

USING CONTENT ANALYSIS TO INVESTIGATE THE RESEARCH PATHS CHOSEN BY SCIENTISTS OVER TIME¹

Chiara Franzoni*, Christopher L. Simpkins**, Baoli Li **, Ashwin Ram **

* *Correspondence: Chiara Franzoni, DISPEA, Polytechnic of Turin, Corso Duca degli Abruzzi 24b, Torino, 10129, Italy. tel: +39.011.23414018; e-mail: chiara.franzoni@polito.it*

** *College of Computing, Georgia Institute of Technology, Atlanta, GA, USA.*

Important Note: *This paper was prepared for presentation at the 2007 ISSI Conference to be held in Madrid, 25-27 June 2007.*

Abstract

We present an application of a clustering technique to a large original dataset of SCI publications which is capable at disentangling the different research lines followed by a scientist, their duration over time and the intensity of effort devoted to each of them. Information are obtained by means of software-assisted content analysis, based on the co-occurrence of words in the full abstract and title of a set of SCI publications authored by 650 American star-physicists across 17 years. We estimated that scientists in our dataset over the time span contributed on average to 16 different research lines lasting on average 3.5 years and published nearly 5 publication in each single line of research. The technique is potentially useful for scholars studying science and the research community, as well as for research agencies, to evaluate if the scientist is new to the topic and for librarians, to collect timely biographic information.

Keywords

Content analysis, academic scientists, semantic search, research trajectories, knowledge development.

¹ This work was supported by the CERIS, National Research Council of Italy and was done while one of the authors (Chiara Franzoni) was kindly hosted as visiting scholar at the Andrew Young School of Policy Studies (Georgia State University, Atlanta, GA). The authors wish to thank Paula Stephan and Francesco Lissoni for comments and suggestions. All usual disclaims apply.

Introduction

In recent years an increasing number of studies have concentrated on analyzing the work and behavior of individual scientists within large databases of scientific publications. The approach of taking single individuals as the main unit of analysis in large field studies took off in the 1970s within the Sociology of Science and is becoming increasingly widespread for applications of Economics of Science and Innovation and for studies of Labor Market for research.

Common bibliometric applications make use of indicators derived from scientific publications to characterize the attributes of single scientists. Citations received (Cole and Cole, 1967; Narin and Hamilton, 1996; Garfield, 1979), co-authorship (Hicks and Hamilton, 1999) and co-citation analyses (Peritz, 1992), for instance, are used to add non-subjective information to the profile of single scientists and of their academic production.

Scientific publications, of course, contain many more information than their sheer author and citations, but the bulk of information is hidden into their text and understandable only to peer-readers. To overcome these problems, it is possible to make use of Software-Assisted Content Analysis, which consists in extracting and organizing non-structured information from plain text, by means of semantics, into a standardized format suitable for several different uses. Among the other things, this set of techniques allow inferring certain characteristics and meaning of full texts, and at the same time offer the advantages of being replicable for very large sets of data in relatively short times, allowing non-subjective and unskilled reading.

In the last decades, applications of Content Analysis to scientific publications (for instance co-word and semantic analysis), have quite extensively been used to map the state and evolution of single or multiple scientific disciplines (Courtial et al., 1997; Klavans and Boyack, 2005), but have not been used so far to characterize individual scientists in terms of interests and topics of inquiry, and of their evolution over time.

In this work we propose an application of clustering algorithms to analyze the topic of inquiry (and their evolution) followed by individual scientists across a time-span of a 15 years by applying Content Analysis in a convenient way. The storage of texts both in cross-sectional and longitudinal dimensions is suitable for work that aims at analyzing the state and evolution of a single scientist's research. For instance, it makes possible to identify the different lines of research followed by a scientist at a specific point in time and along the years, to spot when a scientist enters a new topic (either new to him, or new to the entire set of documents), or abandons it, and to appreciate the amount of effort devoted to different streams of one's production.

Applications of this methodology are useful in a large number of areas, including works on careers in science (Fox; 1983; Stephan and Levin, 1992), on the production and dissemination of knowledge along research trajectories (Hackett et al., 2004), on the functioning of the scientific communities (Crane, 1972), on policy of science and research (Godin, 2003) and more generally on the dynamics of science and technology (Gibbons et al., 1994; Leydesdorff, 2002). In principle, this methodology can be used for several purposes, other than research. It can be used by librarians seeking to obtain biographic information on the topics and subfields addressed by a single author and it can be useful for granting agencies, which are generally interested in knowing the response of scientists to the choice of program funded, for instance to see whether or not receiving funds in a certain program is likely to divert the natural path of research.

The paper is organized as follows: in the next section we address the choice of research paths followed by individual scientists. We then present the dataset collection procedure and final information stored (section "Data"). In the "Methodology" section we describe the clustering algorithm adopted and we lastly describe and comment the results in the final section.

Research Paths followed by Scientists in Academia

The choice of a research line in academia is a delicate process of balancing curiosity and opportunity. On the one side, academics enjoy the freedom of choosing their topics of enquiry according to their attitude and curiosity. Novelty in science and research pays-off with the highest rewards, as of course originality is among the general objectives of all studies (Hagstrom, 1965). On the other side, once a path has been chosen, shifting towards new interests is costly and time consuming and, in many cases, non advisable over short periods.

Indeed, one feature of research that looks immediately evident while observing science directly is that the majority of scientists follow quite linear research paths along their careers (Ziman, 1968), in the sense that they exploit research lines over medium-to-long horizons and move to new topics only gradually.

