A cluster analysis of scholar and journal bibliometric indicators

Massimo Franceschet

Department of Mathematics and Computer Science – University of Udine
Via delle Scienze, 206 – 33100 Udine (Italy)
massimo.franceschet@dimi.uniud.it

Abstract. We investigate different approaches based on correlation analysis to reduce the complexity of a space of quantitative indicators for the assessment of research performance. The proposed methods group bibliometric indicators into clusters of highly inter-correlated indicators. Each cluster is then associated with a representative indicator. The set of all representatives corresponds to a base of orthogonal metrics capturing independent aspects of research performance and can be exploited to design a composite performance indicator. We apply the devised methodology to isolate orthogonal performance metrics for scholars and journals in the field of computer science and to design a global performance indicator. The methodology is general and can be exploited to design composite indicators that are based on a set of possibly overlapping criteria.[1]

1 Introduction

There is a general agreement in bibliometrics that research quality is not characterised by a single element of performance [van Raan, 2006b]. Two potential dangers of condensing down quality of research to a single metric are: (i) a person may be damaged by the use of a simple index in a decision-making process if the index fails to capture important and different aspects of research performance, and (ii) scientists may focus on maximizing that particular indicator to the detriment of doing more justifiable work.

Several performance indicators have been proposed to assess the quality of research of scholars and journals. In particular, metrics used for scholars can be classified as follows:

- *productivity metrics*, including number of (cited) papers, number of papers per academic year, number of papers per individual author;
- *impact metrics*, comprising total number of citations, number of citations per academic year, number of citations per individual author;
- *hybrid metrics*, consisting of average number of citations per paper as well as h index [Hirsch, 2005] and its variants: the m quotient [Hirsch, 2005], the g index [Egghe, 2006], the contemporary h index [Katsaros et al., 2006], and the individual h index [Batista et al., 2006]. These indicators aim to capture both productivity and impact in a single figure.

Proposals to assess journal performance comprise the impact factor (roughly, the number of recent citations per paper) [Garfield, 1979], h-type indexes for journals [Braun et al., 2006], and prestige-oriented metrics [Bollen et al., 2006].

While it is wise to use a variety of metrics, it is unsystematic and confusing to have too many of them, in particular when metrics are highly correlated. We have a strong feeling that the (ever growing) space

[1] Accepted for publication in Journal of the American Society for Information Science and Technology 60(10), 1950-1964, 2009.

of bibliometric indicators can be partitioned into clusters of indicators represented by orthogonal metrics. Each cluster contains indicators that are mutually strongly correlated and hence it may be represented by a single representative metric. Different clusters (representatives) capture independent aspects of research performance. Ideally, research performance should be characterized by different orthogonal aspects and each performance aspect should be clearly captured by a single performance metric.

Our paper is a contribution toward this ideal situation. We investigate three methods based on correlation analysis to group bibliometric indicators into sets of pairwise strongly correlated indicators; we call these sets *correlation clusters*. The first method is based on the analysis of the structural properties of the correlation graph associated with the correlation matrix of the involved bibliometric variables. The second method uses hierarchical agglomerative clustering techniques. Here, we devise three different strategies to merge sets of indicators during the clustering process. The last method exploits principal component analysis, a technique used in statistics to reduce a multi-dimensional data space to a lower dimension. Techniques to elect a representative indicator of a correlation cluster are suggested. The set of representative indicators forms a reduced *base of orthogonal indicators* capturing the information of the entire bibliometric space. We take advantage of the clustering structure to design a global performance indicator that considers independent aspects of performance.

We apply the clustering methods to computer science literature. We analyse both scholar and journal publications as stored in Google Scholar and Web of Science. We use two data samples: (i) the most brilliant computer science scholars working at the Department of Mathematics and Computer Science of Udine in Italy, (ii) the top-20 computer science journals in subject category *theory and methods* according to the number of received citations as recorded in Web of Science database. We perform a correlation analysis using 13 bibliometric indicators for the first sample and 6 meaningful metrics for the second sample.

The rest of the paper is organized as follows. Section 2 discusses and compares related work. Section 3 illustrates the methods that we propose to compute correlation clusters of bibliometric indicators. In Section 4 we test the goodness of the proposed methods by analysing both scholar and journal publications as recorded in Google Scholar and Web of Science (Subsections from 4.1 to 4.3). Subsection 4.4 uses the outcomes of our study to build a global performance indicator. Section 5 concludes the paper.

2 Related literature

We have found five studies that focus on the clustering of bibliometric indicators on the basis of the statistical correlation among indicators[2].

Costas and Bordons apply principal component analysis to bibliometric indicators at the level of individual scholars [Costas and Bordons, 2007]. The authors study the publications of scientists at the Spanish Research Council in the area of Natural Resources as recorded in Web of Science. The analysis includes productivity indicators, like number of published papers, observed impact indicators, like total number of citations and average number of citations per paper, expected impact indicators, based on the impact factor of journals in which the scholar has published, and the h index. They found that the first factor is associated with number of documents, number of citations, and h index, the second has high loadings on relative citation rates, that are individual citation measures as compared with those of the publication journal, the third factor groups average number of citations per paper and percentage of highly cited

[2] One of them – Leydesdorff's paper – was pointed out by an anonymous referee. All these studies are very recent; in particular, two of them were published after the submission of this paper and one is unpublished at the writing moment and available as a pre-print.

papers, and the last factor comprises expected impact indicators. The authors claim that the h index is a size-dependent indicator, as previously noticed by van Raan [van Raan, 2006a].

Bornmann et al. compare nine variants of the h index, including m quotient, g index, h(2) index, a index, m index, r index, ar index and hw index, using a dataset of biomedicine from Web of Science [Bornmann et al., 2008]. The results of the principal component analysis indicate two types of indexes: the first type (h, g, h(2) indexes and m quotient) describes the quantity of the most productive core of the output of a scientist, the second (a, m, r, ar, and hw indexes) depicts the impact of the papers in the core.

Hendrix uses principal component analysis to compare indicators at the institutional level [Hendrix, 2008]. The author analyses information about publications and citations contained in Web of Science, about funding provided by the National Institutes of Health, and about faculty size for schools that are members of the Association of American Medical Colleges. The author clusters the variables in three distinct groups: the first cluster refers to gross research productivity and comprises total number of papers, total number of citations, and total number of faculty. The second factor reflects the impact of research and includes average number of citations per article, impact index (h index divided by a factor of the number of published articles), and number of uncited papers. The third group describes research productivity and impact at individual level, like number of papers/citations per faculty member. Funding demonstrates a significant relationship with all three factors.

Leydesdorff adopts principal component analysis to cluster journal indicators [Leydesdorff, 2009]. The analysis comprises total number of papers, total number of citations, number of self-citations, total number of cited papers, h index, journal PageRank, impact factor (two and five years), Scimago journal rank, immediacy index and cited half-life. Moreover, the study includes social network centrality measures like degree, betweeness, and closeness centrality both in the cited and citing dimensions. The dataset was collected from both Web of Science and Scopus. Size-dependent measures including total papers and total cites are clustered in the first factor. Size-independent measures like impact factor, Scimago journal rank as well as immediacy index group in the second component. The h index loads on both the first and second factors, but mainly on the second one. Social network centrality indicators, including the network-based indicator PageRank, mainly populate the third component, with the exception of betweeness cited; finally, cited half-life is isolated.

Finally, Bollen et al. study the aggregation of various journal indicators including bibliometric and social network indexes computed on both citation and usage networks [Bollen et al., 2009]. The citation network is based on Web of Science data, whereas the usage graph is constructed from usage log data available at web portals of scientific publishers and institutional library services. Principal component analysis as well as hierarchical and repeated partition (K-means) clustering techniques are used, leading to similar conclusions. Usage-based measures are close together; impact factor is among others surrounded by Scimago cites per paper, Scimago journal rank, and JCR immediacy index; citation degree centrality metrics, total cites measures, and h index are clustered together, and, finally, PageRank and betweeness centrality indicators are mutually correlated. The authors observe that the first principal component separates usage from citation measures, while the second principal component seems to discriminate between popularity-based measures and prestige-based ones.

The present contribution applies both hierarchical clustering and principal component analysis to group bibliometric measures at both scholar and journal levels. It differs from the above mentioned studies for the following reasons:

- at scholar level, our study distinguishes from related ones because we include new indicators. In particular, cited papers, papers per year, papers per author, citations per year, citations per author, contemporary h index, and individual h index are not considered in the above mentioned papers;

– at journal level, we include in the study g index and individual h index that are not comprised in any of the related studies;
– we use data from both Web of Science and Google Scholar. In the concluding section, we highlight the differences we have found in our study between the two data sources. By contrast, no related study uses Google Scholar as a data source;
– besides using principal component analysis, we take advantage of three original hierarchical agglomerative clustering strategies we have devised for the problem at hand. Bollen et al. use clustering techniques as well. However, the two analyses differ: Bollen et al. evaluate the similarity between two metrics by computing the Euclidean distance of the measure correlation vectors, whereas we use the Pearson correlation coefficient as a similarity metric.

