Journal of Informetrics 10 (2016) 789–813

Contents lists available at ScienceDirect

Journal of Informetrics

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

Gazing at the skyline for star scientists

A. Sidiropoulos [a], A. Gogoglou [b], D. Katsaros [c,d,*,1], Y. Manolopoulos [e]

[a] Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, Greece
[b] Department of Informatics, Aristotle University of Thessaloniki, Greece
[c] Electrical Engineering Department & Yale Institute for Network Science, Yale University, United States
[d] Electrical and Computer Engineering Department, University of Thessaly, Greece
[e] Department of Informatics, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 8 April 2016
Accepted 10 April 2016
Available online 14 June 2016

Keywords:
Skyline operator
Perfectionism index
Hirsch index
Scientometrics

ABSTRACT

Admittedly, despite the plethora of scientometric indices proposed to rank scientists, none of them can fully capture the performance and impact of a scientist, since each index quantifies only one or a few aspects of his/her multifarious performance. Therefore, the task of scientometric ranking can be seen as a multi-dimensional ranking problem, where the different indices comprise the dimensions. The application of the skyline operator comes then as a natural solution to the problem. In this article we apply the skyline operator to scientist ranking to identify those scientists whose performance cannot be surpassed by others' with respect to all attributes. This technique can be used as a tool for short-listing distinguished researchers in case of award nomination.

1. Introduction

The field of scientometric ranking has a long history starting with the introduction of the Garfield's famous Impact Factor (Garfield, 1955) for journal ranking, and continuing with recent indices that quantify an individual's performance such as the very popular *h*-index (Bornmann, Mutz, Hug, & Daniel, 2014; Hirsch, 2005). We focus here on the family of indices that use a single number to measure a scientist's performance. Members of this family are some straightforward measures such as the average and total number of citations, the number of citations in the elite set of articles (Vinkler, 2011), variations of the *h*-index, such as the contemporary index (Sidiropoulos, Katsaros, & Manolopoulos, 2007), the *e*-index (Zhang, 2009), the *f* index (Katsaros, Akritidis, & Bozanis, 2009) and many more (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Wildgaard, Schneider, & Larsen, 2014). Their advantages and disadvantages have been documented in various studies, and the overall conclusion is that each one focuses on one (or more) but not all of the aspects of an individual's performance (project ACUMEN Wildgaard et al., 2014). For instance, the *h*-index is a proxy for the cumulative impact and productivity achievement, the contemporary *h*-index takes into account how contemporary the articles that comprise the *h*-index, the *e*-index complements the *h*-index by accounting for the ignored excess citations, etc. Therefore, it becomes evident that a fair evaluation of a scientist's work based on quantitative data must take into account multiple, uncorrelated indicators (Bornmann, Mutz, Hug, & Daniel, 2011).

* Corresponding author at: Electrical and Computer Engineering Department, University of Thessaly, Greece. Tel.: +30 2421074975.
E-mail addresses: asidirop@gmail.com (A. Sidiropoulos), agogoglou@csd.auth.gr (A. Gogoglou), dkatsar@inf.uth.gr (D. Katsaros), manolopod@csd.auth.gr (Y. Manolopoulos).

[1] Work done while the author was on sabbatical leave at Yale University.

