become stationary and then decreases. The curve of LEP, after it has been closed (2000), begins to decrease. Both series in the early years show a convex shape, which is the form that is observed for the LHC project, so that sooner or later, we expect a change of concavity and then a phase of stationarity and then of obsolescence. With regard to the paper of the literature citing the paper of the experiments, as was already noted, the phase of obsolescence has not yet been observed even for LEP which was closed for more than 10 years. This is even more evident from Fig. 3(b). Even in this case, LEP presents a concavity facing downwards and looks very close to the stationary phase. Tevatron seems still in a phase of expansion and LHC has an exponential growth.

To better see these trajectories, we report the same ratios for each experiment of the various projects in Fig. 4.

4. Towards the modeling of knowledge propagation in High Energy Physics (HEP)

A model which describes and provides predictions about the knowledge propagation in HEP is formulated by analyzing the citation distribution of papers of projects and its derivations. In the following paragraph we show an overview of such analysis by selecting a subclass of papers.

We selected three remarkable papers for the HEP physics community in terms of important discoveries, one paper for each project:

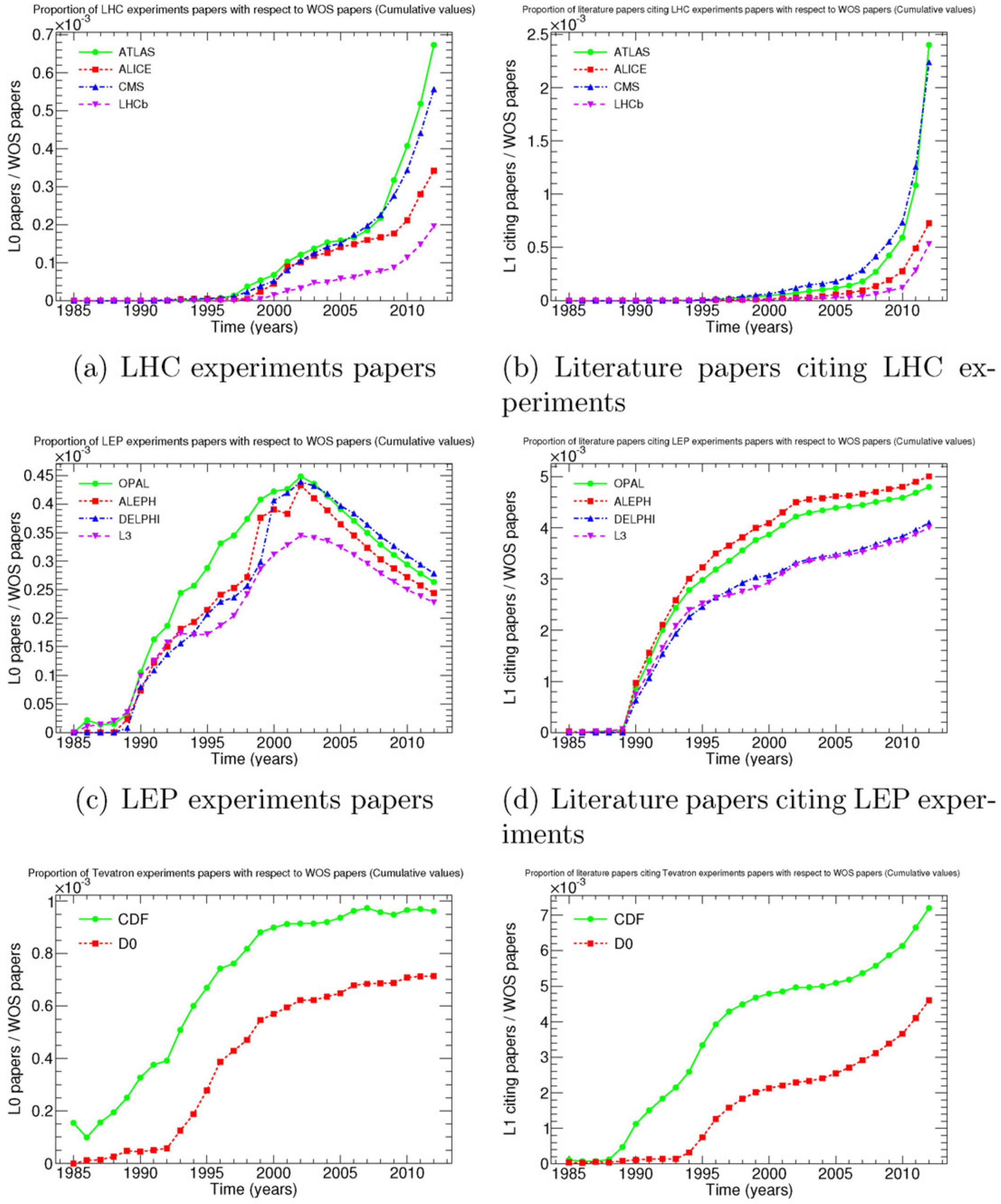
- LHC: the Higgs boson discovery by ATLAS Aad et al. (2012);
- Tevatron: the observation of top quark production by CDF Abe et al. (1995); and
- LEP: the determination of the number of light neutrinos species by ALEPH Decamp et al. (1989).

In Fig. 5 we show the absolute distribution of citations obtained from the respective *level 1* papers over time. We observe similarities between LEP and Tevatron distributions: there is a citation peak close to the publication date and a diffusion tail. Moreover, considering all the three distributions, we observe a strong correlation between the date of publication, the maximum number of citations and the width of the peak region. The impact of a remarkable paper in the scientific community is proportional to publication age: modern papers generate a strong wave of *level 1* papers, and the wave of knowledge continues longer in time. A possible explanation for the observed trend can be assigned to the continuous growth of the scientific community and its effort to achieve such remarkable results.

Table 3, shows for each of the three papers presented above, a summary with the total number of *level 1* publications and the H-index computed using their respective *level 1* papers. However, the original H-index definition does not take into account the *age* of an article. Ref. (Sidiropoulos et al., 2007) proposes the *contemporary H-index* (cH-index) in which the number of citations that an article has received is divided by the *age* of the article. The information reported by these estimators is fundamental to the construction of a model.

A generalization of the results presented above, for each paper in our database, provides a complete sample of HEP data from which we can extract a model. The model includes social factors, like how

[3] We query Web of Science (apps.webofknowledge.com): Advanced Search — Research Area Physics (SU = Physics).



(a) LHC experiments papers

(b) Literature papers citing LHC experiments

(c) LEP experiments papers

(d) Literature papers citing LEP experiments

(e) Tevatron experiments papers

(f) Literature papers citing Tevatron experiments

Fig. 4. Same as Fig. 3 but for single experiments.

the community propagates knowledge, and technological factors, e.g. project time, its lifetime cycle and the information diffusion. Such a model can determine and predict the impact of funding research infrastructures.