Of course, to a great extent, linearity is a function of specialization, which comes along with the ever-increasing complexity of research.

In all disciplines, the formal training given in undergraduate and graduate programs comprises all the foundations of a topic. Yet, specialties are chosen by young scientists quite early in their career, at the latest upon completion of graduate studies, when they are required to work at their PhD dissertations. Whereas in the early years PhD students may try several research lines, later on, specialization prevails over change, which will become rather infrequent and mostly gradual (Hagstrom, 1965).

The reasons behind this tendency are to be searched in the fact that switching to a new line of research is extremely costly for a scientist in several different ways. First, it requires time and effort. For instance, scientists interviewed in several sub-disciplines of Life Sciences indicated that addressing a new topic required at least one year of work before any result could be published (Hackett, 2005). This circumstance is particularly uncomfortable in highly competitive job environments, where the score of publications is extremely important for personal careers. To face the problems of discontinuity of research lines, mature scientists often keep open several different lines of research simultaneously, which helps diversifying the risk of being tied to a single unfruitful research.

Second, in almost all hard sciences, every specific research field requires investing in a set of instruments and techniques and establishing effective collaborations. In this respect, the change of a topic is not only slow because it requires learning the foundations and the state of art of a partially new subject, but also because it requires a general adjustment of techniques, team of research, and equipment (a set of assets that scholars of science address as “research ensemble”), that claims for some enduring interest (Hackett et al., 2004).

Third, the linearity in one’s research paths are mirrored and reinforced by the fact that, in pursuing their single career strategies, scientists see great benefits from building up and maintaining a strong personal identity within their community of reference. Within hard sciences, there are at least two sorts of benefits associated to establishing a clear identity. First, within a research groups or a laboratory, the identity of an individual is based heavily on the track of tasks and duties that he or she has accomplished over the years, including the technical skills accrued at the bench (Hackett, 2005). Acquiring a relevant experience and being knowledgeable in a specific subject is an advantage in the job-market and often the mobility of younger scientists trained in leading institutions is motivated by the desire of acquiring knowledge on certain cutting-edge processes. Second, for senior scientists, a clear identity serves to gain legitimacy on the eyes of colleagues, when addressing a certain theme: working heavily on a topic, posting contributions on journals, participating at conferences and events where the research base is nurtured and results disseminated, are necessary to establish a reputation among the colleagues and peers. As well known, this mechanism was heavily discussed by Merton and colleagues, which highlighted and empirically proved the importance of cumulativeness in science brought by personal recognition, which encourages repeated participation and persistence (Merton, 1968; Allison and Stewart, 1974). Thus, an additional reason for career-minded scientists to build cumulatively on their past work, stands in the fact that being regarded for few distinctive features of their preparation, interests and achievements, as validated by the scientific community, allows certain status benefits that are otherwise unavailable to people new to a subject. A clear identity serves essentially to gain legitimacy when addressing a certain theme, and enhances the probability of obtaining support and credit in research, both in terms of research grants and support (for instance all funding agencies appreciate the fitness of a curriculum with the stated objectives of a research), and in

terms of visibility vis-à-vis their community of reference (being invited to speak at conferences, write on special issues, sit in journals editorial boards, etc.).

Lastly, because of the well documented fads and fashion existing in science, persistence along an established domain offers a low risk strategy in research contexts characterized by strong competition (Garner, 1979).

Having addressed the rationale behind novelty and specialization, we expect that the position in the spectrum going from the one end to the other is chosen by scientists quite carefully and that we can hence characterize the scientific production of researchers with respect to the rate at which they enter into new topics, the portfolio of research lines they pursue simultaneously, the duration of those research lines and the intensity of contributions made in every specific line.

In the following sections we offer an empirical assessment of these characteristics by making use of a novel technique.

Data

The original database used for the study comprises a large group of scientists doing research in American universities in all the various subfields of Physics. The collection of data was started from a list of names obtained from the American Physical Society (APS) archive of Fellows. The APS has a very large membership base, and virtually all USA physicists belong to it since their doctoral studies. The status of “Fellow” is a life-long honorary title given year after year to a very small number of scientists (by internal regulation, to a maximum of 0.5% of the members), in recognition of their scholarly merits. The selection of Fellows is made annually in a peer-review fashion, starting with the nomination of meritorious scientists from the APS members worldwide and ending with the appointment of awards by the subgroup field to which the individual has mostly contributed. The use of the APS Fellows archive hence offers the advantage of having a good, unbiased indication of the subfield of Physics to which a scientist belongs, along with a synthetic description of the major contributions for which the individuals are regarded, which may prove to be of special interest for future research.