The algorithms we devise in this paper belong to the class of hierarchical agglomerative clustering algorithms [Dubes and Jain, 1988]. The main differences between the clustering strategies we propose in this paper and standard hierarchical agglomerative methods are: (i) we use a similarity function based on statistical correlation between variables (bibliometric indicators); (ii) we exploit the peculiarities of such a similarity function in the clustering process in order to define a global measure of correlation for a set of indicators, and (iii) we stop the merging process as soon as it is not possible to proceed unless the objective function value decreases below the given threshold. Hence, we do not necessarily generate the entire clustering dendrogram.

Clustering methods have a long tradition in bibliometrics as tools for grouping bibliometric units (e.g., publications or concepts) on the basis of similarity properties measuring the distance between them [Small, 1973,Kessler, 1963,Callon et al., 1983]. Once the similarity strength between bibliometric units has been established, bibliometric units are typically represented as graph nodes and the similarity relationship between two units is represented as a weighted edge connecting the units, where weights stand for the similarity intensity. Such visualizations are called *bibliometric maps* and resemble the correlation graphs that we exploit in our analysis. At this point, standard hierarchical clustering algorithms can be exploited to find an optimal number of clusters of bibliometric units. When the units are publications or concepts, the identified clusters represent in most cases recognizable research fields [van Raan, 2006b].

3 Methodology

In this section we describe in full detail the proposed methodology, while in Section 4 we apply it to computer science literature.

3.1 Data sources, data samples and bibliometric indicators

We chose two types of actors for our evaluation, scholars and journals, and two data sources, Google Scholar[3] and Thomson Scientific Web of Science[4]. While Web of Science database contains mainly journal publications, Google Scholar finds different types of sources, including journal papers, conference papers, books, theses and reports [Meho and Yang, 2007]. We used the following two data samples:

– 13 computer science scholars of the Department of Mathematics and Computer Science of Udine, Italy. We analysed the papers published until July 2008. We call this sample the scholar sample;

[3] `http://scholar.google.com`
[4] `http://scientific.thomson.com/products/wos/`

– the top-20 computer science journals in subject category *theory and methods* according to the number of received citations recorded in Web of Science database in July 2008. For these journals, we analyzed the published articles during years 2005 and 2006 and the citations they have received until July 2008. We refer to this sample as the journal sample.

We opted for the first sample mainly to address the problem of homonymy for scholars. The sampled computer science researchers work in the department of the writing author. This gave to him the possibility to carefully check the association of bibliographic items with authors either by using his domain knowledge or by directly consulting the local scholars. We computed the following bibliometric indicators on the scholar sample:

1. *papers* (pap). The number of papers.
2. *cited papers* (cp). The number of papers with at least one citation.
3. *papers per year* (ppy). The number of papers divided by the academic age of the scholar.
4. *papers per author* (ppa). The number of papers per author. This is computed by dividing each paper unit by the number of authors of that paper and summing the results over all papers.
5. *citations* (cit). The number of received citations.
6. *cites per year* (cpy). The number of citations divided by the academic age.
7. *cites per author* (cpa). The number of citations per author. This is computed by dividing each citation count of a paper by the number of authors of that paper and summing the results over all papers.
8. *cites per paper* (cpp). The number of citations divided by the number of papers.
9. *h index* (h). The highest number h of papers that have each received at least h citations [Hirsch, 2005].
10. *g index* (g). The highest number g of papers that received together at least g^2 citations [Egghe, 2006].
11. *m quotient* (m). The h index divided by the academic age [Hirsch, 2005].
12. *contemporary h index* (hc). An age-weighted h index obtained by giving more weight to recent papers [Katsaros et al., 2006]. In particular, citations to papers published k years ago are weighted 4/(k+1). The h index is then computed as usual on the weighted citation counts.
13. *individual h index* (hi). The h index divided by the mean number of authors in the set of papers contributing to the h index [Batista et al., 2006].

Our initial experience with citation analysis of journals highlighted two problems. The first one is a limit problem of Google Scholar: it outputs at most 1000 items. On the other hand, most of the journals that we analysed contained more than 1000 articles. A second problem that we encountered is a subject problem: different journals belonging to the computer science ISI categories have strong overlapping with research areas that are far from computer science. As a typical example, Bioinformatics belongs to the following subject categories: (i) biochemical research methods, (ii) biotechnology & applied microbiology, (ii) computer science, interdisciplinary applications, (iv) mathematical & computational biology, and (v) statistics & probability. This is problematic since our goal is to restrict the analysis to (pure) computer science literature. The chosen journal sample solved both the problems. The use of a 2-year temporal interval solved the limit problem. Moreover, journals of computer science in subject category *theory and methods* are pure computer science journals with few significant relationships with other research areas. On this sample we computed the following indicators: total number of papers, total number of citations, average number of citations per paper, h index, g index, individual h index. The indicator cites per paper is strongly correlated to the 2007 impact factor as computed by Thomson Scientific[5].

[5] Pearson correlation 0.91 with p-value $1.435 \cdot 10^{-8}$.

3.2 Data collection and correlation

Data collection, in particular for scholars, was complicated by the well known *name problem*: scholars are usually recorded using initials and surname, e.g., "M. Franceschet", but sometimes the full name is used, e.g., "Massimo Franceschet". Moreover, journals are stored either using the full journal name, like "ACM Transactions on Graphics" or using some abbreviation, like "ACM T Graphic"[6]. Using the abbreviated name for a target increases the probability of homonymy, but using the full name may cut off those bibliographic items that contain only the abbreviated form of the name. To address the name problem for scholars, we decided to sample scholars from the department of the writing author, whose research publications are known to the writing author. Moreover, in order to retrieve journal papers from Google Scholar, we wrote queries containing both the full journal name and its ISO abbreviation. For Web of Science we wrote the full journal name as recorded in Thomson Scientific Journal Citation Report.

Next, we computed values for the chosen bibliometric indicators on the extracted bibliographic data. For Google Scholar, we took advantage of Publish or Perish[7], a free software that queries Google Scholar and computes several citation statistics for scholars and journals. The statistics computed by Thomson Scientific for Web of Science database do not cover all indicators we are using in this paper. Moreover, to the best of our knowledge, there is no automatic tool similar to Publish or Perish for Web of Science. For these reasons, we implemented a software tool that computes all statistics we need, and in particular all indicators computed by Publish or Perish, on the basis of the data given by Web of Science. The outcome of this step is a *bibliometric matrix* for each data source under consideration. Such a matrix contains, for each actor (scholar or journal) under investigation, a row with the values of the different bibliometric indicators for that actor. We represented the bibliometric matrices in XML format. This allowed to query the data using XQuery and transform them into an HTML web page with XSLT [Harold and Means, 2004].

For each bibliometric matrix, we computed a corresponding *correlation matrix*. This is a squared matrix containing the correlation coefficient for each pair of indicators. We used three standard correlation methods: Pearson product moment correlation coefficient, Spearman rank correlation coefficient and Kendall rank correlation coefficient [Moore, 2006]. The last two are non-parametric tests that measure the correlation of the ranks of the samples instead of that of the actual values. By computing the (Pearson) correlation of the three pairs of correlation matrices, we discovered that the three statistical methods are highly correlated on our samples, although Kendall coefficients are lower than the other two. Therefore, in this paper we report on the results according to Pearson correlation only. For statistical computations we took advantage of R, a free software [R Development Core Team, 2007].

3.3 Correlation clusters, partitions, and bibliometric bases

The correlation of a pair of indicators is defined by the correlation coefficient between the indicators in the pair. However, how do we measure the correlation of a set of indicators? Different measures are possible. Let I be a set of bibliometric indicators (that we call the *bibliometric space*) and $R_I = [\rho_{i,j}]$ be its correlation matrix. Given a set $X \subseteq I$ containing $n > 1$ indicators, the *correlation* $cor(X)$ of X can be defined as follows:

- the minimum of the absolute binary correlation coefficients of indicators in X, that is

$$cor(X) = min(X) = min_{\{i,j\} \subseteq X} \{|\rho_{i,j}|\}$$

[6] Each journal has an abbreviation defined by the ISO.
[7] Available at `http://www.harzing.com/pop.htm`

– the average of absolute binary correlation coefficients of indicators in X, that is

$$cor(X) = avg(X) = \frac{\sum_{\{i,j\} \subseteq X} |\rho_{i,j}|}{n \cdot (n-1)/2}$$

– the maximum eigenvalue of the correlation matrix R_X of indicators in X divided by n, that is

$$cor(X) = \frac{max\{\lambda_i\}}{n}$$

The last measure deserves some explanation. It is known that $tr(R_X) = \sum_{i=1}^{n} \lambda_i$, where $\lambda_i \geq 0$ are the eigenvalues of the correlation matrix R_X and $tr(R_X)$ is the trace of R_X, that is, the sum of the elements on the diagonal. Since all elements on the diagonal of a correlation matrix are equal to 1, then $tr(R_X) = \sum_{i=1}^{n} \lambda_i = n$. High values for the defined measure are shown when there is some dominant eigenvalue (close to n) and all other eigenvalues are smaller (close to 0). Low values for the defined measure are obtained when all eigenvalues have similar values (close to 1). By virtue of principal component analysis [Jolliffe, 2002], the former situation characterizes highly correlated variables, while the latter situation denotes uncorrelated variables.