http://dx.doi.org/10.1016/j.joi.2016.04.009
1751-1577/

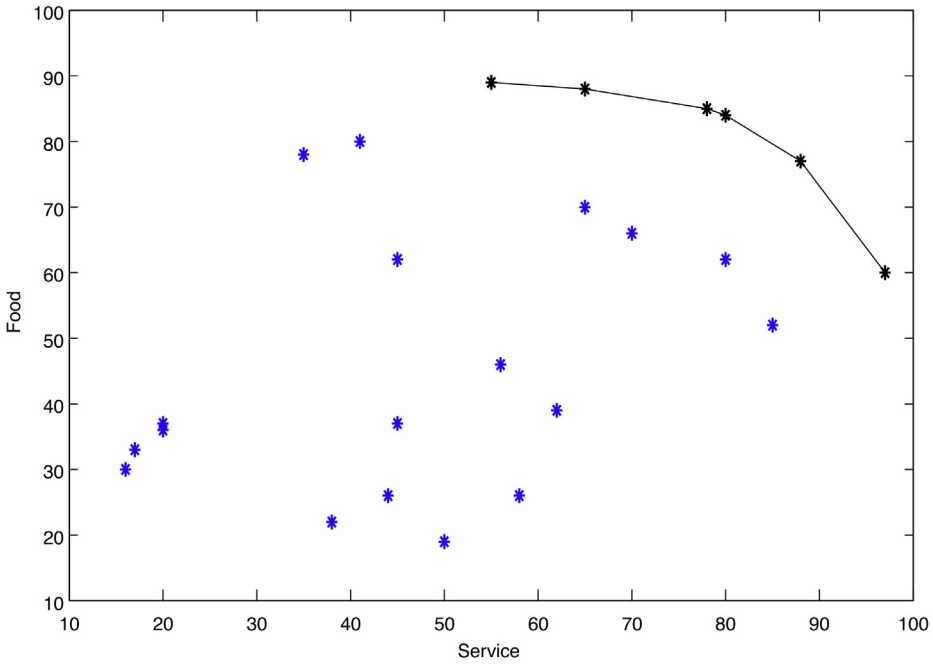

Fig. 1. Skyline plot for restaurants.

A straightforward way to do this is to define a set of weights for the set of indices and compute a weighted average score. A precondition for this process is the normalization of the scores in such a way that they are comparable. This is usually impossible, since most of the indices are not upper-limited. In addition to that, the definition of the corresponding weights will certainly be arbitrary. An alternative is to address the problem as a "rank aggregation" problem (Fagin, Lotem, & Naor, 2001; Langville & Meyer, 2014), and fuse the ranking lists produced by each indicator by appropriately adapting methods such as those reported in Tsai (2014). Still, the selection of the fusion algorithm will raise questions about its appropriateness and fairness.

We argue here that we do not need to produce a single ranked list; we should simply identify those individuals that have not been surpassed by than any other individual with respect to all considered indicators. This is the concept described as *skyline*, and calculated by the respective operator (Börzsönyi, Kossmann, & Stocker, 2001). The resulting set of distinguished individuals is called the *skyline set*.

We will explain how this works by presenting an example from Chomicki, Godfrey, Gryz, and Liang (2003). Assuming that we want to choose a restaurant based on its service and food quality. Having a rating for each restaurant's service and their food quality we can produce two rank tables, one for each evaluation metric (service,food). It is difficult to produce a global rank table by combining the existing two. That is because we cannot define the relation between service and food. Attempts to define such a relation will be prove to be arbitrary. The skyline set notion enables us to detect the best restaurants (given the attributes) by combining the two metrics (or more metrics). The skyline set consists of the set of restaurants (generally the set of objects) none of which can be surpassed with regards to any of the attributes by any other restaurant (or object in general). A geometrical view is shown in Fig. 1. In this plot every point represents a restaurant (an object). The coordinates of each object are defined by the score of the object for each metric. Each metric corresponds to one dimension. Since the higher the score of the two metrics (service and food) the better the object (restaurant), the objects that can surpass all the other objects are distinguished as the top choices. A two metrics rank can be presented with a 2D plot.

2. Definition and calculation of a dataset's skyline set

In this section we will present the original definition of skyline set and a basic, efficient algorithm for its computation, as presented by Borzsonyi adding the mathematical notation.

Definition 1 *(Dominance relationship).* Given two multidimensional points s_1 and s_2 with attributes (dimensions) α from a space D, if s_1 is equal to or better than s_2 in all dimensions, and s_1 is better than s_2 in at least one attribute, we would say that s_1 dominates s_2 and write $s_1 \succ s_2$. That is:

$$s_1 \succ s_2 : (\forall \alpha \in D, s_1.\alpha \geq s_2.\alpha) \wedge (\exists \alpha \in D, s_1.\alpha > s_2.\alpha).$$

Definition 2 *(Skyline set).* The skyline set comprises the set of points not dominated by any other point.