5. Clustering of papers based on citation patterns

Starting from the results of the previous section we tried to get a predictive knowledge output model for each paper in our database. We noticed that not all papers are equal in terms of citation trajectory. So it is not immediate to identify a parametric function. Moreover, for each experiment there are papers with different weights. The weight classifies the behavior from excellent to mediocre papers in terms of propagation impact. In principle, the weight distribution can be extracted from data. There are two issues we are working on:

1. Try to group the papers; and
2. Try to figure out if there are covariates that explain the different clusters.

The cluster of papers could depend on some covariates, such as the characteristics of the scientific community that produced them, and the number of authors involved. We deal with this point in the discussion section. We focus here on a methodology for the construction of clusters of papers based on the shape of their distribution of citations over time.

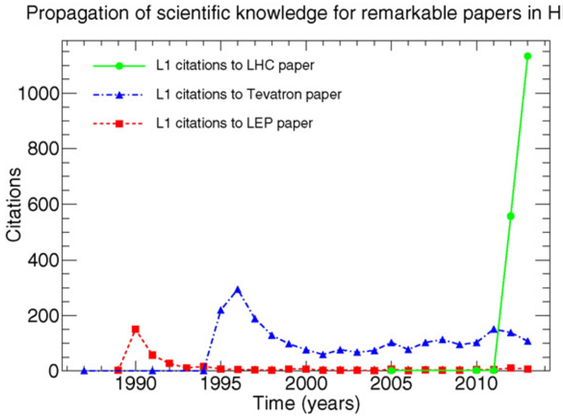


Fig. 5. Absolute distribution of citations over time for three remarkable papers for each project.

Paper citations distribution is normalized and shifted in order to compare papers published (and cited) in different time periods:

- shifting: the timeline of papers citations is shifted in such a way that all the citations are reported to a temporal range $t_0, t_1, \ldots, t_{n-1}, t_n$, where t_0 is the first year when a paper has been cited; and
- normalization: the number of citations C_p^y received by a paper p in the year y is normalized as follows:

$$norm\left(C_p^y\right) = \frac{C_p^y \cdot K}{C_{y.}}$$

where C_y is the total number of citations observed in the year y and K is a normalization factor.

5.1. Cluster methodology

We define a cluster of papers C_i as

$$C_i = \{p_1, p_2, \ldots, p_n\}, \tag{1}$$

where i is the index which identifies the cluster, p_j with $j = 1, \ldots, p_{n_i}$ is the n_i elements of the cluster i, i.e. the papers contained in C_i.

The cluster analysis of time series is a well-known problem studied in the literature (Nagin, 2009; Xie et al., 2010; Manrique-Vallier et al., 2014; Ho et al.,). Most of the relevant contributions on this problem start from the Group-based Trajectory Modeling (GBTM) (Nagin, 2009). GBTM provides a non-parametric statistics for distinguishing the developmental trajectories of sub-populations in sets. It is based on using mixed models for the prediction of different trajectories in the data. In particular Xie et al. (2010) present an evolution of GBTM for multidimensional outcomes and Manrique-Vallier et al. (2014) used the idea of mixed membership to relax the within-class homogeneity assumption. GBTM algorithm, while having the advantage of being able to include covariates both stationary and time dependent, has many limitations. First of all it assumes a priori a model for the response variable and uses polynomial models to estimate the trajectories; secondly, the number of groups must be fixed as well as the order of the polynomials that are assumed for each different trajectory. Finally, from the computational point of view, the model proves inefficient in the presence of a very large number of papers, and resulting in a large number of clusters. Ho et al. develop a probabilistic model for latent role analysis in time-varying networks, as well as an efficient variational EM algorithm for approximate inference and learning. Here we use Affinity Propagation (AP), by the messaging passing algorithm presented in Frey & Dueck (2007) where the authors show its impressive capability of grouping data with complex structure. The choice of this particular algorithm is motivated by its capability of determining automatically the number of final clusters without requiring as input *a prior* knowledge or guess of the number of clusters.

Table 3
Additional scientometric information for papers (Aad et al., 2012; Abe et al., 1995; Decamp et al., 1989).

Project	Paper	*L1* papers	H-index	cH-index
LHC	Aad et al. (2012)	1696	43	82
Tevatron	Abe et al. (1995)	2280	105	63
LEP	Decamp et al. (1989)	348	55	22

The clustering procedure that we adopt consists of the following steps:

- Data pre-processing: before starting the clustering procedure, we apply a pre-selection criterion for the input ensemble of papers. We define an ensemble of papers

$$E_k = \left\{p_i : N_{total}^{cit}(p_i) \geq k\right\} \tag{2}$$

where $N_{total}^{cit}(p_i)$ is the total number of citations that p_i received since its publication and k is a threshold value defined to filter the items of the ensemble. In our analysis we limited the threshold values to $k = 10{,}50{,}100{,}500$.

- Distance definition: there are several different definitions to quantify the similarity between elements of a given ensemble E_k of papers. In the AP framework, we construct a similarity matrix, defined as

$$S_{i,j} = -d\left(p_i, p_j\right), \tag{3}$$

where $d(p_i, p_j)$ is the distance estimator defined by the user. We performed the present cluster analysis with two different distance definitions: the dynamic time warping (DTW) (Müller, 2007) and the squared euclidean distance between points. For the DTW distance we use the raw distribution of citation for each paper, meanwhile for the squared euclidean distance we apply the normalization procedure presented at the beginning of this section.

- AP clustering: we perform the AP clustering with the damping factor $\lambda = 0.5$, a maximum of 200 iterations and 15 iterations with no change in the number of estimated clusters that stop the convergence.
- Multiple passes: due to the large number of elements that we are considering, the construction of large similarity matrices is not possible due to hardware limitations. In order to deal with such limitations we implemented an interactive procedure which compares the similarity between the available exemplars of a given cluster to the remaining papers. We call “pass” each time we compare exemplars to a new chunks of papers. This situation is more pronounced when applying pre-selection criteria where k is small.

5.2. Results

The ensemble of papers used in the clustering procedure presented here is the same as previously described in Section 3. In Table 3 we summarize the clustering results, for each of the four pre-selected ensemble of papers, $k = 10{,}50{,}100{,}500$, we build two similarity matrices based on the distance definitions presented above. We describe in details the features of such cluster in the next section.

Table 4
Summary of the clusters obtained with the affinity propagation method.

Collection	Distance	k	Papers	Clusters (size > 1)	Passes
cut500dtw	DTW	500	1453	107 (73)	1
cut100dtw	DTW	100	18,745	106 (71)	2
cut50dtw	DTW	50	43,595	245 (156)	2
cut10dtw	DTW	10	149,749	69 (47)	3
cut500euclidean	Euclidean	500	1453	70 (24)	1
cut100euclidean	Euclidean	100	18,745	60 (15)	2
cut50euclidean	Euclidean	50	43,595	171 (45)	2
cut10euclidean	Euclidean	10	149,749	436 (76)	2

Table 5
Average semantic and temporal dimensions of the cluster collections.