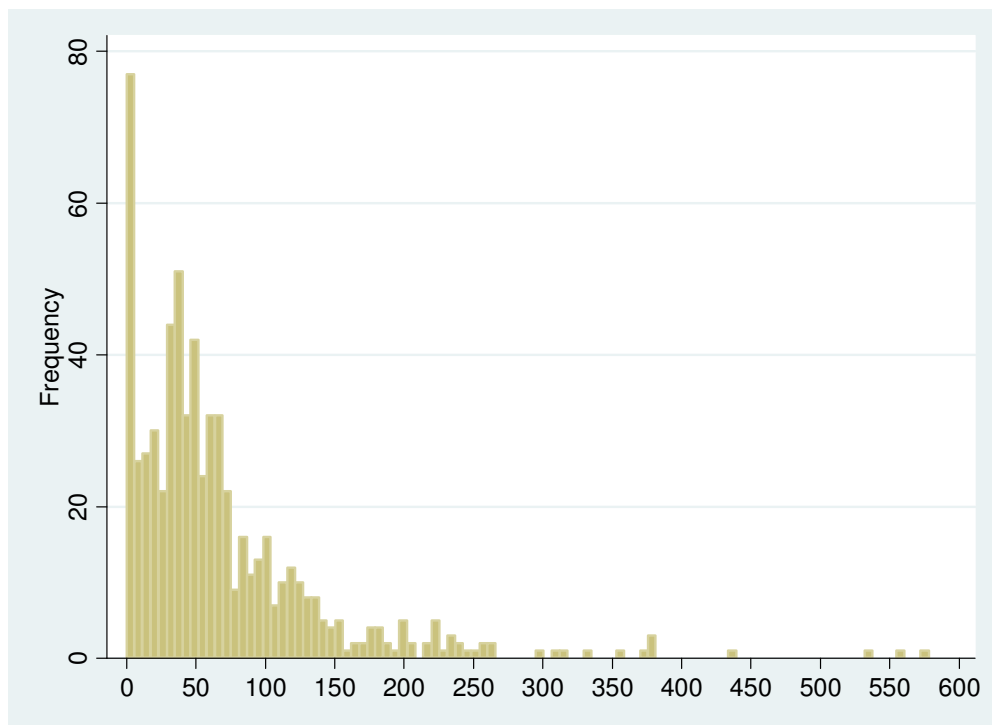


Figure 1. Total number of publication per scientist. Frequency Distribution.

We started by collecting all (1054) names of fellows awarded since 1995 to 2002, which were affiliated to US American universities at the time of the nomination. For 933 of these individuals (88%) we were able to retrieve full CV information through web search. After drops of people retired in the period of observation and of common names, we obtained a list of 650 individuals. From the ISI Science Citation Index we extracted information on all publications made by each individual on scientific journals since 1990 to the beginning of 2006 and kept only publications for which abstracts were available. The database resulted in 45,342 unique combinations of scientist-publications and 38,178 unique SCI publications, all recorded as references and accompanied by the full abstracts content. Figure 1 shows the distribution of the total number of publications per scientist.

Clustering Methodology

The clustering algorithm is described in Figure 2. Each paper constitutes a document and is comprised by title and full abstract. Each document is represented by a vector of term weights based on the classic vector space model (VSM) (Salton 1989). Each cluster of related documents contains a centroid vector which we call the cluster's representative. To compute document vectors we first preprocess the documents to transform them into a form more amenable to vector space analysis. In the preprocessing stage we remove stopwords from documents and stem the remaining words with the efficient regular expression-based Porter stemming algorithm (Porter 1980). Stopwords are frequently occurring function words in a language which have important grammatical roles but carry no meaning, such as prepositions and articles (e.g., a, an, the, of, for, and, or). Stemming is similar in spirit to finding the root forms of words. For example, a stemmed term index might contain only the root form "walk" to represent "walk," "walked," and "walking." Stopword removal and stemming measures reduce the total number of words in the corpus, which helps to reduce the dimensionality of the document vector space and improve efficiency.

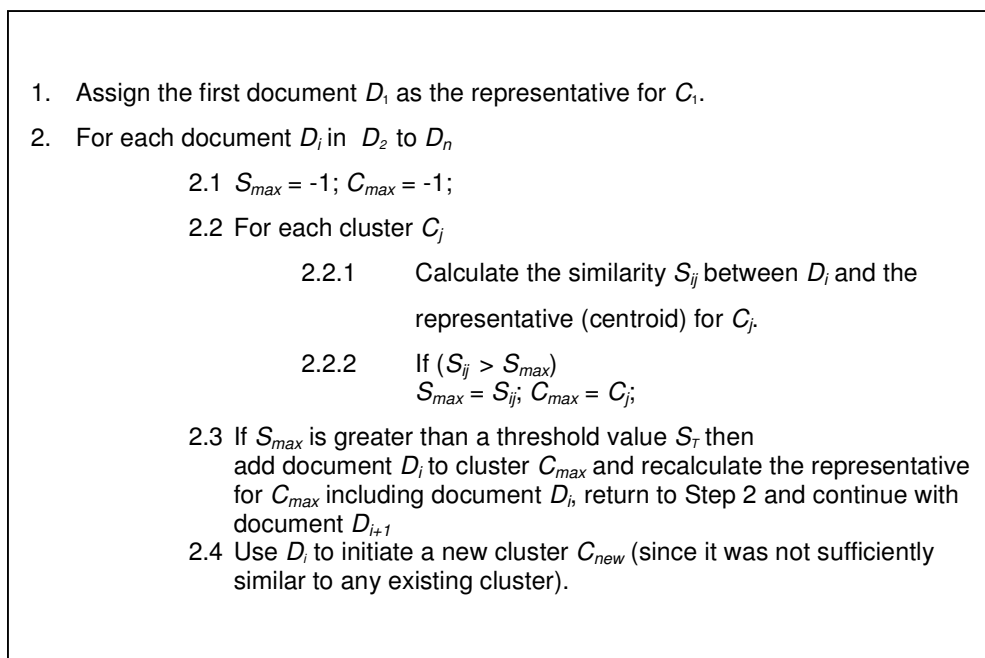


Figure 2. Single Pass Clustering

Once the documents are preprocessed, term weight vectors are computed for each document. The weight w_{t_i, d_j} of a term t_i in a document d_j is calculated by equation (1). TF_{t_i, d_j} (term frequency) is the count of t_i 's occurrence in document d_j . DF_{t_i} (document frequency) is the number of documents in

which the term t_i occurs. N is the total number of documents in the collection. This term weight formula captures both the importance of a term in characterizing a document, and the degree to which a term discriminates between documents. Terms that occur frequently in a given document are assigned higher weights, and terms which occur in many documents receive lower weights to reflect the fact that the terms do not distinguish well between documents.