The generic concept of 'skyline' set as a group of points not 'fully dominated' by any other point dates back several decades. It is similar to the term *Pareto frontier* in the economics studies, which is used to describe a set of options which are 'Pareto efficient'; Pareto efficiency refers to a set of resource allocations to a population where it is not possible to improve the allocation to one individual without hurting someone else. Therefore, the notions of Pareto frontier and skyline set are similar, but our application area is not related to any resource allocation problem. Nevertheless, the concept of (weak/strong) Pareto dominance (Voorneveld, 2003) for options with multiple features as used in multicriteria optimization and decision making is alike to skyline dominance. The technique of Data Envelopment Analysis (Rousseau & Rousseau, 1997) is also based on Pareto efficiency; it focuses on relative performance based on the equilibrium between resources (personnel, funding, etc.) and resulting productivity (number of published papers, citations acquired, etc.) which is not the case with the skyline set. The skyline operator or skyline algorithm is mostly similar to the maximal vector problem (Kung, Luccio, & Preparata, 1975) which seeks to find the subset of vectors such that each one is not dominated by any of the vectors from the set; in this setting, one vector dominates another if each of its components has an equal or higher value than the other vector's corresponding component, and it has a higher value on at least one of the corresponding components.

Despite the profound similarities between the notions of skyline operator, pareto frontier and data envelopment analysis, we will continue our discussion using the skyline operator approach for two reasons: (a) due to the fact that this notion has several variations which can be further investigated in the bibliometric field, and (b) due to the rich literature on algorithmic/computational aspects of the skyline algorithm and its applications.

There are various skyline algorithms surveyed in Godfrey, Shipley, and Gryz (2007) and Tiakas, Papadopoulos, and Manolopoulos (2015) depending on the complexity and efficiency requirements and in our case we have opted for the Sort-Filter skyline (SFS) algorithm (Chomicki et al., 2003) that is a fairly straightforward algorithm with minimal computational cost. SFS is computationally efficient in the sense that it presorts the data points of the initial set according to their sum of attributes (descending or ascending order) to make sure that no point can be dominated by the ones that come after it. The pseudocode for the SFS algorithm is displayed below, as presented in Chomicki's original work with added notes for clarification of the steps:

Algorithm 1. SFS algorithm.

Require: A dataset T with d dimensions
Ensure: A set of skyline records S

1: Sort all records by the sum of all attributes in descending (or ascending) order leading to dataset T'
2: $S \leftarrow 0$ (Note: Initialize the set of skyline records S)
3: Move the first record from T' into S
4: **while** $T' \neq 0$ (Note: While there are still records in our dataset T) **do**
5: Compare each record t of T' with all the records in S
6: **if** t is dominated by a record in S **then**
7: Remove t from T' (Note: Point t is discarded since it has been dominated by an other one.)
8: **else**
9: move t from T' to S (Note: Point t is a candidate for the final skyline set since it has not yet been dominated by another one.)
10: **end if**
11: **end while**

In the preprocessing state of the algorithm (step 1) the points of our original set T are sorted according to their sum of attributes (T'). Attributes need to be normalized to have the same range of values. Should the attributes represent data where the higher the value the better (like our restaurant example) sorting is performed in descending order. If the attributes represent data where the smaller the value the better (for instance rank order) sorting is performed in ascending order. The set of skyline records S is initialized to be empty and the first record is moved to the set (steps 2–3). Afterwards, each record is compared with the existing points in the skyline set (step 5) and if it is dominated by at least one of them, as defined by the dominance relationship in Definition 1, it is discarded from the set. As can be seen in step 9, while a point t has not yet been dominated by another point in the skyline set it remains a candidate for the final skyline set. Through the recursive execution of steps 6–7, point t will be removed from the skyline set only when it has been dominated by another point. If that does not occur, then the point will remain in the final set of skyline records. This way it is ensured that a point t will only be compared with the ones that could surpass it with respect to the given attributes, avoiding unnecessary comparisons and thus reducing the computational cost.

In the majority of cases the skyline records of a real world dataset are a small percentage of the whole set of points leading to reduced computational cost especially after the initial presorting step. In our scientometric scenario, the sum of attributes represents the rank of the author based on the scientometric indices (i.e., attributes of the skyline operator); the scientists (data points) are sorted in ascending order. This way the authors on top of the list will have acquired a rank higher than the rest of the list, meaning that they are better candidates for the skyline set.