Collection	Size	Size ≥ 5	$avg(\text{size})$	$avg(\mathcal{S}^{C_i})$	$avg(\mathcal{T}^{C_i})$
cut500dtw	107	60	23.066	0.081	0.257
cut100dtw	106	55	339.327	0.169	0.300
cut50dtw	245	121	358.727	0.186	0.321
cut10dtw	69	36	3609.722	0.183	0.278
cut500euclidean	70	21	66.571	0.176	0.306
cut100euclidean	60	9	2075.889	0.247	0.246
cut50euclidean	171	22	1972.909	0.224	0.311
cut10euclidean	436	41	3641.244	0.241	0.332

6. Clusters description

The cluster collections presented in Table 3 have been calculated by working on the distribution of the citations received by papers in time. In other terms, the resulting clusters group together those papers that have been cited in a similar way during their life-cycle. Our hypothesis is that the citation analysis per se is a sufficient criterion for clustering together papers that have an affinity both from a temporal perspective and from a semantic perspective. In particular, we are interested in understanding if the citation behavior is based on the historical period in which the cited papers have been published and/or if it depends on the topics addressed by the papers. A correlation among temporal, semantic, and citation dimensions would justify the choice of the citations as a descriptive criterion for understanding the success of specific scientific topics in time. On the contrary, the discovery of substantial independence of these three dimensions would support the idea that the citation behavior is determined by factors (such as the popularity of author and institutions) that do not depend on the topic and the historical period of publication.

In order to study the cluster collections of Table 3 according to the semantic and temporal dimensions, we define a set of descriptive dimensions for clusters, based on a preliminary activity of semantic indexing of papers and the analysis of their years of publication.

6.1. Semantic indexing

The semantic indexing activity aims at associating each paper with a set of topics, each representing a latent variable in the corpus. We stress the fact that this activity is completely independent from the clustering activity described in Section 5.1. Indexing is based exclusively on the terms extracted from the paper titles, while clustering is based exclusively on the citations received by the papers. Formally, we define the semantic index $I(\mathcal{C})$ of a corpus $\mathcal{C}$ of n papers as follows:

$$I(\mathcal{C}) = \langle (p_1, T_1), (p_2, T_2), \ldots, (p_n, T_n) \rangle,$$

where p_i denotes a paper in $\mathcal{C}$, and $T_i = \{t_0, \ldots, t_k\}$ is a set of topics associated with p_i. In order to calculate $I(\mathcal{C})$, we exploit the well-known indexing approach based on Latent Semantic Analysis, which is often referred to Latent Semantic Indexing (LSI) (Deerwester et al., 1990). In the following, we briefly recall LSI in order to introduce the definition of $I(\mathcal{C})$. For LSI, we are interested in the $M \times N$ term-document matrix C, where rows represent terms and columns represent documents. In our case, terms have been extracted by the paper titles by means of standard natural language normalization techniques, including stemming and stop-words filtering. Documents are papers

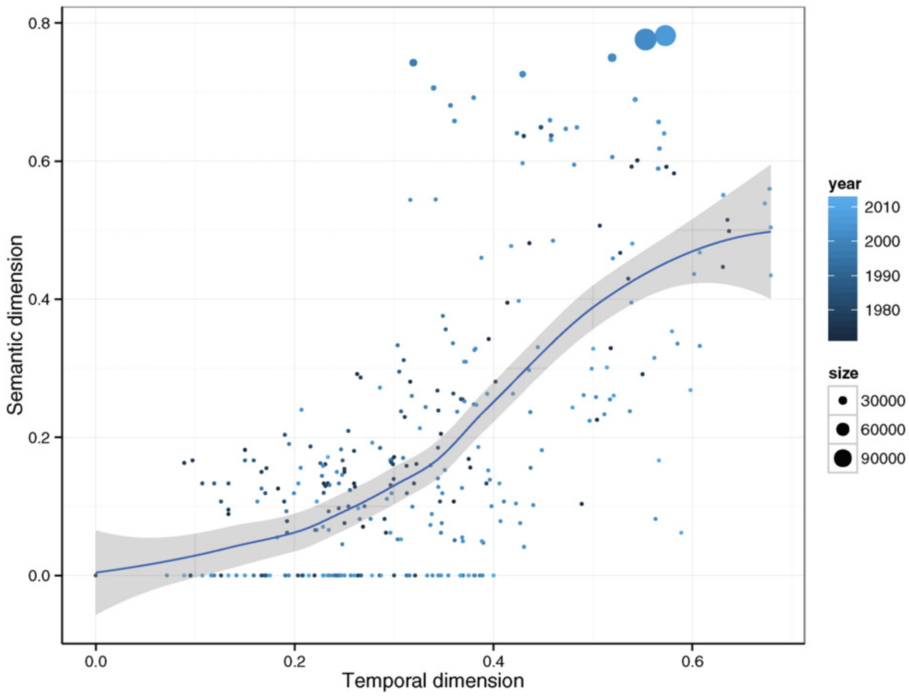


Fig. 6. Correlation between semantic and temporal dimensions in each cluster.

of the corpus $\mathcal{C}$. An entry (i,j) in the matrix C denotes the relevance of the ith term in the jth document, according to the term frequency–inverse document frequency (TfIdf) measure (Aizawa, 2000). According to this model, each paper p_j can be represented as a vector $\overrightarrow{v}(p_j)$. The idea behind LSI is to calculate an approximate version of the matrix C through its Singular Value Decomposition (SVD), such as:

$$C = \mathcal{U}\Sigma V^T,$$

where $\mathcal{U}$ is the $M \times M$ matrix whose columns are the orthogonal eigenvectors of CC^T and V^T is the transpose of the $N \times N$ matrix whose columns are the orthogonal eigenvectors of C^TC. The following step is to reduce the rank of C to an approximation of rank k. To this end, a matrix Σ_k is derived from Σ by replacing by zeros the r-k smallest singular values of the diagonal of Σ in order to compute $C_k = \mathcal{U}\Sigma_k V^T$ (Manning et al., 2008). The rank-k approximation of C can be now used in order to represent each document as a vector $\overrightarrow{v}_k(p_j)$ of k dimensions by mapping its original vector $\overrightarrow{v}(p_j)$ into the new k space as $\overrightarrow{v}_k(p_j) = \Sigma_k^{-1}\mathcal{U}_k^T\overrightarrow{v}(p_j)$. The intuition is that by reducing the number of dimensions we bring together terms with similar co-occurrences. This intuition, together with several empirical experiments made using LSI (Wolfe et al., 1998), leads to the conclusion that the k dimensions of the approximate vector space representation of the corpus can be interpreted as latent topics in the corpus.