$$W_{t_i,d_j} = (\log(TF_{t_i,d_j}) + 1) \times \log\left(\frac{N}{DF_{t_i}}\right) \quad (1)$$

Representing documents as vectors in a Euclidean space allows us to use distance metrics from linear algebra to compute a similarity measure between documents (Manning and Schütze, 1999; Rasmussen, 1992).

Our clustering algorithm, as most text retrieval algorithms, uses cosine similarity, which bases similarity on the angle between document vectors. The smaller the angle between two document vectors, the higher their similarity. Vectors for similar documents will be “near” each other in the vector space for some definition of “near”. Euclidean distance is the most obvious distance metric, but its sensitivity to document length requires document normalization. For example, two documents about the same topic but with different lengths would have a high Euclidean distance despite their semantic similarity. To calculate Cosine similarity measure, we hence converted them in unit vectors firstly, to avoid the need of normalization.

Once all the documents are represented as vectors in a term-weight vector space, our clustering algorithm can compute clusters for the document corpus. Our clustering algorithm, outlined in Figure 1, assigns each document to exactly one cluster. The clustering algorithm is parameterized by a similarity threshold S_r affecting the number of clusters computed for a given corpus, which was set to be 0.08. For example, a high similarity threshold requiring high similarity for documents within a cluster will result in a greater number of clusters and vice versa. The algorithm computes an appropriate number of clusters with respect to the similarity threshold, thus avoiding the need to compute or estimate the number of clusters a priori as in traditional K-means clustering. The Single Pass Clustering assigns a new document to the cluster with which it has the maximal similarity and the similarity is above a predefined threshold.

Because our clustering algorithm computes clusters automatically based only on the textual content of documents, clusters do not correspond perfectly to human-derived categories such as those determined by the American Physical Society. Our clusters thus are free to capture other sources of similarity such as research methodology, experimental instrumentation, and other ways in which research can be considered similar. As a consequence, while a scientist might stay within a particular subfield of Physics, the clusters to which his documents are assigned may reflect, for example, variations in the scientist’s methodology over time. Conversely, a scientist may change subfields but retain similar methods, resulting in a low variability with respect to our automatically computed clusters. In summary, we let the documents speak for themselves in our text analysis.

Results

The clustering algorithm applied on our database of Physicists’s publications resulted in 660 clusters, ideally corresponding to publications with similar content. Some statistics on the resulting clusters are reported in Table 1.

Table 1. Clusters characteristics. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max	Median
size of cluster (# publications)	660	57.84	117.85	1	1173	13
population of cluster (# different individuals)	660	16.42	25.27	1	257	6
duration of cluster (years)	660	7.77	5.69	1	17	7

Clusters can be more or less populated in terms of number of articles grouped by the algorithm. We call “size” the cluster dimension in terms of publications. There are nearly 58 publications on average per cluster, but the size is highly variable, as shown in Figure 3, reporting the frequency distribution for the sizes of clusters. There is a considerable number of clusters having just one publication (111, equal to 17% of clusters), which correspond to the clusters identified by the software as relatively unrelated to the rest of the publications, and a very long right-tail. The right graph in Figure 3 represents the distribution reduced to the 25th – 75th percentiles (values 2 and 58 respectively).