We define the correlation of a singleton (a set with just one element) to be 1. Notice that in any case the correlation is a real number in $[0, 1]$ that measures how strongly the indicators of the set are inter-correlated. Values close to 1 indicate (either positive or negative) high correlation, while values close to 0 represent low correlation. In the experiments presented in Section 4, we adopt the definition of set correlation based on the average.

A *correlation cluster* is a set of indicators with a significant correlation. It contains indicators that are mutually strongly correlated. We can consider a correlation cluster as a macro indicator. That is, we can elect a *representative indicator* for the cluster. Given a cluster X, a simple method to choose a representative of X is to select the indicator i with the maximum average pairwise correlation between i and the other indicators in X. A more involved method is to use principal component analysis [Jolliffe, 2002] for each cluster of indicators and to use the first principal component as representative.

At the opposite side of correlation clusters we have *isolated indicators*. An indicator i is said to be isolated if the average pairwise correlation between i and the other indicators in the bibliometric space is low. It is unlikely that an isolated indicator belongs to a correlation cluster with other indicators.

A *partition* $\mathcal{P}$ of I is a set of subsets of I such that: (i) each subset is not empty; (ii) subsets are pairwise disjoint; (iii) the union of all subsets is I. Two trivial partitions are the one containing only singletons (the *singleton partition*) and the one containing the only set I (the *universal partition*). The cardinality $|\mathcal{P}|$ of a partition $\mathcal{P}$ is the number of subsets it contains. A partition $\mathcal{P}$ is smaller than another partition $\mathcal{Q}$ if the cardinality of $\mathcal{P}$ is smaller than the cardinality of $\mathcal{Q}$. The minimum correlation of a partition is the minimum of correlations of sets in the partition, that is, $min(\mathcal{P}) = min_{X \in \mathcal{P}}\{cor(X)\}$. The average correlation of a partition is the average correlation of sets in the partition, that is, $avg(\mathcal{P}) = \sum_{X \in \mathcal{P}} cor(X)/|\mathcal{P}|$.

Our goal is to simplify the bibliometric universe by computing a partition of it and then forming a *bibliometric base* containing all representatives of the sets in the partition. What is a good partition of the universe of indicators? How can we *efficiently* find good partitions? We consider the following two desiderata for the goodness of a partition:

– the partition should have *small cardinality*. The smaller is the partition, the bigger is the reduction of the bibliometric space;
– the partition should have *high minimum correlation*. The higher is the correlation of the sets in the partition, the better these sets are captured by representative indicators.

A strategy to find a good partition has to compromise between the two desiderata and can be of two types: (i) the strategy attempts to find the partition with the highest minimum correlation and with cardinality smaller than a given threshold, or (ii) the strategy attempts to find the partition with the smallest cardinality and with minimum correlation bigger than a given threshold. Since we do not know a priori the number of clusters of our partition we aim at strategies that operate in the second way.

3.4 Clustering algorithms

Clustering is the process of organizing objects into groups whose members are similar in some way [Anderberg, 1973,Dubes and Jain, 1988]. A cluster is a collection of objects which are similar between them and are dissimilar to objects belonging to other clusters. In the following we propose a clustering algorithm for bibliometric indicators based on three different merging strategies. Our goal is to find partitions of indicators with small cardinality and minimum correlation greater than or equal to a given threshold. We first propose an algorithmic schema. Let fix a correlation threshold $\alpha \in [0, 1]$:

1. start with the singleton partition;
2. union two sets A and B such that the union $A \cup B$ has correlation bigger than or equal to α. Repeat step 2 until there are no more sets whose union has correlation bigger than or equal to α.

The given algorithmic schema guarantees the following properties that are immediate to verify:

1. the algorithm returns a partition with a minimum correlation of at least α;
2. the smaller is the correlation level α, the smaller is the returned partition;
3. if the task of choosing the pair of sets to union has polynomial complexity, then the algorithm has polynomial complexity as well.

The above algorithmic schema must be instantiated with a method of choice of the sets to merge. We propose the following three different strategies for this task; all of them can be implemented in polynomial time:

Strategy S1. This strategy unions the pair of sets such that the correlation of the union of them is the greatest;
Strategy S2. This strategy merges the pair of sets such that the minimum correlation of the resulting partition is the greatest.
Strategy S3. This strategy joins the pair of sets such that the average correlation of the resulting partition is the greatest.

We have implemented the clustering scheme with the mentioned strategies. The code is available as free software at the author's web page. The three outlined strategies are different, as shown in the following examples:

- S1 is different from S2. Let X, Y, Z be sets such that $cor(X) = cor(Y) = 1$ and $cor(Z) = \frac{1}{2}$. Suppose that $cor(X \cup Y) = 1$ and $cor(X \cup Z) = cor(Y \cup Z) = \frac{3}{4}$. Strategy S1 clearly chooses to union the pair X, Y, however strategy $S2$ picks the pair X, Z (or Y, Z). Indeed, if we union the pair X, Z (or Y, Z), then the resulting partition has minimum correlation $\frac{3}{4}$, and if we union the pair X, Y, then the resulting partition has a smaller minimum correlation of $\frac{1}{2}$;

– S1 is different from S3. Let X, Y, Z be sets such that $cor(X) = cor(Y) = \frac{1}{2}$ and $cor(Z) = 1$. Suppose that $cor(X \cup Y) = \frac{3}{8}$ and $cor(X \cup Z) = cor(Y \cup Z) = \frac{3}{4}$. Strategy S1 clearly opts for the pair X, Z (or Y, Z), however strategy $S3$ chooses the pair X, Y. Indeed, if we union the pair X, Z (or Y, Z), then the resulting partition has average correlation of $\frac{10}{16}$, and if we union the pair X, Y, then the resulting partition has a bigger average correlation of $\frac{11}{16}$;
– S2 is different from S3. Use the previous example. Strategy S2 settles on the pair X, Z (or Y, Z): if we union the pair X, Z (or Y, Z), then the resulting partition has minimum correlation of $\frac{1}{2}$, and if we union the pair X, Y, then the resulting partition has a smaller minimum correlation of $\frac{3}{8}$. However, as shown above, strategy S3 selects the pair X, Y.

A natural question is: do these strategies output the smallest partition among all partitions with a minimum correlation greater or equal to a given threshold? The answer is no, as shown in the following counter-example. Consider the correlation graph depicted in Figure 1 and a correlation threshold of 0.5. Nodes are indicators and edges are labelled with the binary correlation coefficients. If two nodes are not connected by an edge then the correlation for that pair of nodes is 0. Define the correlation of a set of indicators X either as $cor(X) = min(X)$ or as $cor(X) = avg(X)$. It turns out that strategies S1, S2, and S3 return the partition $\{\{A, B\}, \{C, D\}, \{E, F\}\}$. However, the smallest partition with a minimum correlation of 0.5 is $\{\{A, B, C\}, \{D, E, F\}\}$.

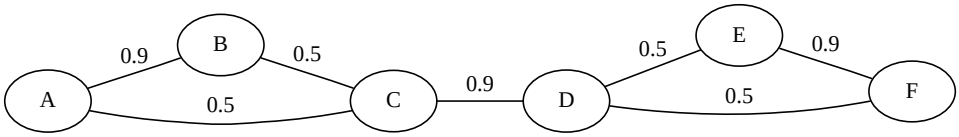

Fig. 1. A counter-example for the coarsest partition problem

Is there any efficient (polynomial) algorithm to find the smallest partition with a given minimum correlation threshold? This is very unlikely. Indeed, let us define a dissimilarity function dis between two bibliometric indicators as an arbitrarily function such that (a) $dis(i, j) \geq 0$, (b) $dis(i, i) = 0$, and (c) $dis(i, j) = dis(j, i)$. Assume that we measure the dissimilarity of a set of bibliometric indicators X either as the maximum dissimilarity among pairs of indicators or as the average dissimilarity among pairs of indicators (the dissimilarity of singletons is 0). Then, we have that the problem of computing the smallest partition with a given maximum dissimilarity threshold is NP-complete. The proof follows from results in [Gonzalez, 1985,Gonzalez and Sahni, 1976], in particular from the fact that P2 form of the clustering problem defined in [Gonzalez, 1985] is NP-hard. NP-completeness for a problem means that, unless the complexity classes P and NP are equal, which is not the current conjecture, there is no polynomial algorithm that solves the problem in the general case [Cormen et al., 2001].