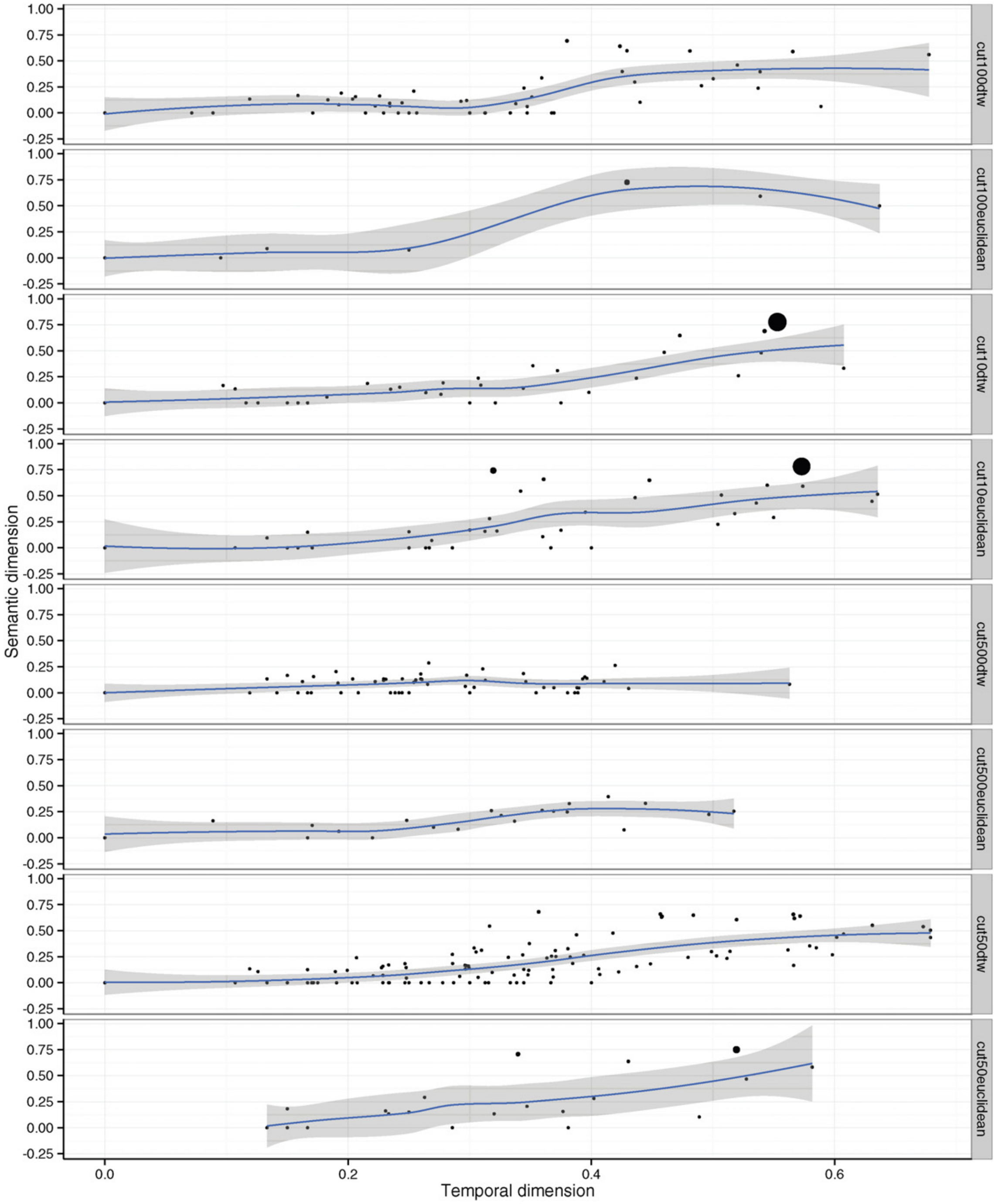


Fig. 7. Correlation between semantic and temporal dimensions with respect to different cluster collections.

In our process of indexing, we define a vector space of 400 dimensions (i.e., $k = 400$), which has been recommended as a good choice for LSI (Bradford, 2008). Given a paper p_i and its corresponding approximate vector $\vec{v}_k(p_i)$ with $k = 400$, we denote as $\vec{v}_k(p_i)[j]$ the contribution of p_i to the latent topic represented by the jth dimension of the matrix C_k. The idea is that the higher is the absolute value of $\vec{v}_k(p_i)[j]$, the higher is also the relevance of the topic t_j for the paper p_i. Following this intuition we empirically determined a threshold $th = 0.2$ in order to choose the topics to associate with p_i in the semantic index $I(\mathcal{C})$ as follows:

$$I(\mathcal{C})[p_i] = (p_i, T_i), where\ T_i = \left\{t_j, |\vec{v}_k(p_i)[j]| \geq th\right\}.$$

6.2. Descriptive dimensions

Our descriptive semantic ($\mathcal{S}^{C_i}$) and temporal ($\mathcal{T}^{C_i}$) dimensions provide a measure of the homogeneity of a cluster C_i with respect to topics and years of publication, respectively.

6.2.1. Semantic dimension

Given a cluster C_i, its semantic dimension $\mathcal{S}^{C_i}$ is calculated through the semantic index $I(\mathcal{C})$. In particular, we first determine the set $T(C_i)$ of topics involved in C_i as follows:

$$T(C_i) = \bigcup_{j=1}^{|C_i|} T_j | \exists \left(p_j, T_j\right) \in I(\mathcal{C}) : p_j \in C_i,$$