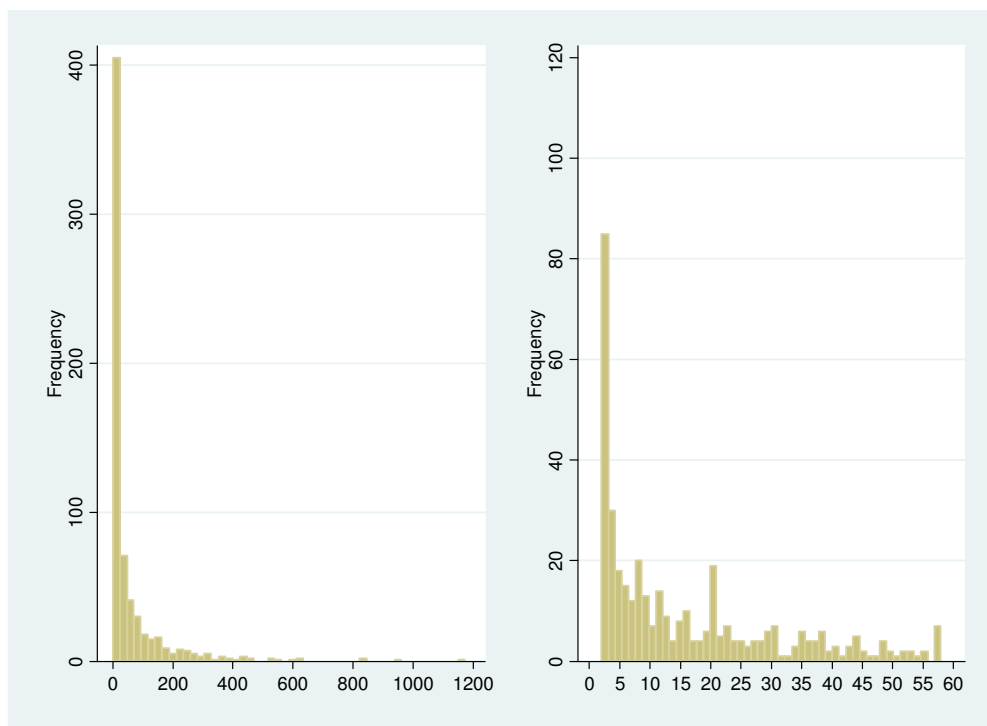


Figure 3: Left: Cluster Size (number of publications per cluster) Frequency Distribution. Right: Cluster Size (number of publications per cluster) Frequency Distribution of 25th - 75th percentile.

Clusters also differ in the number of scientists in our dataset that contributed to it. To account for this differences, we counted the number of database IDs that have at least one publication in a cluster. Because the cluster algorithm works on a unique corpus of text, we expect clusters in which many different scientists have published to be more general in content than clusters more concentrated on a single scientist, holding the number of publications constant. Figure 4 shows the distribution of clusters according to the number of single individuals that contributed to it (restricted to the 95 percentile, equal to 67 IDs per cluster).

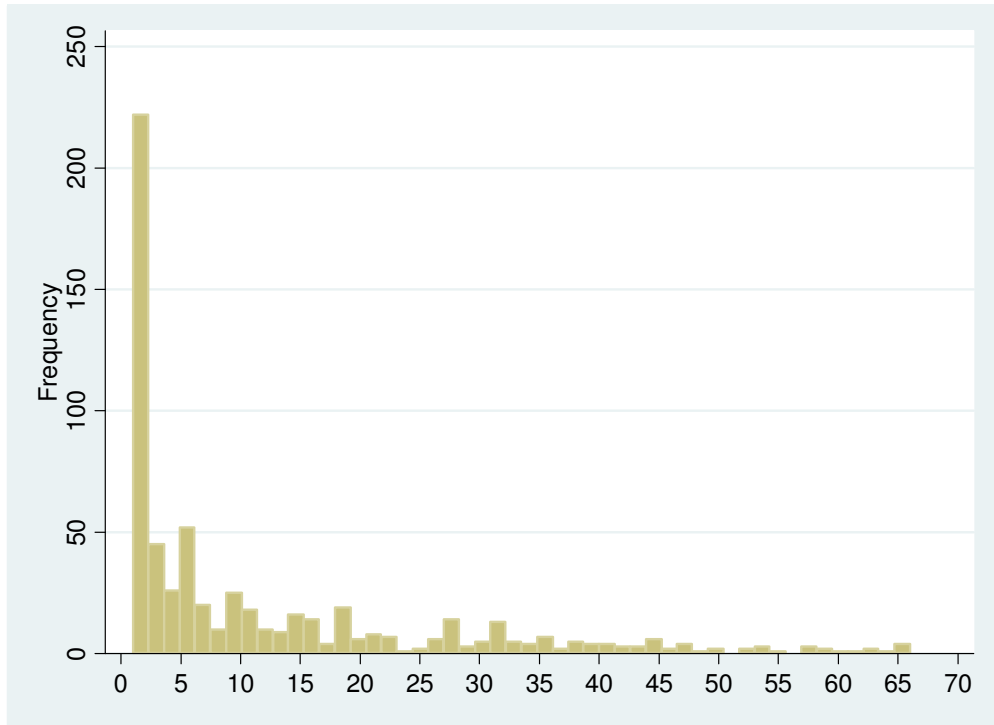


Figure 4: Population of clusters (# of different IDs in a cluster). Frequency Distribution (restricted to 95th percentile)

In terms of duration, clusters differ in the number of years elapsed since the earliest publication to the latest publications grouped in the same cluster.