4 A case study for computer science literature

In this section we apply the methodology proposed in Section 3 to computer science literature.

4.1 Computer science scholars on Google Scholar

In this section we make a correlation analysis of the scholar sample on Google Scholar according to the methods outlined in Section 3. The Pearson correlation matrix is given in Table 1. A corresponding correlation graph is depicted in Figure 2; nodes represent indicators and edges correspond to correlations between them. The lightness and thickness of edges is proportional to the correlation strength between the connected variables. For the sake of readability, only correlations greater than or equal to 0.8 are showed.

	h	**g**	**pap**	**cp**	**cit**	**cpp**	**m**	**hc**	**ppy**	**cpy**	**ppa**	**cpa**	**hi**
h	1.00	0.73	0.88	0.87	0.82	0.38	0.19	0.87	0.80	0.81	0.74	0.71	0.66
g	0.73	1.00	0.69	0.73	0.98	0.85	-0.21	0.73	0.36	0.92	0.60	0.91	0.42
pap	0.88	0.69	1.00	0.98	0.77	0.23	-0.08	0.67	0.78	0.74	0.93	0.73	0.67
cp	0.87	0.73	0.98	1.00	0.82	0.30	-0.05	0.68	0.77	0.80	0.92	0.76	0.67
cit	0.82	0.98	0.77	0.82	1.00	0.76	-0.10	0.78	0.49	0.95	0.69	0.91	0.51
cpp	0.38	0.85	0.23	0.30	0.76	1.00	-0.11	0.56	0.01	0.75	0.16	0.70	0.10
m	0.19	-0.21	-0.08	-0.05	-0.10	-0.11	1.00	0.36	0.40	0.15	-0.11	-0.12	0.12
hc	0.87	0.73	0.67	0.68	0.78	0.56	0.36	1.00	0.67	0.84	0.50	0.62	0.41
ppy	0.80	0.36	0.78	0.77	0.49	0.01	0.40	0.67	1.00	0.61	0.75	0.49	0.68
cpy	0.81	0.92	0.74	0.80	0.95	0.75	0.15	0.84	0.61	1.00	0.68	0.90	0.54
ppa	0.74	0.60	0.93	0.92	0.69	0.16	-0.11	0.50	0.75	0.68	1.00	0.78	0.81
cpa	0.71	0.91	0.73	0.76	0.91	0.70	-0.12	0.62	0.49	0.90	0.78	1.00	0.69
hi	0.66	0.42	0.67	0.67	0.51	0.10	0.12	0.41	0.68	0.54	0.81	0.69	1.00

Table 1. Pearson correlation matrix for the scholar sample on Google Scholar

The correlation graph highlights two correlation clusters: a paper-based group containing papers, cited papers, and papers per author, with computed cluster correlation at 0.94 and papers as representative indicator, and a citation-based set comprising g, cites, cites per year, cites per author, with computed cluster correlation at 0.93 and cites as representative indicator (notice that this cluster corresponds to a *clique* in the correlation graph). The m quotient is an isolated indicator, meaning that is not connected with any other indicator in the correlation graph. On the other hand, the h index is the most central indicator, that is, it has the highest number of links to other indicators in the correlation graph (six links).

We now consider the output of the clustering algorithms that we proposed in Section 3. We use (a generalized version of) a dendrogram to visualize the output of our hierarchical clustering algorithms. A *dendrogram* is a forest of binary trees in which nodes are sets of objects (indicators in our case). Leaves corresponds to singletons. Each internal (non-leaf) node has two children that are the sets that have been merged to form that node during the clustering process. The partition of indicators is formed by taking all roots of trees in the dendrogram. We labelled each internal node with the timestamp of the merging operation. This allows to track the evolution of the merging process of the algorithm. The granularity of the partition can be controlled with the correlation threshold parameter. The extreme cases are when the threshold is 1 (the forest contains only singleton trees corresponding to the singleton partition) and 0 (the forest is a unique tree corresponding to the universal partition). In general, the lower is the correlation threshold, the coarser is the resulting partition.

Figure 3 depicts the dendrogram of clustering strategy S1 with a correlation threshold at 0.9. We printed trees bottom-up, from leaves to the root; singleton trees (trees with one node) are not shown

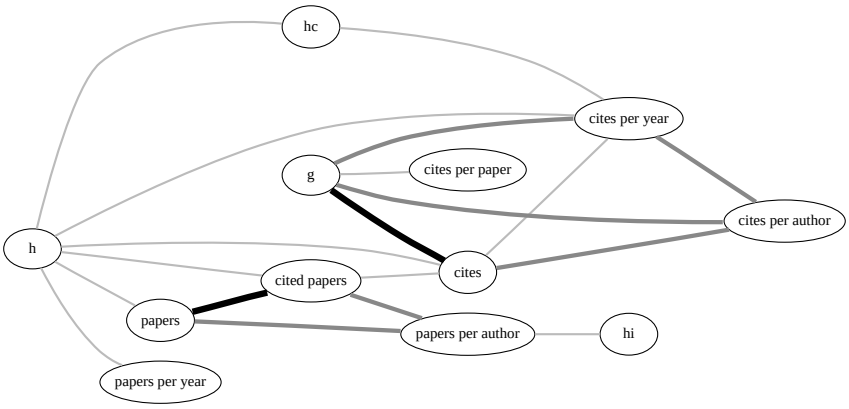

Fig. 2. Correlation graph for the scholar sample on Google Scholar: darker, thicker edges correspond to higher correlations

in the dendrogram. The resulting partition contains the two main clusters identified by the analysis of the correlation graph. Figure 4 shows the dendrogram of clustering strategy S2 with a lower correlation threshold at 0.8; notice that the resulting partition is coarser. Finally, Figure 5 contains the dendrogram of clustering strategy S3 with correlation threshold at 0.8. The resulting partition contains one main cluster containing both paper-based and citation-based metrics. Notice that the merging process of strategy S3 is different from those of strategies S1 and S2. In the clustering approach, isolated indicators correspond to objects that are reluctant to join a cluster. These indicators can be identified by setting a low correlation threshold and searching for indicators that join a cluster with a high timestamp. The quotient m is an example.

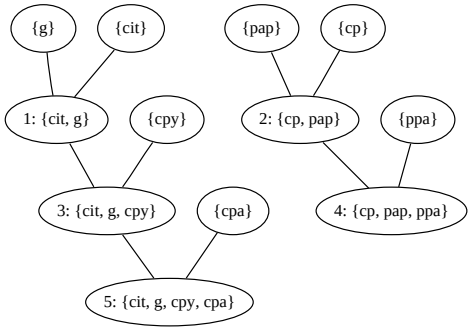

Fig. 3. Dendrogram for clustering strategy S1 (threshold at 0.9)

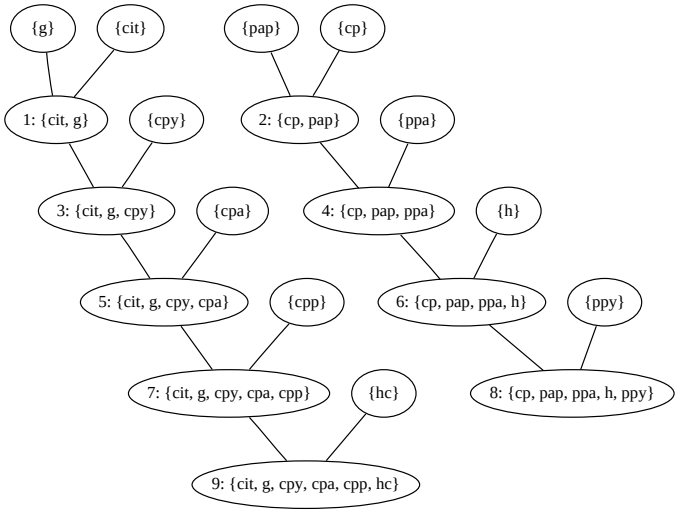

Fig. 4. Dendrogram for clustering strategy S2 (threshold at 0.8)