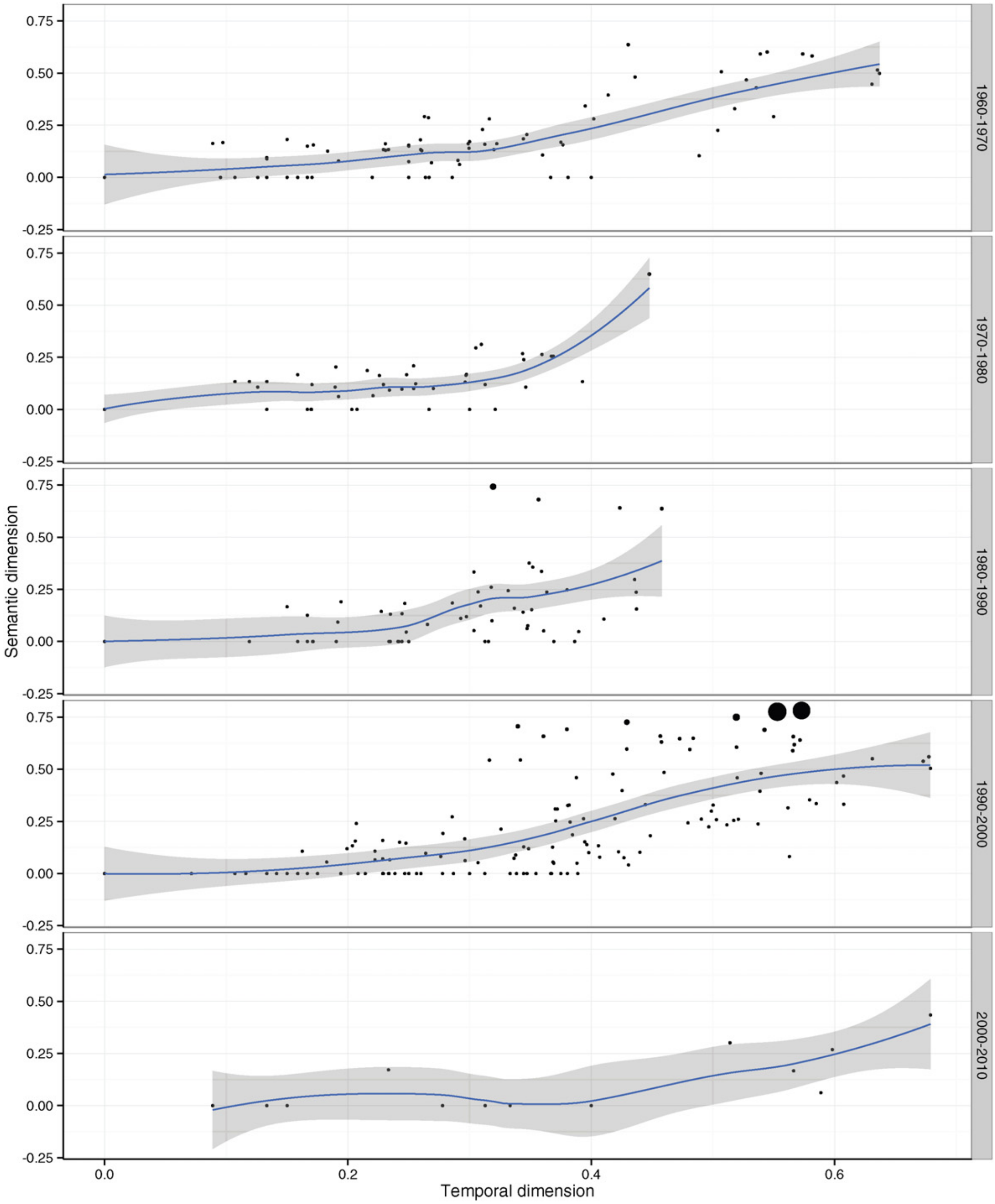


Fig. 8. Correlation between semantic and temporal dimensions in time.

where $|C_i|$ is the cardinality of C_i. Then, we associate with each topic $t_j \in T_j$ the number $N(t_j, C_i)$ of papers in C_i that corresponds to the topic t_j. In such a way, we obtain a distribution of papers in C_i over the set of topics T_j. On top of this distribution, we calculate the semantic dimension $\mathcal{S}^{C_i}$ of a cluster C_i as the Gini coefficient (Atkinson, 1970). Since it is basically a measure of inequality among values of the frequency distribution, low values of $\mathcal{S}^{C_i}$ represent an almost equal distribution of papers over the topics and, thus, a low level of semantic homogeneity of the cluster. On the contrary, when $\mathcal{S}^{C_i}$ is high, it means that there is a relatively small number of topics which is associated with many papers in C_i and, as a consequence, the cluster is homogenous from the semantic point of view.

6.2.2. Temporal dimension

Similarly to semantic dimension, the temporal dimension is based on the frequency distribution of papers over the years of publication. Also in this case, the temporal dimension $\mathcal{T}^{C_i}$ of a cluster C_i is calculated as the Gini coefficient of such a distribution. Low values represent an equal distribution over different years, while high values represent the presence of a limited number of years with a prevalence of papers.

6.3. Cluster analysis

According to the semantic and temporal dimensions described above, we analyze the cluster collections described in Table 4. In particular, for each collection, we calculate the semantic and temporal dimensions of all the clusters grouping at least 5 papers. This choice is motivated by the fact that we need a minimal number of papers in a cluster in order to adopt our dimensions based on the paper distribution over topics and years, respectively. The number of clusters involved in the analysis, as well as the average values of the semantic and temporal dimensions, is reported for each cluster collection in Table 5.

As we can see from Table 5, the clusters seem to be generally more characterized by the temporal rather than by the semantic dimension, as seen by the higher values of $\mathcal{T}^{C_i}$ with respect to $\mathcal{S}^{C_i}$. This result suggests that citations depend more on the year of publication of papers than on their topics. A more detailed analysis of the semantic and temporal dimensions is shown in Fig. 6.

As expected, we note a correlation between the semantic and the temporal dimensions: clusters grouping together papers published in the same year tend to be also homogeneous in terms of topics. This is due to the emergence of paradigms and specific topics in specific periods of time. However, there is also an interesting group of clusters with high levels of semantic homogeneity which are weakly homogeneous in terms of time. We note also that this group is composed by the largest clusters. This suggests the emergence of popular topics that produce a large number of papers for long periods of time.

The correlation between semantic and temporal dimensions by different cluster collections is shown in Fig. 7.

Here, we note that low cut thresholds (i.e., 10 and 50 citations) seem to produce results where the correlation is more evident and, in general, the level of semantic homogeneity is higher. In particular, those collections focus on highly cited papers only (i.e., cut equal to 500 citations) seem to be inadequate to capture both the temporal/semantic correlations and to produce semantically homogeneous clusters. A correlation between temporal and semantic homogeneity seems to be independently confirmed in case of clusters associated with different time periods, as shown in Fig. 8.

A final interesting result is given by the analysis of the correlation between semantic dimension and cluster size shown in Fig. 9.

In fact, one could expect that large clusters result in low levels of semantic homogeneity due to the high probability of clustering together papers addressing very different topics. Of course, the limited number of topics (i.e., 400) with respect to the size of the largest clusters determines the fact that topics are associated with many papers. But the relevant thing here is that the distribution is also highly unequal, which means that some topics prevail clearly over the others. The fact that the level of semantic homogeneity increases with the cluster size suggests the interesting consideration that the citations as a criterion of clustering are useful also for clustering together papers with the same or similar topics: a first (initial) confirmation of the hypothesis that the way papers are cited depends on the topics the papers address.

7. Summary and discussion

In this analysis, we examined publication trends and citations for various experiments related to major research infrastructures.

The aggregated analysis carried out indicates a regularity in the pattern of publications and citations for research infrastructures. First is a pre-experiment phase, in which the literature papers referred to by experiments are more numerous than the papers produced by the group that conducted the experiment. When the experiment starts, the experiment papers grow and from a certain point begin to increase alongside the literature papers mentioning the experiment. When the experiment produces the first results, there is usually a peak in internal publications and literature papers. From that moment, the number of publications begins to grow, eventually reaching a saturation point. We were only able to observe this phase for the LEP experiments. We note that the number of literature papers that cite other literature papers that cite experiment papers has not declined, even more than ten years after the experiments ended.

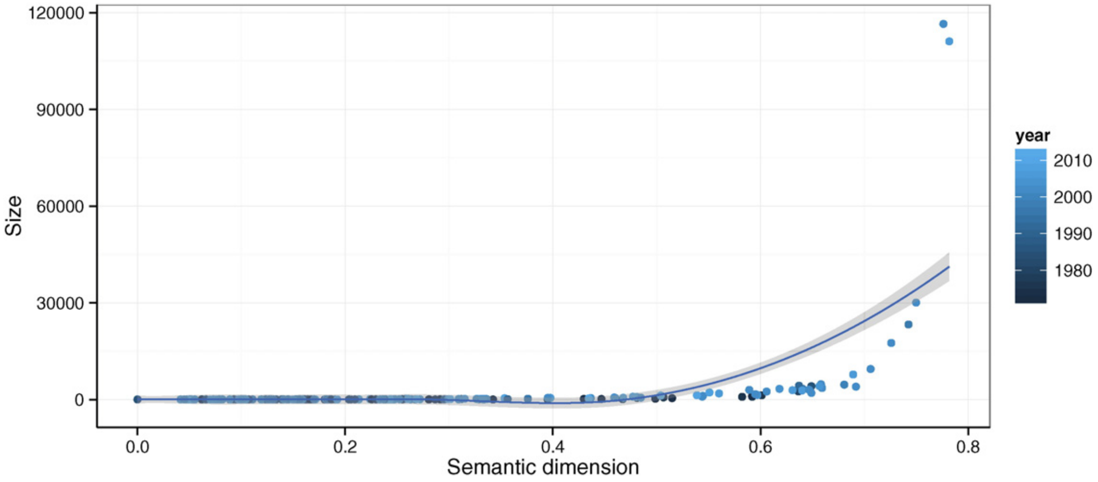


Fig. 9. Correlation between semantic dimension and cluster size.