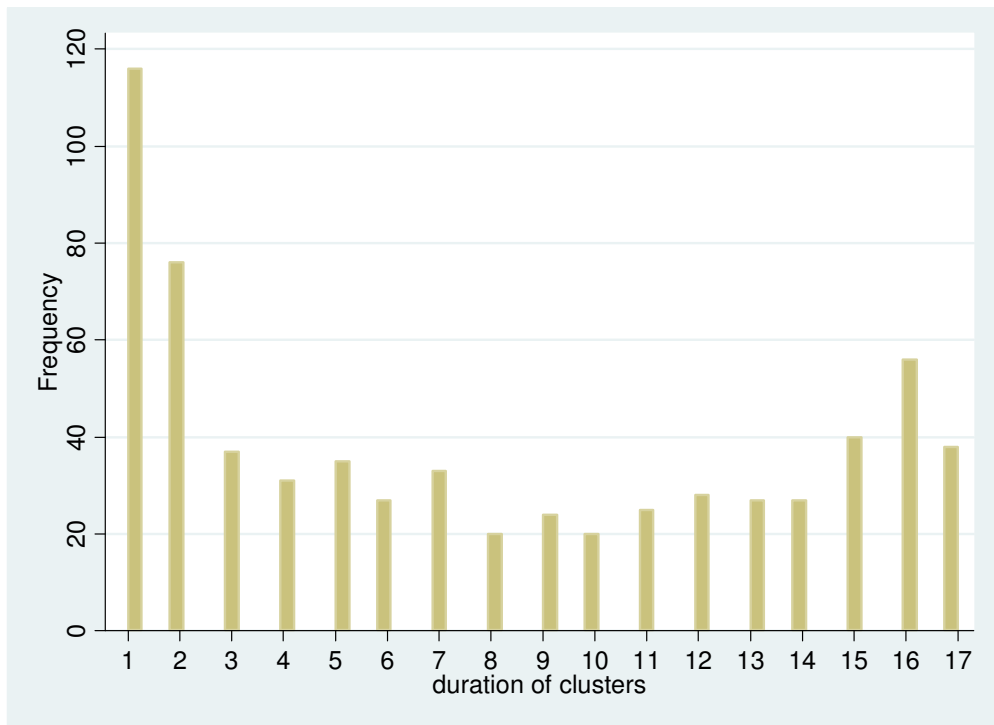


Figure 5: Duration of Clusters (in 17 years). Frequency Distribution.

Our publications were collected during 2006 and dated back to 1990 (first year for which ISI records the full article abstracts). Given the construction of the dataset, the distribution is both left and right censored, in the sense that clusters more active towards the two ends of the observation period have more chances to be not fully observed. Hence cluster duration is here given as a summary statistics to check the consistency of the clustering algorithm, but it should not be taken as an estimate of duration of single topics. The distribution of clusters duration is shown in Figure 5.

The clusters with duration equal to 1 year are heavily inflated by the 111 clusters having just one publication in it. Nonetheless, there is also a high proportion of clusters having a longer duration (16 or 17 years), indicating that there are topics which have a longer publication-cycle than our time-span.

Having presented statistics on the results generated by the clustering algorithm, we now show a possible use of the cluster data to obtain information on single individuals' scientific productivity over time. We constructed three indicators as follows:

1. **Diversification of interest:** number of clusters in which a scientist has at least one publication across the time-span.
2. **Intensity of interest:** number of papers a scientist makes on average in each single cluster (calculated for IDs whose diversification \geq 1).
3. **Duration of interest:** number of years a scientist stays on average in each single cluster (calculated for IDs whose diversification \geq 1)..

The indicators of Intensity and Duration of interest is calculated only for the 603 (93%) scientists that had at least one publication (and one cluster).

Table 2 shows general statistics for the three indicators and on the total number of publications stored in the database. All indicators are calculated over the 17 years time-span.

Table 2. Diversification, Intensity and Duration of Interest. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max	Median
Productivity (# papers)	650	67.93	73.12	0	578	48
Diversification of interest (# clusters)	650	16.22	14.54	0	118	13
Intensity of interest (# papers per cluster)	603	4.73	6.38	1	82.57	3.37
Duration of interest (# years)	603	3.41	1.45	0	13	3.23

The number of different clusters in which a scientist has made at least one contribution gives a measure of the ranges of interests he had during the years, a reason why we call this measure "Diversification" of interests. The higher the number, the wider the spectrum of different lines of research and topics on which he/she was active. Figure 6 shows the distribution of our diversification indicator. The average scientist in our dataset had a little more than 16 different research lines during 17 years (the median scientist 13): nearly one new line per year. This is overall plausible, and in line with the fact that the scientists in our database are certainly head of big university labs and typically supervise (and co-author) the work of several post-doc and junior scientists working on different research lines simultaneously.

On average the scientists published almost 5 papers on each of their different research clusters, although this measure is highly variable per cluster and per scientist, going from 1 to nearly 83 paper per cluster. The frequency distribution of the indicator is shown in Figure 7.

Finally, the frequency distribution of the duration of the interests is shown in Figure 8. The average duration is 3.41 years per single cluster and is distributed almost normally.