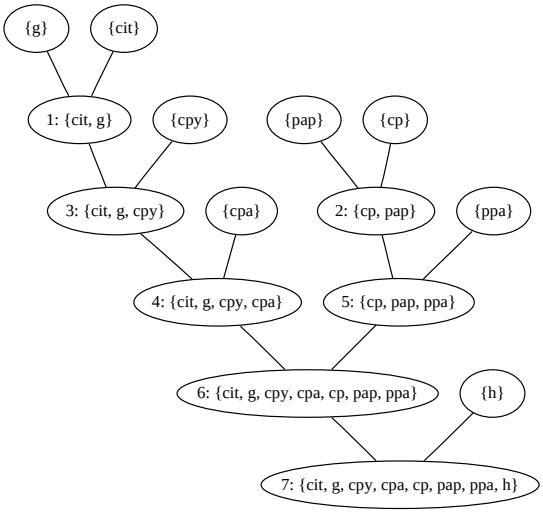

Fig. 5. Dendrogram for clustering strategy S3 (threshold at 0.8)

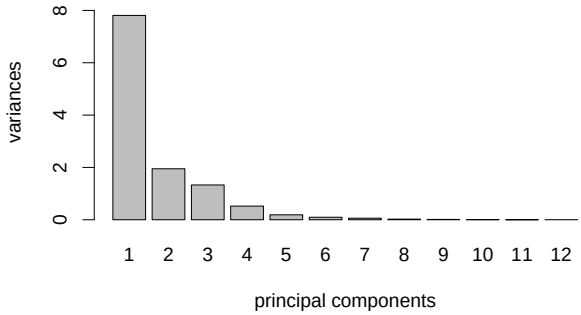

Fig. 6. Screeplot of the variances

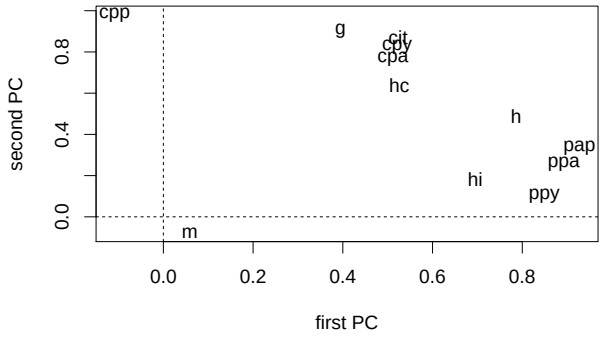

Fig. 7. Scatterplot of scholar indicators on the two principal components

Finally, we apply principal component analysis (PCA) to the bibliometric matrix[8]. PCA is a multivariate statistics technique used to reduce a multi-dimensional space to a lower dimension [Jolliffe, 2002]. Figure 6 shows the screeplot for the variances explained by the principal components. According to Kaiser method (eigenvalues greater than 1), the first three principal components are selected; they explain 92% of the variance, whereas the first two account for 81% of it. The scree method also suggests the use of three principal components – the scree in the screeplot starts just after the third component. Table 2 shows the component loadings for the three selected components, rotated using variance maximizing (varimax) method, and Figure 7 shows the projection of bibliometric indicators on the bi-dimensional plane identified by the first two principal components. Two main groups are clear: one contains the impact measures cites, cites per author, cites per year, g and contemporary h, and the other comprises the productivity measures papers, papers per author, papers per year. Interestingly, h and individual h are between these two clusters, but closer to the productivity group. Cites per paper and m quotient are isolated.

PC	**h**	**g**	**pap**	**cit**	**cpp**	**m**	**hc**	**ppy**	**cpy**	**ppa**	**cpa**	**hi**
PC1	**0.79**	0.39	**0.93**	0.52	-0.11	0.06	0.53	**0.85**	0.52	**0.89**	0.51	**0.70**
PC2	0.49	**0.90**	0.33	**0.84**	**0.98**	-0.07	**0.64**	0.10	**0.82**	0.26	**0.77**	0.18
PC3	0.19	-0.16	-0.11	-0.07	-0.03	**0.99**	0.38	0.36	0.18	-0.14	-0.09	0.10

Table 2. Loadings of the first three principal components (varimax rotation method)

4.2 Computer science scholars on Web of Science

In this section we analyze the correlation for the scholar sample on Web of Science. Table 3 shows the Pearson correlation matrix. The corresponding correlation graph is depicted in Figure 8. The following correlation cluster is evident from the correlation graph: {papers, cited papers, papers per year, papers per author}, with correlation at 0.92 and papers as representative indicator (the cluster is a clique in the correlation graph). Moreover, two additional inter-related clusters can be noticed in the lower part of the graph. They are: {h, g, cites, cites per year, cites per author} and {h, g, cites, hc} both with correlation at 0.87 and cites as representative indicator. Cites per paper and m quotient are isolated metrics, and cites is the most central indicator with five adjacent nodes.

Moving to cluster analysis, Figure 9 shows the dendrogram of clustering strategies S1 and S2, while Figure 10 is the dendrogram of clustering strategy S3. In each case the correlation threshold is set to 0.8. The upper part of the correlation graph, related to productivity metrics, and the lower part, related to impact metrics, are clearly identified in the dendrograms by the left and right trees, respectively. Notice the difference between strategies S1 (or S2) and S3: S1 first discovers the cluster {h, g, cites, hc} and then joins cites per year and cites per author, while S3 works the other way around, first discovering {g, cites, cites per year, cites per author} and then merging h and hc.

Turning to principal component analysis, Figure 11 shows the screeplot for the principal components. Both Kaiser and scree methods agrees on the number of components to select – two components; they express 86% of the total variance. Table 4 contains the varimax rotated component loadings and Figure 12

[8] We removed cited papers from the matrix to avoid system singularity (13 variables, 13 observations). We already know from previous analyses that cited papers is highly correlated to papers.

	h	**g**	**pap**	**cp**	**cit**	**cpp**	**m**	**hc**	**ppy**	**cpy**	**ppa**	**cpa**	**hi**
h	1.00	0.85	0.75	0.76	0.91	0.40	0.66	0.87	0.52	0.82	0.71	0.71	0.60
g	0.85	1.00	0.46	0.45	0.95	0.74	0.53	0.78	0.24	0.89	0.43	0.81	0.70
pap	0.75	0.46	1.00	0.99	0.67	-0.15	0.60	0.70	0.89	0.64	0.96	0.54	0.43
cp	0.76	0.45	0.99	1.00	0.67	-0.16	0.60	0.69	0.88	0.63	0.95	0.53	0.40
cit	0.91	0.95	0.67	0.67	1.00	0.57	0.61	0.84	0.46	0.94	0.65	0.85	0.70
cpp	0.40	0.74	-0.15	-0.16	0.57	1.00	0.22	0.46	-0.29	0.56	-0.15	0.58	0.42
m	0.66	0.53	0.60	0.60	0.61	0.22	1.00	0.64	0.71	0.77	0.54	0.65	0.55
hc	0.87	0.78	0.70	0.69	0.84	0.46	0.64	1.00	0.50	0.79	0.70	0.76	0.61
ppy	0.52	0.24	0.89	0.88	0.46	-0.29	0.71	0.50	1.00	0.55	0.83	0.41	0.32
cpy	0.82	0.89	0.64	0.63	0.94	0.56	0.77	0.79	0.55	1.00	0.62	0.93	0.74
ppa	0.71	0.43	0.96	0.95	0.65	-0.15	0.54	0.70	0.83	0.62	1.00	0.60	0.57
cpa	0.71	0.81	0.54	0.53	0.85	0.58	0.65	0.76	0.41	0.93	0.60	1.00	0.83
hi	0.60	0.70	0.43	0.40	0.70	0.42	0.55	0.61	0.32	0.74	0.57	0.83	1.00