The concept of skyline operator is very powerful and its variations may have many applications in scientometrics. For instance, using our techniques we can easily implement the 'personalized' skyline operator (Lee, You, & Hwang, 2009) which aims at identifying 'interesting' objects based on user-specific preferences and/or retrieval size *k*. In our scientometric scenario, for a skyline operator with 2 dimensions (e.g., *h*-index and Perfectionism index (Sidiropoulos, Katsaros, & Manolopoulos, 2015)), one can retrieve those scientists whose weighted combination of ranks with respect to these two indicators belongs to the skyline set. Furthermore, we can even relax the concept of dominance and ask it to occur in only some of the dimensions (Chan, Jagadish, Tan, Tung, & Zhang, 2006), or even implement techniques for retrieving 'thick'

skyline sets (Jin, Han, & Ester, 2004), where apart from the true skyline points, one retrieves the points which are within ϵ-distance from the true skyline set.

As a final comment on the origins and variations (Chomicki, Ciaccia, & Meneghetti, 2013) of skyline operators, we need to say that even though we can examine various aspects of a skyline operator such as those mentioned above or additional ones such as dimension significance and diversity (Magnani, Assent, & Mortensen, 2014), in this article we use its plainest version (Börzsönyi et al., 2001).

3. Skyline operator in scientometrics

In principle, the application of the skyline operator for recognizing the excellent individuals is appealing, but first we need to establish that a distinguishable skyline does exist. If all points (i.e., scientists) are concentrated in a huge cluster and their differences are minor, then a clearly distinguishable skyline set would be difficult to find. Moreover, the existence of a distinguishable skyline set should be confirmed for different numbers and combinations of dimensions (i.e., scientometric indicators). Additionally, the *stability* of skyline set should be examined, since a stagnant skyline set with no differences (in terms of size and/or membership) over the years presents limited interest. Such a skyline set would indicate that no "newcomers", having recently risen to top scientists, have entered the skyline set. Moreover, scientists that have failed to maintain their status and have been surpassed by younger scientists are expected to be left out of the skyline set eventually. In the following two subsections we provide details about the datasets used in our study (results of Sections 3.2 and 5), and we investigate the usefulness of skyline operator as a scientometric tool.

3.1. Data collection

We collected data from Microsoft Academic Search (MAS)[2] to test these issues. More specifically, we extracted full citation data starting from year 1980 up to 2013[3] concerning the top-500 scientists in the Databases domain and the Networks domain as well as a larger set of scientists regarding the more general Computer Science field, and calculated several single-number performance indicators. The main reason we have chosen MAS for our data collection is that it is open access and provides an API with structured metadata. This allows us to access data for scientists based publication field, their accredited h-index as calculated by MAS and their number of total publications and citations. Moreover, it provides a detailed domain categorization, where each domain is comprised of many other subdomains that allow for an efficient field specific search. For the purposes of this study, the domains Databases and Networks as well as the field of Computer Science correspond to the categorization provided by MAS. In the experiments presented in this article we have used the following datasets:

1 *Databases dataset (DB)*: This dataset consists of top researchers working on the Databases domain. It has been acquired by querying the MAS for the top 500 researchers of the database domain based on their h-index as computed by the MAS system.
2 *Networking dataset (NET)*: As with the previous dataset, it has been acquired by selecting the top 500 researchers of the Networking field by querying the MAS API.
3 *Computer Science dataset (CS)*: This dataset contains all the Computer Science people, 29,856 scientists. It has been acquired by selecting all researchers with h-index greater or equal to 8 from all researchers belonging to the CS domain (as reported by MAS).

Regarding the CS dataset, a choice of threshold for the h-index was necessary to restrict the data set of the entire Computer Science field and focus on the most prolific scientists based on the h-index. Other values of the h-index could be considered in the range of [5, 9]. However, given the particular set of authors selecting a lower h-index did not restrict the number of authors as desired and failed to exclude from the data set authors with very few publications or those that ceased being active for the majority of our given years. A choice of h-index higher than 8 seemed to restrict the set of authors to less than a third of the original sample size, which we considered to be too restrictive. Moreover, an h-index of 8 is deemed a common value in the field of Computer Science for consistently publishing scientists, but this value can be adjusted to the particular characteristics of any other field of study.

After defining these datasets, we extended them by acquiring all the publications for the corresponding researchers and then all the citations to their publications, leading to a data set of over 3 million publications. To run the experiments, we had to build rank tables for each year. We built rank tables for each year starting from 1980 until 2013. This means that the temporal information for each publication (as well as for each citation) is essential. Unfortunately this information is missing for 11.2% of the publications in our datasets on average. To complete this missing information, we gathered data from DBLP[4] by using the XML search DBLP API. This way, we recovered about 6% of the missing information. After that step, the publications with missing year were ignored. It needs to be noted that the missing years constitute only 6% of the total

[2] We appreciate the offer of Microsoft to gratis provide their database API.
[3] Citation data of 2013 are not very rich though.
[4] http://dblp.uni-trier.de/.