The analysis of clusters of papers based on the shape of their distribution of citations over time shows a correlation between the semantic and the temporal dimensions. Moreover we discover important correlations between semantic dimension and cluster size; the level of semantic homogeneity increases with the cluster size. So, seems that using the citations as a criterion of clustering is useful also for clustering together papers with the same or similar topics. These conclusions are obviously valid for High Energy Physics. It is our intention to find out what happens instead in other disciplines, it will certainly be interesting.

Further developments can be achieved by: *i*) analyzing more in depth the clusters composition, also the co-citation network between the authors; *ii*) identifying clusters based on semantic topics and compare these collections with the ones obtained using the citations; *iii*) examining the cluster characteristics and connections and create a scientific map of HEP physics; *iv*) applying the clustering methodology to other fields; *v*) selecting possible covariates that explain the citation pattern for each cluster; and, last but not the least, *vi*) defining a theoretical model to describe and predict the growth of knowledge and the diffusion of project results and its uncertainty.

Acknowledgments

We are grateful for comments on earlier versions of the manuscript to Massimo Florio (Università degli Studi di Milano), Stefano Forte (Università degli Studi di Milano), Diana Hicks (Georgia Institute of Technology), Alessandro Sterlacchini (Università Politecnica delle Marche) and several others. We are also grateful to the anonymous referees, thanks to their comments the paper is substantially improved. This paper has been produced in the frame of the project Cost–Benefit Analysis in the Research, Development and Innovation Sector sponsored by the EIB University Research Program (EIBURS). S.C. is also supported by the HICCUP ERC Consolidation grant (614577). The findings, intepretatations and conclusions presented in the paper should not be attributed to the EIB or other institutions.

Appendix A. Descriptive Tables

Table A.6
Physics Articles (source: Web of Science).

	Year	Papers
1	1985	45,325
2	1986	45,559
3	1987	50,133
4	1988	54,246
5	1989	56,876
6	1990	59,760
7	1991	63,399
8	1992	64,352
9	1993	67,934
10	1994	72,256
11	1995	73,060
12	1996	80,813
13	1997	84,107
14	1998	83,547
15	1999	88,515
16	2000	88,375
17	2001	89,550
18	2002	94,631
19	2003	97,234
20	2004	103,074
21	2005	107,002
22	2006	112,565
23	2007	114,623
24	2008	118,945
25	2009	117,542
26	2010	117,978
27	2011	125,548
28	2012	125,883

Table A.7
ALEPH data.

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
20	1989	6	6	133	1	1	17	1	96
21	1990	17	17	160	302	29	18	216	450
22	1991	23	23	170	279	7	29	42	343
23	1992	20	20	148	341	6	15	105	207
24	1993	26	26	165	386	2	4	41	215
25	1994	21	20	166	431	10	7	30	286
26	1995	28	28	187	366	4	4	45	363
27	1996	39	37	179	461	4	11	54	517
28	1997	30	30	218	414	3	9	46	766
29	1998	40	38	159	455	3	6	65	477
30	1999	134	127	164	521	8	6	67	476
31	2000	52	49	136	451	2	6	49	464
32	2001	36	26	80	613	4	21	75	1038
33	2002	101	100	54	658	3	8	64	692
34	2003	13	11	40	519	0	6	44	422
35	2004	13	11	54	498	1	11	14	474
36	2005	9	3	28	551	2	15	36	414
37	2006	11	8	10	544	3	6	48	388
38	2007	3	1	13	589	0	9	4	481
39	2008	1	0	21	632	0	3	0	648
40	2009	7	3	12	662	0	7	4	601
41	2010	3	1	13	654	1	5	18	557
42	2011	1	1	9	829	0	0	0	810
43	2012	0	0	8	866	0	0	0	1998

Table A.8
DELPHI data.

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
20	1989	2	2	153	1	0	3	1	111
21	1990	23	23	176	195	22	44	89	407
22	1991	16	16	170	204	2	14	6	269
23	1992	19	19	165	273	3	4	109	185
24	1993	19	19	173	306	1	7	26	152
25	1994	22	22	181	329	6	3	21	198
26	1995	34	34	200	292	5	12	45	254
27	1996	35	33	209	333	15	15	38	383
28	1997	25	25	241	334	2	15	38	628
29	1998	40	38	209	368	5	15	24	356
30	1999	67	64	193	366	1	4	31	335
31	2000	146	143	144	308	4	9	30	258
32	2001	76	52	92	383	14	18	150	425
33	2002	67	64	83	493	0	9	5	450
34	2003	36	33	74	416	3	5	54	421
35	2004	29	25	68	426	4	13	26	421
36	2005	18	11	28	433	2	16	26	307
37	2006	26	22	12	470	3	4	52	388
38	2007	11	8	11	515	4	6	5	436
39	2008	6	5	16	608	0	3	5	536
40	2009	10	6	12	597	0	7	6	567
41	2010	3	2	9	591	0	4	7	478
42	2011	3	3	9	769	0	0	1	791
43	2012	0	0	9	836	0	0	0	2026

Table A.9
L3 data.

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
20	1989	5	5	150	9	0	6	3	111
21	1990	22	22	181	218	28	32	79	447
22	1991	16	16	180	210	8	9	46	270
23	1992	22	22	146	281	10	18	104	183
24	1993	19	19	157	330	5	4	16	170
25	1994	11	11	177	329	1	5	14	202
26	1995	14	13	204	260	0	3	31	263
27	1996	26	25	210	288	1	11	41	342
28	1997	31	30	203	260	19	24	36	391
29	1998	51	51	178	286	4	17	23	307
30	1999	67	65	192	317	10	16	40	322

(continued on next page)

Table A.9 (*continued*)

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
31	2000	57	53	138	363	10	30	64	359
32	2001	57	47	103	467	5	29	87	590
33	2002	58	52	80	505	2	13	18	422
34	2003	29	28	57	420	3	10	51	305
35	2004	36	28	58	415	7	12	30	356
36	2005	24	18	37	426	5	17	34	310
37	2006	18	14	24	464	1	7	46	347
38	2007	11	8	21	481	0	9	3	381
39	2008	3	2	16	587	0	3	0	495
40	2009	7	3	14	579	0	8	4	532
41	2010	4	2	13	568	0	4	0	464
42	2011	6	5	11	743	0	2	1	780
43	2012	3	3	12	816	0	0	0	1959

Table A.10
OPAL data.