Table 3. Pearson correlation matrix for the scholar sample on Web of Science

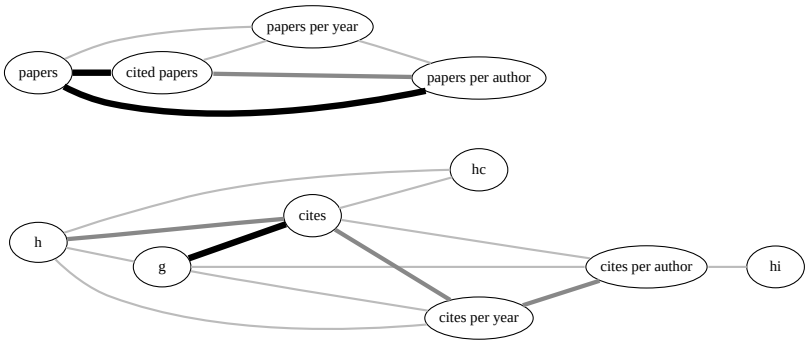

Fig. 8. Correlation graph for the scholar sample on Web of Science

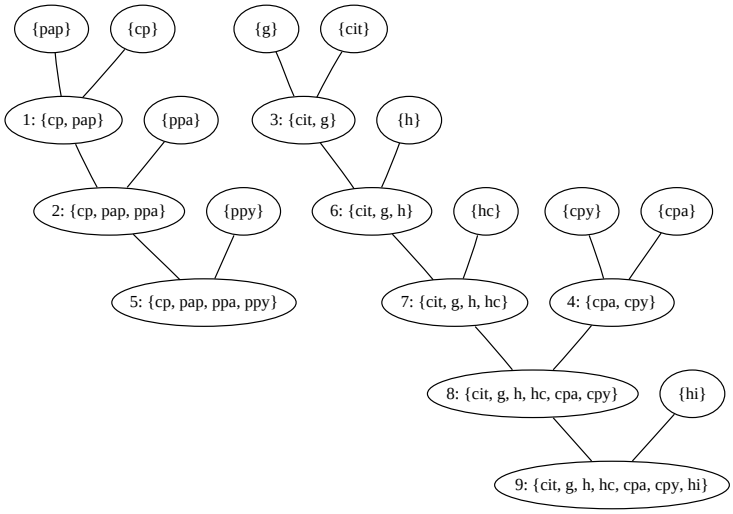

Fig. 9. Dendrogram for clustering strategies S1 and S2 (threshold at 0.8)

shows the scatterplot of scholar indicators using the first two principal components. Comparing with Google Scholar, h and individual h are closer to impact metrics rather than to productivity ones; moreover, m quotient is between productivity and impact indicators and does not define a separate performance dimension. Cites per paper is still in the back of beyond. Costas and Bordons [Costas and Bordons, 2007] also aggregate the h index with citations and separate them from cites per paper.

PC	**h**	**g**	**pap**	**cit**	**cpp**	**m**	**hc**	**ppy**	**cpy**	**ppa**	**cpa**	**hi**
PC1	**0.76**	**0.98**	0.30	**0.90**	**0.83**	**0.47**	**0.71**	0.09	**0.85**	0.29	**0.79**	**0.67**
PC2	0.55	0.17	**0.95**	0.43	-0.42	**0.48**	0.51	**0.91**	0.40	**0.92**	0.32	0.25

Table 4. Loadings of the first two principal components (varimax rotation method)

Finally, for each indicator, we computed the average pairwise correlation between the indicator and the other indicators both on Google Scholar and on Web of Science. The results are shown in Table 5.

The most inter-correlated indicators for both data sources are cites per year, cites and h. Isolated indicators are m quotient and cites per paper in case of Google Scholar, and cites per paper in case of Web of Science. These results confirm the analysis made on the structural properties of the correlation graph. The only significant difference between the two data sources is for m quotient – 0.04 on Google Scholar and 0.59 on Web of Science– as already observed using principal component analysis.

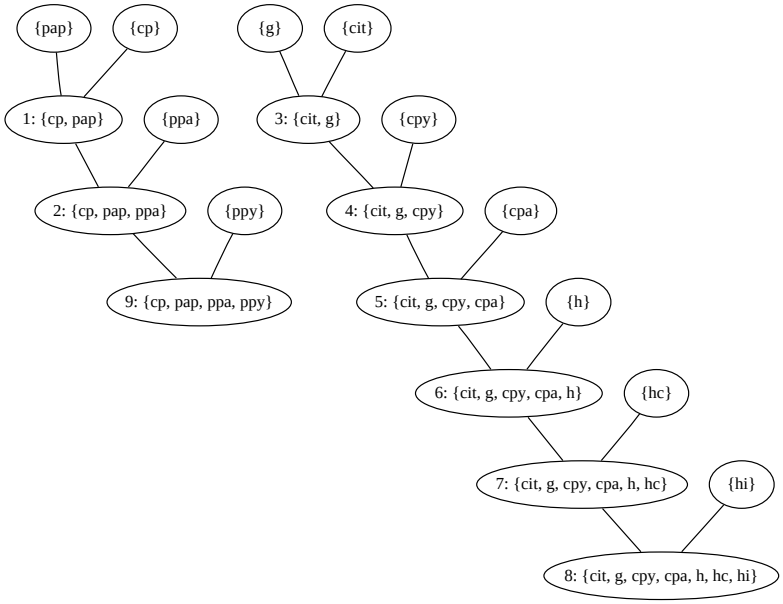

Fig. 10. Dendrogram for clustering strategy S3 (threshold at 0.8)

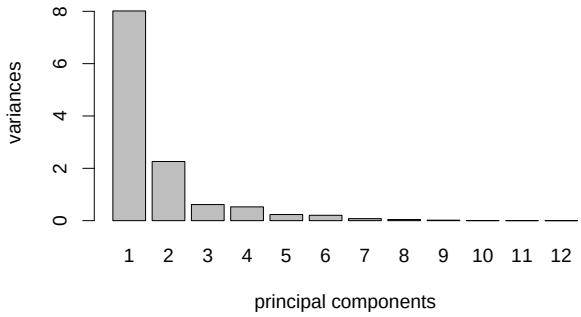

Fig. 11. Screeplot of the variances of principal components

source	h	g	pap	cp	cit	cpp	m	hc	ppy	cpy	ppa	cpa	hi
Google Scholar	0.71	0.64	0.67	0.69	0.70	0.39	0.04	0.64	0.57	0.72	0.62	0.67	0.52
Web of Science	0.71	0.65	0.62	0.61	0.73	0.26	0.59	0.69	0.50	0.74	0.62	0.68	0.57

Table 5. Average correlation for scholar indicators

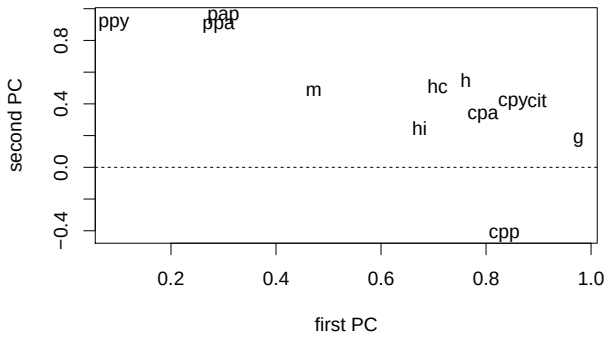

Fig. 12. Scatterplot of scholar indicators on the two principal components

4.3 Computer science journals

In this section we study and compare the correlation for the journal sample on both Google Scholar and Web of Science.

We start with the analysis of Google Scholar. Table 6 contains the Pearson correlation matrix. The corresponding correlation graph is depicted in Figure 13. Indicators cites, h, g, and hi are strongly correlated with each other, with a correlation at 0.91. Papers is moderately correlated with cites (0.79), but it shows lower correlation with respect to other measures. Cites per paper is isolated. The main correlation cluster is immediately discovered by the clustering procedures we have devised using any of the three joining strategies.

	pap	**cit**	**cpp**	**h**	**g**	**hi**
pap	1.00	0.79	-0.33	0.60	0.41	0.52
cit	0.79	1.00	0.12	0.93	0.86	0.83
cpp	-0.33	0.12	1.00	0.36	0.49	0.49
h	0.60	0.93	0.36	1.00	0.95	0.94
g	0.41	0.86	0.49	0.95	1.00	0.92
hi	0.52	0.83	0.49	0.94	0.92	1.00

Table 6. Pearson correlation matrix for the journal sample on Google Scholar

The results for principal components analysis are as follows. Figure 14 shows the screeplot for the principal components: the first component explains 70% of the variance and the first two components account for 95% of the variance. Table 7 shows the component loadings for the first two principal components and Figure 15 shows the scatterplot of journal indicators on the first two principal components. Indicators h, g, cites and hi are close together, whereas papers and cites per paper are far from this group and from each other. Notice that while papers significantly loads on both components, cites per paper mainly loads only on the second one. This difference keeps apart the two indicators in the scatterplot in Figure 15.