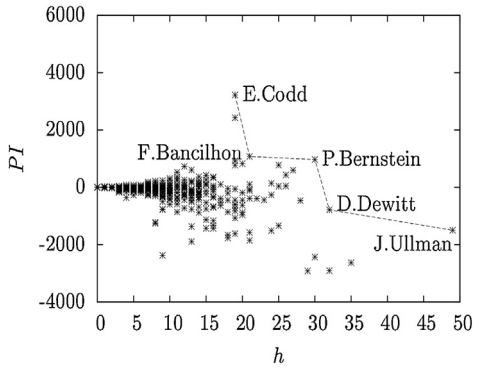

Fig. 2. 2D skyline set (*h*, *PI*) for top-500 Databases researchers for 1995.

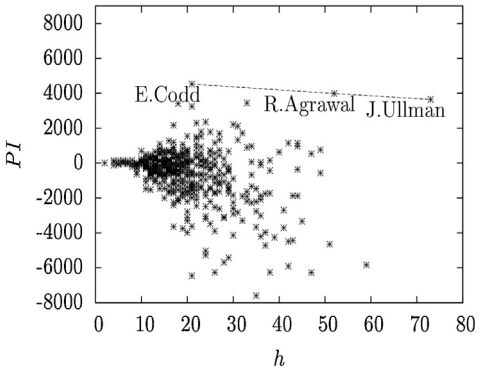

Fig. 3. 2D skyline set (*h*, *PI*) for top-500 Databases researchers for 2005.

publications for all scientists included and that could perhaps change the value of the bibliometric indices for a small number of authors at a specific year or two, but their overall scientific status as well as their placement in the skyline set over the years should not change.

3.2. Skyline operator's value as a scientometric tool

For the sake of presentation clarity, we present 2-dimensional skyline sets for database scientists, based on the DB dataset. In a later Section 5.1, we will elaborate on that. We examined various combinations of indicators all of which produced distinguishable skyline sets. In Figs. 2 and 3 we present the 2D skyline set for the years 1995 and 2005 respectively, with dimensions corresponding to the *h*-index and the *Perfectionism* index, *PI* (Gogoglou, Sidiropoulos, Katsaros, & Manolopoulos, 2015; Sidiropoulos et al., 2015). Perfectionism index is a recently proposed indicator which aims at differentiating among laconic and mass-producing scientists. It is defined as the summation of citations in the *h*-core and the excess area minus the citations "lost" in the complement area of the tail area, thus combining information about the whole citation curve of a scientist. We have shown that *PI* is uncorrelated to *h*-index. Each star represents a scientist and its projection to the *x*-axis (*y*-axis) depicts his/her *h*-index (*PI*). The line connects the points (scientists) who comprise the skyline set.

For year 1995 (Fig. 2), the skyline set consists of E. Codd, F. Bancilhon, P. Bernstein, D. DeWitt, and J. Ullman. For year 2005 (Fig. 3), the skyline set consists of E. Codd, R. Agrawal and J. Ullman. As expected, the performance of scientists (in terms of *h*-index and *PI*) has changed. The *h*-index is a non-decreasing indicator, whereas *PI* may increase or decrease, but here it has increased for all the designated scientists. However, the really interesting thing is that both the size and composition of the skyline set has changed. We have "new-comers" (i.e., R. Agrawal), "departers" (i.e., F. Bancilhon, P. Bernstein, D. DeWitt), and "stable stars" (i.e., E. Codd, J. Ullman). The common feature of the plots is the existence of a large, compact cluster and of the distinguishing scientists located in the right and upper part of the plots which are members of the skyline set or candidates for a future skyline set. This observation indicates that given a set of attributes without strong correlations with each other, a distinguishable skyline set can be produced over the years. Of course, when other features are added and the datasets become bigger larger sets of authors are expected to be distinguished. The main reason why the skyline set consists of a relatively small number of scientists is because the dataset only consists of 500 authors and it is only a small example of the distinguishing power of the skyline algorithm. In the experiments' Section 5 we are going to test this indication for different combinations of indexes and contemplate how the different choices of indices can provide a larger set as skyline, depending of course on the size of the original data set.