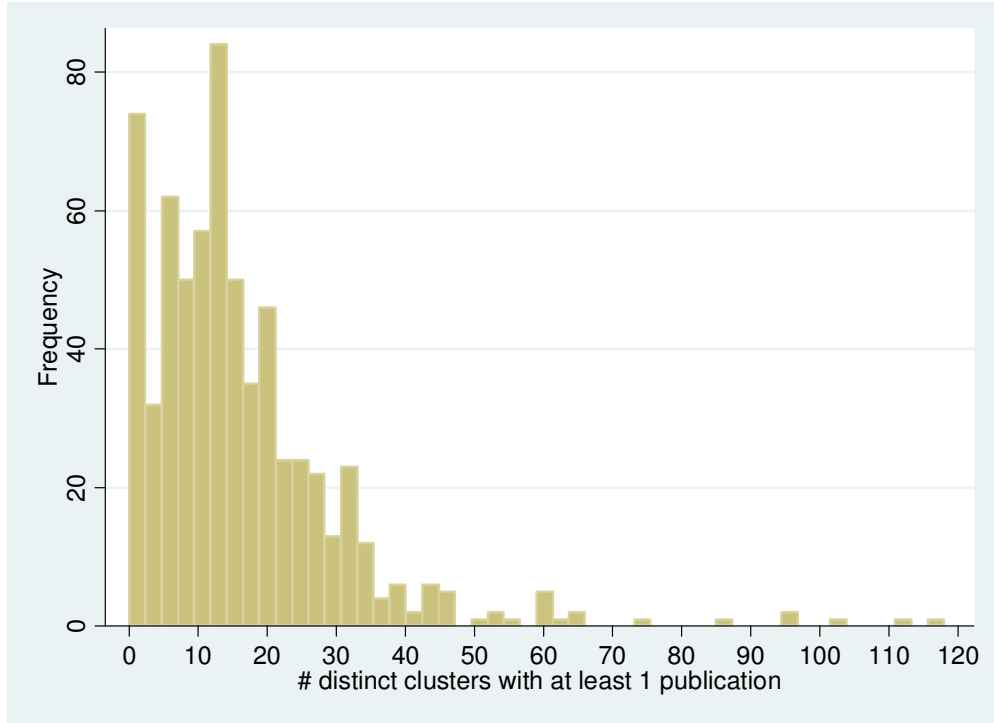


Figure 6: Diversification of interest. Frequency Distribution

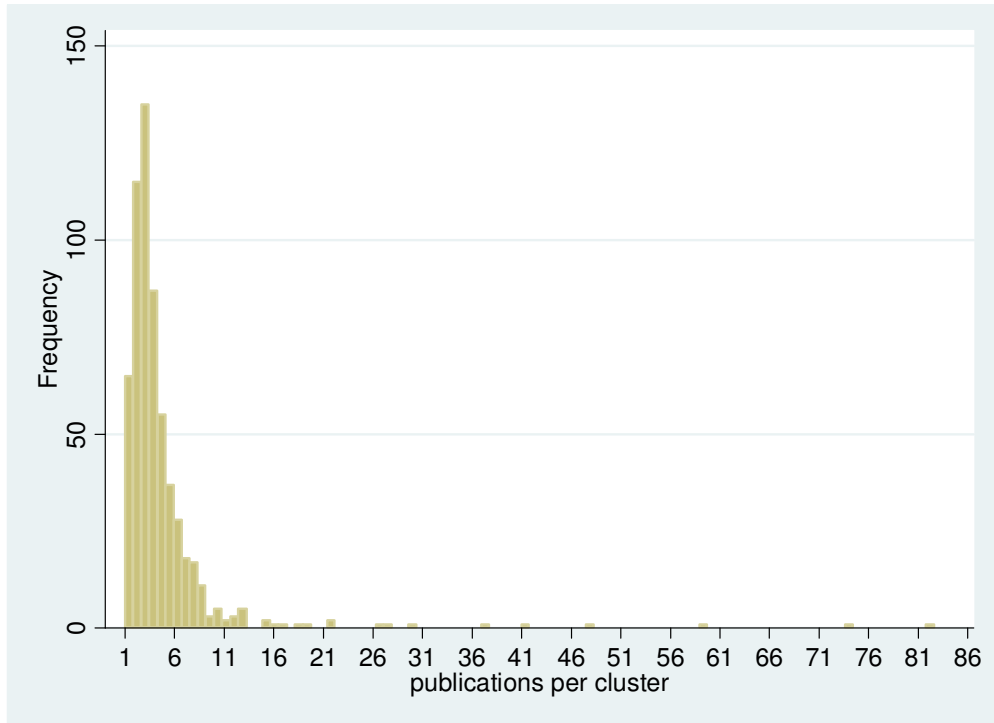


Figure 7: Intensity of Interest. Frequency Distribution

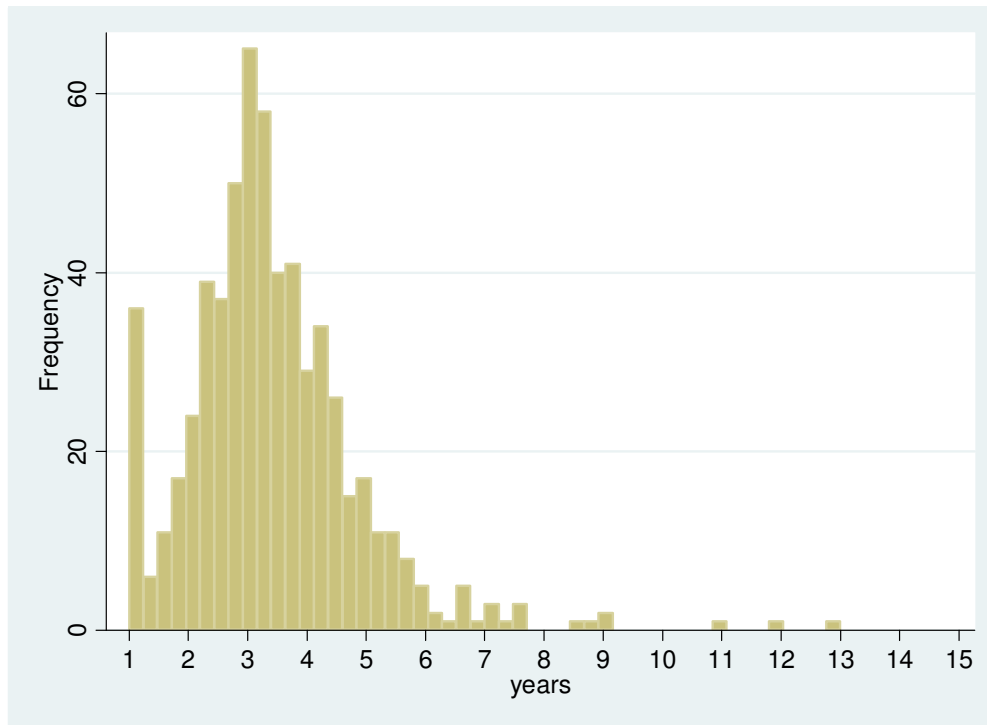


Figure 8: Duration of Interest (during 17 years). Frequency Distribution

Conclusions

We presented a methodology of Software-Assisted Content Analysis that is capable of extracting information on the characteristics of a scientist’s productivity in terms of diversification of interests, intensity and duration over time, by clustering the title and abstracts of scientific publications.