Turning to Web of Science, Table 8 shows the Pearson correlation matrix. The corresponding correlation graph is given in Figure 16. As found for Google Scholar, indicators cites, h, g, and hi form a

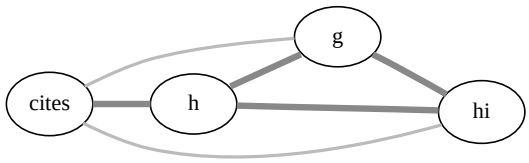

Fig. 13. Correlation graph for the journal sample on Google Scholar

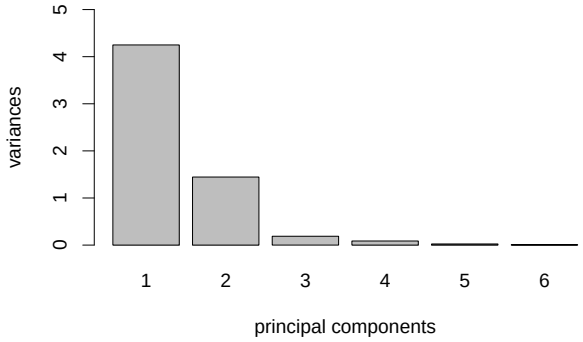

Fig. 14. Screeplot of the variances of principal components

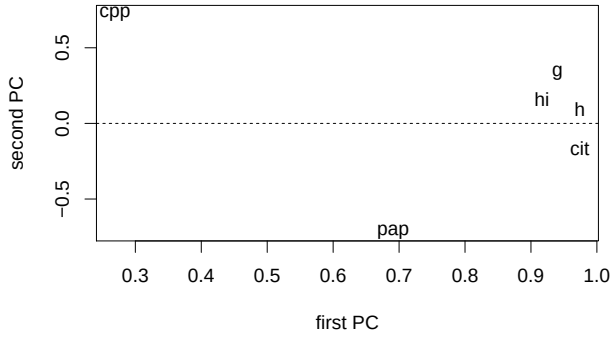

Fig. 15. Scatterplot of journal indicators on the two principal components.

PC	**h**	**g**	**cit**	**hi**	**pap**	**cpp**
PC1	**0.97**	**0.94**	**0.97**	**0.92**	**0.69**	0.27
PC2	0.10	0.33	-0.17	0.16	**-0.72**	**0.72**

Table 7. Loadings of the first two principal components (varimax rotation method)

correlation cluster; the degree of correlation is, however, somewhat lower (0.84 compared to 0.91). Papers is moderately correlated with cites (0.74) and otherwise scarcely associated; cites per paper is isolated. The main correlation cluster is soon detected by the clustering procedures we have proposed using any of the three joining strategies.

	pap	**cit**	**cpp**	**h**	**g**	**hi**
pap	1.00	0.74	-0.28	0.44	0.34	0.27
cit	0.74	1.00	0.24	0.89	0.81	0.77
cpp	-0.28	0.24	1.00	0.51	0.65	0.58
h	0.44	0.89	0.51	1.00	0.95	0.82
g	0.34	0.81	0.65	0.95	1.00	0.80
hi	0.27	0.77	0.58	0.82	0.80	1.00

Table 8. Pearson correlation matrix for the journal sample on Web of Science

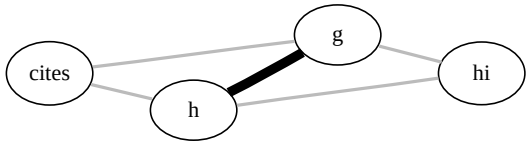

Fig. 16. Correlation graph for the journal sample on Web of Science

As for principal components analysis, Figure 17 shows the screeplot for the principal components: the first component explains 68% of the variance and the first two components account for 92% of the variance. Table 9 shows the varimax rotated component loadings and Figure 18 depicts the scatterplot of journal indicators on the first two principal components. Indicators h, g, and hi are clustered. Citations lies between this group and papers, marking a difference with respect to Google Scholar. Cites per paper is still cut off. These results are in accordance with the findings of Leydesdorff [Leydesdorff, 2009] and those of Bollen et al. [Bollen et al., 2009] with one minor difference: in Leydesdorff's analysis citations and papers are closer together than in our study.

PC	**h**	**g**	**cit**	**hi**	**pap**	**cpp**
PC1	**0.87**	**0.93**	0.61	**0.77**	0.03	**0.79**
PC2	0.46	0.31	**0.78**	0.35	**0.92**	-0.32

Table 9. Loadings of the first two principal components (varimax rotation method)

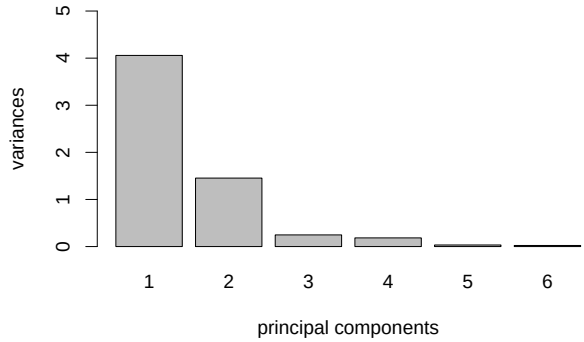

Fig. 17. Screeplot of the variances of principal components

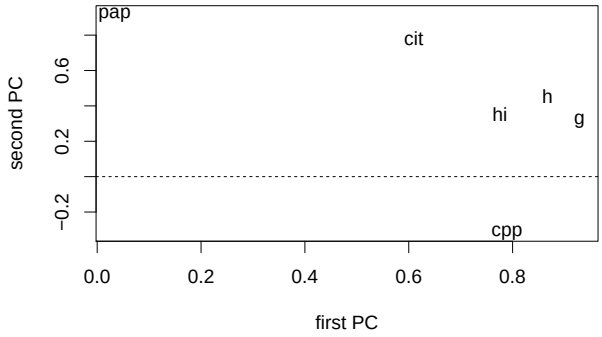

Fig. 18. Scatterplot of journal indicators on the two principal components

The average correlation for each indicator with respect to other indicators on Google Scholar and on Web of Science is shown in Table 10.

source	pap	cit	cpp	h	g	hi
Google Scholar	0.40	0.71	0.22	0.76	0.73	0.74
Web of Science	0.30	0.69	0.34	0.72	0.71	0.65

Table 10. Average correlation for journal indicators

The outcomes confirm the correlation graph analysis: papers and cites per paper are the most isolated indicators, while cites, h, g, and hi are highly correlated. The degree of association is higher for Google Scholar.

4.4 A global performance indicator

In the following we show how the outcomes of our experiments can be exploited to design a *fair* and global performance metric for scholars and journals. By fair indicator we mean an indicator that equally weights orthogonal and important aspects of research performance. The need for such an indicator is illustrated in the following example. Suppose we want to design a global indicator using partial indicators A, B, C, and D. We might blindly take the arithmetic mean of the four indicators. However, suppose that A and B are strongly correlated, while C and D are independent from A and B and among each other. Hence, both A and B measure roughly the same performance aspect, while C and D measure different aspects. By taking the arithmetic average, we would weight twice the performance aspect related to A and B to the detriment of the aspects measured by C and D. A better solution is to take the weighted average in which A and B are loaded one-half.

We follow such intuition to design a *global performance indicator* (gpi) that takes into account the correlation structure among partial performance indicators. In particular, we use the loadings of the principal component analysis to weight the various indicators. Let B be a bibliometric matrix with n bibliometric units and m bibliometric indicators. We compute the gpi as follows:

1. the bibliometric matrix B is standardized to allow equal scales among the different indicators;
2. $k < m$ principal components are computed along with the corresponding loading matrix L in which the entries are the absolute values of the original figures;
3. the matrix product $P = B \times L$ is performed. The resulting matrix P has n rows and k columns. Each row contains the k principal scores for the corresponding bibliometric unit;
4. finally, a global performance score for the i-th bibliometric unit is given by the (arithmetic) average of the i-th row of P. In this way, we equally weight each independent aspect.

Table 11 shows the rankings of scholars according to the above defined global performance indicator computed on Google Scholar and Web of Science. The Spearman association between the two compilations is 0.78, p-value 0.0024. Table 12 gives global performance rankings of journals both on Google Scholar and Web of Science. The Spearman association between the two compilations is 0.73, p-value 0.00035.

As a further example of application of the techniques proposed in this paper, we considered the two most popular college and university rankings: THE-QS World University Rankings (THE-QS) and Academic

Google Scholar		Web of Science	
gpi	**scholar**	**scholar**	**gpi**
5.90	FH	GF	4.71
2.26	LC	LC	3.33
1.88	GF	FH	2.92
0.66	AP	SM	1.41
0.64	AM	AP	0.55
-0.54	CT	AM	0.13
-0.55	AD	VR	-0.58
-0.86	SM	CT	-0.62
-1.12	CP	CP	-0.66
-1.41	MF	AD	-1.15
-1.93	ML	MF	-3.03
-2.07	MM	MM	-3.14
-2.86	VT	ML	-3.89