4. Selection of dimensions for the skyline operator

In the previous section's short experiment we explained the value of the skyline operator concept in scientometric analysis by selecting its dimensions intuitively and ad-hoc. Clearly, the selection (number and identity) of skyline operator's dimensions needs to be performed methodically, and this section's goal is to address this issue, i.e., to analyze the correlations between single-number scientometric indicators and select the representative ones.

For this purpose, we examined the 38 indicators seen in Table 1 which focus on modeling the size and/or shape of the (whole or parts of the) citation curve and capturing the productivity and/or impact of a scientist. These are indices that characterize author-level performance and were chosen for the purposes of the present work primarily because they are found to be popular in the literature and have been analyzed in various other studies (as mentioned in Section 1). Additionally, they that can be calculated with publically available information regarding citations and years of publications as well as number of co-authors, which can be available in a scientist's personal page or any on-line citation database. There are many other indices that would require different input to be calculated that is not so easily available or a higher computational cost. Also we attempted to include indices concerning all the areas (h-core, tail, excess are) or the axes (citations–publications) of the citation curve.

Our goal is to apply unsupervised learning methodologies to identify the groups of indices that convey the same information about the citation curve and in this way provide an easily interpretable categorization of the 38 indicators reducing the dimension of the original feature space. Unsupervised learning has also been employed in Bollen, van de Sompel, and Chute (2009) to group indices, but that study focused on graph and network features rather than on individual scientists. In the present work, the resulting clusters of indices represent the areas of the citation curve each indicator focuses on and the qualities of publishing behavior that they emphasize.

For our analysis we have opted for a dataset that can be considered statistically diverse and representative of the publishing behaviors of a variety of authors. The set (CS dataset) was chosen so that it contains different categories of authors from the Computer Science field, i.e., productive authors with a large number of publications, highly distinguished authors with a

Table 1
This set of 38 single-number, author-level scientometric indices was examined as potential skyline dimensions.

	Indicator	Abbreviation	Reference
1	number of citations	C	
2	number of publications	P	
3	citations/paper	C/P	
4	h index (Hirsch)	h	Hirsch (2005)
5	a parameter (Hirsch)	aHirsch	Hirsch (2005)
6	g index (Egghe)	g	Egghe (2006)
7	Perfectionism index	PI	Sidiropoulos et al. (2015)
8	m-quotient (Hirsch)	mHirsch	Hirsch (2005)
9	excess index (Zhang)	e	Zhang (2009)
10	A index (Jin)	A	Jin, Liang, Rousseau, and Egghe (2007)
11	R index (Jin)	R	Jin et al. (2007)
12	AR index (Jin)	AR	Jin et al. (2007)
13	Age-Weighted Citation Rate	JAWCR	Jin et al. (2007)
14	Contemporary h index	hcont	Sidiropoulos et al. (2007)
15	h_2	h_2	Kosmulski (2006)
16	m of Bornmann	mBor	Bornmann, Mutz, and Daniel (2008)
17	h individual index	hI	Batista, Campiteli, and Kinouchi (2006)
18	h individual index (normalized)	hInor	Harzing (2007)
19	hm (Schreiber)	hm	Schreiber (2008)
20	f (Tol)	f	Tol (2009)
21	t (Tol)	t	Tol (2009)
22	h index weighted	hw	Egghe and Rousseau (2008)
23	x (Kosmulski)	x	Kosmulski (2007)
24	hg (Alonso)	hg	Alonso, Cabrerizo, Herrera-Viedma, and Herrera (2010)
25	q_2	q_2	Cabrerizo, Alonso, Herrera-Viedma, and Herrera (2010)
26	tapered h index	htap	Anderson, Hankin, and Killworth (2008)
27	normalized h index	hnor	Sidiropoulos et al. (2007)
28	metric1: (sqrt(C)-h)/sqrt(C)	metric1	Gogoglou et al. (2015)
29	metric2: (P-h)/P	metric2	Gogoglou et al. (2015)
30	metric3: (citations_in_tail/C)	metric3	Gogoglou et al. (2015)
31	power law coefficient	plaw	Gogoglou et al. (2015)
32	rationalized h index	hrat	Ruane and Tol (2008)
33	v (Vihinen)	v	Riikonen and Vihinen (2008)
34	w (Wu)	w	Wu (2010)
35	h of Miller	hmill	Miller (2006)
36	s (Silagdze)	s	Silagadze (2010)
37	hmock	hmock	Prathap (2010)
38	Wohlin	Wohlin	Wohlin (2009)

Table 2
Variance explained by the first 5 principal components.