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
19	1989	5	5	175	6	3	15	1	126
20	1990	25	25	185	260	15	28	120	514
21	1991	28	28	172	254	14	27	43	322
22	1992	22	21	203	353	7	18	95	232
23	1993	42	42	180	354	16	5	33	195
24	1994	26	25	180	380	5	14	19	255
25	1995	39	39	219	332	7	9	41	355
26	1996	57	55	234	389	29	49	31	512
27	1997	42	39	261	407	5	34	48	821
28	1998	56	55	217	466	2	23	54	479
29	1999	69	67	205	514	0	17	64	515
30	2000	54	51	156	449	2	20	55	424
31	2001	54	43	110	559	4	33	142	826
32	2002	71	68	64	600	4	14	22	586
33	2003	27	26	47	510	2	9	55	430
34	2004	18	14	54	510	1	13	24	453
35	2005	16	8	29	547	4	13	28	378
36	2006	15	9	21	543	0	7	47	431
37	2007	8	5	13	552	0	10	4	448
38	2008	3	2	21	640	0	4	2	607
39	2009	8	2	15	629	0	7	4	611
40	2010	1	0	9	612	0	4	0	502
41	2011	2	1	9	802	0	0	4	779
42	2012	0	0	9	856	0	0	0	2116

Table A.11
CDF data.

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
16	1983	0	0	89	2	0	0	0	82
17	1984	3	3	78	2	0	0	0	45
18	1985	7	7	86	5	0	0	3	29
19	1986	2	2	95	2	0	0	0	23
20	1987	13	13	121	4	0	1	0	68
21	1988	16	16	101	13	15	10	5	21
22	1989	25	25	150	96	14	11	83	145
23	1990	41	39	160	230	11	8	77	211
24	1991	40	39	153	216	7	14	6	233
25	1992	31	31	130	242	9	7	50	126
26	1993	86	86	149	284	2	3	35	142
27	1994	98	90	185	411	27	17	177	365
28	1995	97	89	219	677	36	26	260	860
29	1996	116	108	280	700	21	83	82	1255
30	1997	86	78	280	629	22	39	87	1184
31	1998	133	115	273	540	20	54	73	730
32	1999	156	134	286	583	11	37	76	843
33	2000	108	97	237	532	15	16	57	768
34	2001	107	96	210	504	12	21	39	729
35	2002	107	89	232	604	14	62	24	887
36	2003	109	89	238	485	6	31	63	740
37	2004	142	102	244	555	36	58	151	1091
38	2005	182	125	180	681	52	30	144	787
39	2006	194	149	210	732	60	48	221	1126
40	2007	216	130	174	925	61	59	229	1220
41	2008	184	85	227	1039	65	136	244	1662
42	2009	169	94	164	1249	70	64	380	1873
43	2010	186	150	170	1265	58	76	247	2313
44	2011	188	131	235	1948	120	205	684	6501
45	2012	134	101	215	2142	75	222	508	8711

Table A.12
D0 data.

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
15	1983	2	1	50	0	1	3	0	56
16	1984	0	0	40	2	0	0	0	13
17	1985	0	0	38	2	0	0	0	4
18	1986	1	1	57	1	0	0	0	14
19	1987	1	1	66	5	0	0	0	31
20	1988	3	3	53	1	1	1	0	4
21	1989	7	7	90	11	0	0	4	41
22	1990	2	2	97	14	0	0	0	48
23	1991	5	5	103	17	0	0	8	29
24	1992	6	6	103	12	0	0	0	43
25	1993	39	38	131	10	2	3	0	58
26	1994	63	46	130	114	8	11	79	143
27	1995	77	73	146	303	22	22	236	517
28	1996	111	102	180	433	20	23	40	707
29	1997	80	67	191	373	33	53	78	670
30	1998	89	73	182	353	15	27	63	449
31	1999	134	117	194	348	37	19	71	502
32	2000	81	73	166	304	10	19	32	357
33	2001	98	80	165	278	4	11	22	364
34	2002	105	91	172	318	10	23	16	418
35	2003	78	61	186	276	7	13	43	347
36	2004	113	83	184	348	25	36	169	543
37	2005	144	89	161	485	38	28	139	496
38	2006	159	124	171	547	35	40	208	1044
39	2007	158	88	169	686	54	35	237	879
40	2008	154	85	181	731	121	79	367	1014
41	2009	149	84	163	914	96	59	292	1264
42	2010	156	125	139	992	67	71	340	1962
43	2011	156	98	211	1469	135	204	434	5316
44	2012	135	95	173	1710	90	173	478	7442

Table A.13
ALICE data.

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
26	1993	2	2	57	0	0	0	0	235
27	1994	1	1	65	2	0	0	0	210
28	1995	0	0	65	1	0	0	0	359
29	1996	1	0	73	3	0	0	0	753
30	1997	1	1	90	2	0	0	0	1066
31	1998	1	1	114	2	0	0	0	555
32	1999	19	19	130	5	0	2	2	566
33	2000	24	24	147	5	0	0	1	839
34	2001	74	57	174	6	1	4	0	1065
35	2002	23	22	159	8	0	0	1	862
36	2003	34	34	162	9	1	13	3	1058
37	2004	32	23	180	19	1	8	5	895
38	2005	46	37	153	28	3	5	5	786
39	2006	39	30	146	30	1	18	7	889
40	2007	56	36	153	53	0	5	2	769
41	2008	43	33	148	90	6	9	8	926
42	2009	62	40	154	129	7	14	6	751
43	2010	112	95	157	202	59	66	159	1078
44	2011	604	184	222	527	72	131	129	1348
45	2012	240	184	137	630	226	160	213	1153

Table A.14
ATLAS data.

	Year	l0	l0_published	l1cited	l1citing	X0to0	X0to1	X1to0	X1to1
26	1993	1	1	63	0	0	0	0	398
27	1994	0	0	56	0	0	0	0	766
28	1995	3	3	86	2	0	0	2	1626

Table A.14 (*continued*)

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
29	1996	1	1	80	6	0	0	0	1817
30	1997	6	6	92	7	0	2	0	2532
31	1998	26	23	110	9	0	1	12	1423
32	1999	20	19	136	9	0	0	1	1675
33	2000	20	20	140	19	0	1	0	1319
34	2001	49	47	157	16	4	2	4	2074
35	2002	36	33	179	23	0	4	7	1624
36	2003	36	34	198	32	2	1	3	1352
37	2004	42	38	228	28	5	5	6	1620
38	2005	37	24	196	31	0	5	4	1279
39	2006	46	32	266	55	2	3	7	1908
40	2007	93	52	292	88	4	19	31	1762
41	2008	142	84	333	192	51	69	75	2294
42	2009	267	228	284	345	37	21	190	2375
43	2010	265	232	334	410	63	56	85	2972
44	2011	381	303	526	1189	263	289	929	7186
45	2012	1048	438	441	3306	2841	592	5535	13,600

Table A.15
CMS data.