Table 11. Global performance rankings of scholars

Google Scholar		Web of Science	
gpi	**journal**	**journal**	**gpi**
3.27	IEEE T PARALL DISTR	IEEE T EVOLUT COMPUT	4.351
3.25	THEOR COMPUT SCI	FUZZY SET SYST	2.419
2.23	COMMUN ACM	THEOR COMPUT SCI	2.363
1.92	J SYST SOFTWARE	IEEE ACM T NETWORK	1.015
0.93	ACM COMPUT SURV	IEEE T PARALL DISTR	0.987
0.89	FUZZY SET SYST	COMMUN ACM	0.852
0.64	IEEE T EVOLUT COMPUT	ACM COMPUT SURV	0.362
0.36	J ALGORITHM	J SYST SOFTWARE	0.047
-0.30	IEEE T NEURAL NETWOR	SIAM J COMPUT	-0.013
-0.44	IEEE ACM T NETWORK	IMAGE VISION COMPUT	-0.117
-0.65	IMAGE VISION COMPUT	COMPUT METH PROG BIO	-0.227
-0.85	INFORM COMPUT	IEEE T NEURAL NETWOR	-0.734
-0.91	J PARALLEL DISTR COM	J ALGORITHM	-0.860
-1.02	J SYMB COMPUT	INFORM COMPUT	-1.061
-1.11	IBM SYST J	J SYMB COMPUT	-1.181
-1.35	J ACM	J PARALLEL DISTR COM	-1.455
-1.41	J COMPUT SYST SCI	IBM SYST J	-1.609
-1.56	SIAM J COMPUT	EVOL COMPUT	-1.695
-1.81	COMPUT METH PROG BIO	J COMPUT SYST SCI	-1.710
-2.09	EVOL COMPUT	J ACM	-1.734

Table 12. Global performance rankings of journals

Ranking of World Universities (ARWU). THE-QS is compiled by Times Higher Education in association with Quacquarelli Symonds and combines the following six weighted indicators: (1) academic peer review (40%); (2) employer review (10%); (3) faculty student ratio (20%); (4) citations per faculty (20%); (5) international faculty (5%); (6) international students (5%). ARWU is published by Shanghai Jiao Tong University and is based on the following six weighted criteria: (1) alumni winning Nobel Prizes and Fields Medals (10%); (2) staff winning Nobel Prizes and Fields Medals (20%); (3) highly cited researchers (20%); (4) articles published in Nature or Science (20%); (5) publications in Web of Science (20%); (6) per capita performance on the above indicators (10%).

We analysed the 2008 top-100 university rankings in both cases. THE-QS indicators are mostly independent; the higher association degree is between the two international scores (Spearman 0.68). Interestingly, the reviews of worldwide academic peers and of university employers do not show a striking association and correlate less than the international scores (0.55). Since indicators are independent, no clustering is necessary and a global performance index (GPI) can be obtained by taking the arithmetic (unweighted) mean of all indicators. The GPI and the THE-QS rankings correlate at 0.75, with a median change of rank of 12 positions and a maximum rank change of 52 positions (University of St Andrews). Only 4 universities maintain the same rank in both compilations. By contrast, ARWU indicators form two main correlation clusters: one contains alumni and staff price winners (indicators 1 and 2), and the second comprises the three bibliometric indicators (indicators 3, 4, and 5). In particular, highly cited researchers and articles published in Nature or Science have the best correlation (0.82). In this case, we computed a global performance index (GPI) using factor loadings as explained above. The two rankings are similar (correlation 0.98), with a median rank change of 2 positions, a maximum rank change of 32 positions (Ecole Normale Super Paris), and 24 universities that have the same position in both compilations.

5 Conclusion

We proposed different clustering techniques to group bibliometric indicators used in the quantitative assessment of research quality and compared the results with principal component analysis. Our methods can be applied to reduce the complexity of the space of bibliometric indicators when it contains many dependent metrics. In particular, they allow the design a composite performance indicator that considers independent aspects of research performance. We are aware that clustering, being an unsupervised learning approach, constitutes a descriptive and exploratory method. Moreover, our results may be sensible to both data sources and data sets used in our experiments. Nevertheless, the principal component analysis matches well the results of the cluster analysis. More importantly, our experiments are fully reproducible: the methodology is clear and the necessary software tools are freely available (with the exception of Web of Science). This allows interested scholars to compute bibliometric bases in other fields or in the same field on different data sets.

With respect to the evaluation of individual scholars, the clustering method discovers a base of metrics composed of the following indexes:

1. number of papers, measuring scholar productivity;
2. total number of citations, measuring absolute (size-dependent) impact of the scholar;
3. average number of citations per paper, measuring relative (size-independent) impact of the scholar;
4. m quotient, measuring enduring impact over time.

With respect to the assessment of journal performance, orthogonal measures are number of papers, measuring the size of the journal, h index (or total number of citations), capturing the absolute impact of the journal, and average number of citations per paper, accounting for the relative impact of the journal.

We noticed that the h index is a size-dependent indicator which is more correlated to the total number of citations rather than to the number of papers. This holds true for journals over both Google Scholar and Web of Science and for scholars over Web of Science. Interestingly, on the scholar sample evaluated on Google Scholar, the h index is better associated to papers (Pearson 0.88) rather than to citations (Pearson 0.82). This discrepancy is related to the content of the two data sources – Web of Science contains mainly journal publications while Google Scholar finds different types of sources, including conference papers and books [Meho and Yang, 2007,Franceschet, 2009], which are important publication sources in computer science – and to the difference in the two datasets – at scholar level, Google Scholar finds more citing (source) publications as well as more cited (target) papers with respect to Web of Science, while at journal level the possible differences are only in the set of citing publications, since the set of target papers is fixed.

Finally, we observed that the m quotient defines a separate performance dimension only on Google Scholar. We explain this as follows. Recall that m is the ratio between the h index and the academic age of a scholar. We noticed that scholars with high m values on Google Scholar are young scholars with a good h score for their age, mostly obtained thanks to conference papers. These scholars have low values for the other indicators because they published few papers with respect to senior scholars. This explains the low average correlation of m quotient on Google Scholar. The same young scholars have few *journal* papers, which take longer to be written and published, and hence their h and m scores on Web of Science are small. Thus m is better correlated with other indicators when computed on Web of Science. It follows that in the field of computer science and in fields with similar publication patterns, the m quotient can be used to discover talented young researchers when computed on Google Scholar or on similar data sources.

References

[Anderberg, 1973] Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.

[Batista et al., 2006] Batista, P. D., Campiteli, M. G., and Konouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189.

[Bollen et al., 2009] Bollen, J., de Sompel, H. V., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. Retrieved April 9, 2009, from `http://arxiv.org/abs/0902.2183`.

[Bollen et al., 2006] Bollen, J., Rodriguez, M. A., and de Sompel, H. V. (2006). Journal status. *Scientometrics*, 69(3):669–687.

[Bornmann et al., 2008] Bornmann, L., Mutz, R., and Daniel, H. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from Biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5):830–837.

[Braun et al., 2006] Braun, T., Glänzel, W., and Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1):169–173.

[Callon et al., 1983] Callon, M., Courtial, J.-P., Turner, W., and Brain, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22:191–235.

[Cormen et al., 2001] Cormen, T. H., Stein, C., Rivest, R. L., and Leiserson, C. E. (2001). *Introduction to Algorithms*. McGraw-Hill Higher Education.

[Costas and Bordons, 2007] Costas, R. and Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3):193–203.

[Dubes and Jain, 1988] Dubes, R. C. and Jain, A. K. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

[Egghe, 2006] Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1):131–152.

[Franceschet, 2009] Franceschet, M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*. Forthcoming.

[Garfield, 1979] Garfield, E. (1979). *Citation indexing: its history and applications in science, technology and humanities*. Wiley, New York.

[Gonzalez, 1985] Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306.

[Gonzalez and Sahni, 1976] Gonzalez, T. F. and Sahni, S. (1976). P-complete approximation problems. *Journal of the ACM*, 23:555–565.

[Harold and Means, 2004] Harold, E. R. and Means, W. S. (2004). *XML in a Nutshell*. O'Reilly, 3rd edition.

[Hendrix, 2008] Hendrix, D. (2008). An analysis of bibliometric indicators, National Institutes of Health funding, and faculty size at Association of American Medical Colleges medical schools, 1997-2007. *Journal of the Medical Library Association*, 96(4):324–334.

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

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. Springer.

[Katsaros et al., 2006] Katsaros, C., Manolopoulos, Y., and Sidiropoulos, A. (2006). Generalized h-index for disclosing latent facts in citation networks. Retrieved December 19, 2008, from `http://arxiv.org/abs/cs.DL/0607066`.

[Kessler, 1963] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25.

[Leydesdorff, 2009] Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*. Forthcoming.

[Meho and Yang, 2007] Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125.

[Moore, 2006] Moore, D. (2006). *Basic Practice of Statistics*. WH Freeman Company, 4th edition.

[R Development Core Team, 2007] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science and Technology*, 24:265–269.

[van Raan, 2006a] van Raan, A. F. J. (2006a). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502.

[van Raan, 2006b] van Raan, A. F. J. (2006b). Measuring science. Capita selecta of current main issues. In Moed, H. F., Glänzel, W., and Schmoch, U., editors, *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*, pages 19–50. Kluwer Academic Publishers.