Our approach is novel to the extent that clustering algorithms are here used to extract information on single scientists, rather than mapping overall scientific fields. Our results produce information for single IDs over time and are hence stored in panel format.

We expect this methodology to be useful for a number of applications ranging from the studies of the development of the knowledge needed to cultivate Science and Innovation Policy, to applications for librarians and for granting agencies that may obtain timely biographic information upon a scientist’s work for large masses of data, without a subjective evaluation.

We apply the methodology on an original sample of publications, based on a dataset of 650 American star-scientists in the field of Physics for which all SCI publications were collected from 1990 to the beginning of 2006. Scientists in our sample are highly recognized by their scientific community and we expect them to be head of big university labs, supervising the work of several junior researchers. Our results estimated that scientist were working on average on 16 different research lines during the time-span, publishing nearly 5 articles per cluster, and that research lines were active on average for 3,4 years.

References

- Allison P.D.; Stewart J.A. (1974), “Productivity Differences Among Scientists: Evidence for Accumulative Advantage” *American Sociological Review*, 39(4), 596-606.
- Cole S., Cole J.R: (1967), “Scientific Output and Recognition: A study in the Operation of the Reward System in Science”, *American Sociological Review*, 32(3):377-390.
- Courtial, J. P., Sigogneau, A., and Callon, M. (1997), “Identifying strategic sciences and technologies through scientometrics”, in W. B. Ahston and R. A. Klavans (eds.), “Keeping abreast of science and technology, technical intelligence for business”, Columbus, OH: Battelle Press.

- Crane D (1972), "Invisible Colleges. Diffusion of knowledge in scientific communities", Chicago & London: The University of Chicago Press.
- Fox M.F. (1983), "Publication Productivity among Scientists: A critical Review", *Social Studies of Science*, 13(2), 285-305.
- Garfield, E. (1979), "Citation Indexing: Its theory and application in science, technology and humanities", New York, John Wiley.
- Garner C.A. (1979), "Academic Publication, Market Signaling and Scientific Research Decisions", *Economic Inquiry*, 17(4):575-584.
- Gibbons M., Limoges C., Nowotny H., Schwartzman S., Scott P., Trow M. (1994), "The new production of knowledge. The Dynamics of Science and Research in Contemporary Society", Sage Publications.
- Godin B. (2003), "The emergence of S&T indicators: why did governments supplement statistics with indicators?", *Research Policy*, 32, 679-691.
- Hackett E.J., Conz D., Parker J., Bashford J., DeLay S. (2004), "Tokamaks and turbulence: research ensembles, policy and technoscientific work", *Research Policy* 33(5):747-767.
- Hackett E.J (2005), "Essential Tensions: Identity, Control, and Risk in Research, *Social Studies of Science*, 35(5):787-826.
- Hagstrom W.O. (1965), "The Scientific Community", Basic Books Inc., New York, London.
- Hicks D., Hamilton K. (1999), "Does University-Industry Collaboration Adversely Affect University Research?", *Issues in Science and Technology Online*, Summer 1999, 74-75, http://www.nap.edu/issues/15.4/images/realnum_big.jpg.
- Klavans R., Boyack, K.W. (2005), "Generation of large-scale maps of science and associated indicators", SANDIA Report SAND2005-7538, 01 Dec 2005.
- Leydesdorff L. (2002), "Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports, *Scientometrics*, 53(1):131-159.
- Manning C., and Schütze H. (1999), "Foundations of Statistical Natural Language Processing", Cambridge: MIT Press.
- Merton R. K. (1968), "The Matthew Effect in Science", *Science*, New Series, 159(3810):56-63.
- Moed H.F., Burger W.J.M., Frankfort J.G., Van Raan A.F.J. (1985), "The use of bibliometric data for the measurement of university research performance", *Research Policy*, 14, 131-149.
- Narin F., Hamilton K.S. (1996), "Bibliometric performance measures", *Scientometrics*, 36(3):293-310.
- Peritz, B.C. (1992), "On the Objectives of Citation Analysis: Problems of Theory and Method", *Journal of the American Society for Information Science*, 43(6):448-451.
- Porter M. F. (1980), "An algorithm for suffix stripping", *Program* 14(3):130-137.
- Rasmussen E. (1992), "Clustering Algorithms", in: Frakes N. and Baeza-Yates (eds.), "Information Retrieval: Data Structures & Algorithms", New Jersey: Prentice Hall.
- Salton G. (1989), "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", Reading, MA: Addison Wesley.
- Stephan P.E., Levin S.G. (1992), "Striking the Mother Lode in Science: The Importance of Age, Place, and Time", Oxford University Press.
- Ziman J.M. (1968), "Public knowledge: an essay concerning the social dimension of science, London, Cambridge U.P.
- Courtial, J. P., Sigogneau, A., and Callon, M. (1997), "Identifying strategic sciences and technologies through scientometrics", in W. B. Ahston and R. A. Klavans (eds.), "Keeping abreast of science and technology, technical intelligence for business", Columbus, OH: Battelle Press.