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
27	1993	1	1	67	0	0	0	0	440
28	1994	1	1	57	0	0	0	0	740
29	1995	1	1	85	3	0	0	0	1546
30	1996	2	1	96	8	0	0	0	1873
31	1997	6	4	107	6	0	2	0	2544
32	1998	12	12	99	17	0	0	8	1117
33	1999	17	17	142	17	0	0	1	1471
34	2000	18	18	148	17	0	0	0	1310
35	2001	39	38	164	35	0	0	9	2033
36	2002	41	39	195	46	1	2	22	1515
37	2003	40	37	197	53	5	4	2	1330
38	2004	44	38	204	31	0	3	6	1500
39	2005	43	29	225	51	0	1	5	1393
40	2006	77	54	246	91	1	6	8	1726
41	2007	98	63	270	140	20	17	83	1780
42	2008	126	79	315	281	28	40	62	2526
43	2009	155	129	320	327	18	32	11	2775
44	2010	242	178	376	456	44	72	187	3186
45	2011	579	265	461	1279	889	320	1148	7462
46	2012	572	334	441	2516	702	512	3366	14,324

Table A.16
LHCb data.

	Year	*l0*	*l0*_published	*l1*cited	*l1*citing	X*0to0*	X*0to1*	X*1to0*	X*1to1*
25	1993	0	0	25	0	0	0	0	168
26	1994	0	0	25	0	0	0	0	271
27	1995	0	0	31	0	0	0	0	469
28	1996	0	0	29	0	0	0	0	768
29	1997	0	0	30	0	0	0	0	824
30	1998	3	3	40	1	0	0	2	508
31	1999	1	1	58	4	0	0	0	699
32	2000	12	12	55	2	0	0	1	732
33	2001	14	14	60	9	0	3	2	1059
34	2002	11	11	70	3	5	5	0	990
35	2003	23	23	83	4	0	0	1	1119
36	2004	7	7	84	8	0	2	1	1144
37	2005	28	20	99	12	0	0	5	916
38	2006	16	13	141	8	0	1	0	1810
39	2007	46	27	151	26	2	9	2	1558
40	2008	19	18	135	41	0	6	5	1462
41	2009	37	28	138	72	0	3	7	1384
42	2010	82	67	157	72	9	10	14	1735
43	2011	127	92	218	384	52	78	131	1507
44	2012	158	135	200	632	175	208	421	2265

References

Aad, G., et al., 2012. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys. Lett. B 716, 1–29. http://dx.doi.org/10.1016/j.physletb.2012.08.020 (arXiv:1207.7214).

Abe, F., et al., 1995. Observation of top quark production in $\overline{p}p$ collisions. Phys. Rev. Lett. 74, 2626–2631. http://dx.doi.org/10.1103/PhysRevLett.74.2626 (arXiv:hep-ex/9,503,002).

Aizawa, A., 2000. The feature quantity: an information theoretic perspective of tfidf-like measures. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 104–111.

Atkinson, A.B., 1970. On the measurement of inequality. J. Econ. Theory 2 (3), 244–263.

Bradford, R.B., 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, pp. 153–162.

de Solla Price, D., 1986. Little Science, Big Science… and Beyond. Columbia University Press, New York.

Decamp, D., et al., 1989. Determination of the number of light neutrino species. Phys. Lett. B 231, 519. http://dx.doi.org/10.1016/0370–2693(89)90704–1.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. Technol. 41 (6), 391–407.

M. Florio, S. Forte, E. Sirtori, Cost–Benefit Analysis of the Large Hadron Collider to 2025 and Beyond arXiv:1507.05638.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 315 (5814), 972–976.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A. 102 (46), 16,569–16,572.

Q. Ho, L. Song, E. P. Xing, Evolving Cluster Mixed-Membership Blockmodel for Time-Varying Networks

Manning, C.D., Raghavan, P., Schütze, H., et al., 2008. Introduction to Information Retrieval. vol. 1. Cambridge University Press, Cambridge.

Manrique-Vallier, D., et al., 2014. Longitudinal mixed membership trajectory models for disability survey data. Ann. Appl. Stat. 8 (4), 2268–2291.

Martin, B.R., 1996. The use of multiple indicators in the assessment of basic research. Scientometrics 36 (3), 343–362.

Martin, B.R., Irvine, J., 1984. Cern: past performance and future prospects: I. Cern's position in world high-energy physics. Res. Policy 13 (4), 183–210.

Müller, M., 2007. Dynamic time warping. Information Retrieval for Music and Motion, pp. 69–84.

Nagin, D., 2009. Group-Based Modeling of Development. Harvard University Press.

Sidiropoulos, A., Katsaros, D., Manolopoulos, Y., 2007. Generalized hirsch h-index for disclosing latent facts in citation networks. Scientometrics 72 (2), 253–280.

Wolfe, M.B., Schreiner, M., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K., 1998. Learning from text: matching readers and texts by latent semantic analysis. Discourse Process. 25 (2–3), 309–336.

Xie, H., McHugo, G.J., He, X., Drake, R.E., 2010. Using the group-based dual trajectory model to analyze two related longitudinal outcomes. J. Drug Issues 40 (1), 45–61.

Stefano Carrazza is a research fellow at CERN in Theoretical Physics. He graduated in Particle Physics and Field Theory at Ecole Normale Superieure de Lyon, and he received his Ph.D. in Theoretical Particle Physics in 2015 at the University of Milan, Italy, under the supervision of Prof. S. Forte. His thesis topic is focused on the precision study of the nucleon structure using the new data produced by the LHC at CERN, he is currently a member of the NNPDF collaboration. His research interest is focused on collider physics, LHC phenomenology and electroweak corrections to perturbative QCD.

Alfio Ferrara is an associate professor of Computer Science at the University of Milan, where he received his Ph.D. in Computer Science in 2005. His research interests include database and semi-structured data integration, Web-based information systems, data analysis, and knowledge representation and evolution. On these topics, he worked in national and international research projects, including the recent EU FP6 BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) project, the FP6 INTEROP NoE (Interoperability Research for Networked Enterprises Applications and Software) project, and the ESTEEM (Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents) PRIN project funded by the Italian Ministry of Education, University, and Research. He is also an author of several articles and papers in international journals and conferences about ontology management and matching.

Silvia Salini is an associate professor of Statistics in the Department of Economics, Management and Quantitative Methods of the the University of Milan. She earned her Bachelor in Statistics from the Catholic University of Milan, Italy, in 1999. She took a Ph.D. in Statistics at the University of Milan-Bicocca, Italy, in 2002. She worked for a long time on data mining methods and big data analytics robust statistics and causal models. She has recently focused her research on the methodological problem of measuring outcomes of the higher education system. She has been involved in several research projects on assessment and evaluation. Among them, the most relevant are: Cost/Benefit Analysis in the Research Development and Innovation Sector — Funded by the European Investment Bank — University Research Sponsorship Program (EIBURS), MIUR PRIN MISURA Multivariate Models for Risk Assessment, Cariplo Foundation project 2009, The Quality of the Education System in Lombardy: Measurement, International Comparisons and Proposals, and FP6 Understanding Privatization Policy: Political Economy and Welfare Effects. On these topics she has publications on many international journals.