	PC1	PC2	PC3	PC4	PC5
% of variance explained	78.032	15.654	5.910	1.952	0.922
Cumulative variance explained	78.032	88.686	94.596	96.548	97.47

Table 3
Bibliometric indices contained in the 3 clusters.

cluster1	v, hnor, mBor, PI/papers, metric1, aHirsch, metric2, metric3, plaw
cluster2	C, C/P, e, A, R, AR, JAWCR, x, hmock, Wohlin, hI, hw, s, htap
cluster3	P, h, g, mHirsch, hcont, h2, hInor, hm, f, t, hg, q2, hrat, w, hmil

high h-index (top 500 authors based on their h-index) and other authors with varying publishing patterns. This dataset was enriched for the purposes of the statistical analysis with a set of 500 authors with h-index<8, so that publishing patterns associated with lower scientific status are also represented. As a result, the set can be considered representative of the correlations between indices for various publishing behaviors.

Firstly, based on visual inspection of the distribution of all the bibliometric indices over the set of authors, we have deduced that it significantly differs from the normal distribution. This fact has also been established in various studies dealing with the distribution of citations, h-index and other bibliometric indexes (Eom & Fortunato, 2011; Gupta, Campanha, & Pesce, 2005), concluding that bibliometric indices follow exponential, Yule, Gumbell and other non-uniform distributions. Then, we examined whether Spearman or Pearson correlation coefficient is appropriate for constructing the correlation table between the indicators. Since Spearman coefficient (ρ) is independent of the normality of the data distribution (and thus it is considered a non-parametric statistic) and moreover it can handle ties, we used Spearman ρ coefficient for the construction of the correlation matrix.

The correlation matrix (38×38) was formulated and most indices were found to be significantly correlated (Spearman $\rho > 0.8$) with each other (significance level 5%), but there were also a few indices that showed no strong correlation (Spearman $\rho < 0.3$) with the majority of other indices. In other words, there seemed to be a highly populated group of indices and a few others that deviated strongly, as far as their distribution was concerned.

The aforementioned correlation analysis lead us to choose between Factor Analysis and Principal Component analysis to reduce the dimension of the 38×38 feature matrix so that we can better interpret the relationship between the indicators. Factor Analysis is considered inappropriate for our purposes, since the correlation matrix is not positive definite, i.e., there are some columns that are a linear combination of other columns (redundant factor). Consequently, we opted for Principal Component Analysis (PCA) to project the feature space of the 38 indicators characterizing a set of authors to a lower dimensional space. The matrix that was subjected to PCA had rows representing the authors in our dataset and 38 columns that contained the bibliometric indices for each author. Table 2 shows the first 5 principal components and the variances explained by them. As can be seen, the first 2 components explain 88.68% of the variance which leads us to choose these two components as our new coordinate system.

Fig. 4 displays the result of the PCA analysis on the coordinate system of the first two principal components. The two principal components offer two different but complementary points of view on the scientific performance. The first principal component expresses the number of papers that formulate the productive h-core, whereas the second principal component indicates the impact of the papers published by the author in relation to complementary areas of the h-core (excess and tail area). This basic categorization of bibliometric indices has been introduced by Bornmann et al. (2008). Here we utilize this categorization as a new coordinate system to further examine the relationships of the bibliometric indices. The new 2-dimensional space allows for our indicators to be clustered into three groups (see Fig. 4 and Table 3).

The three groups (cluster1, cluster2 and cluster3) that were formulated can be explained if we consider the nature of the indices that belong to them. The group with a high coefficient on the second principal component share a common trait, the fact that they take into account the whole citation curve, considering the total number of papers in their calculations and the impact these papers have, such as the Perfectionism index or the normalized h-index (h/total number of papers). Including Hirsch's a parameter (which is essentially the metric1 from Table 1) we can safely deduce that this group also expresses information about the tail of the citation curve. The tail-related metrics are found to penalize the citation curves that are highly skewed and have a long and thick tail (Gogoglou et al., 2015). These are the indices constituting cluster1. As we get closer to the first principal component, the focus shifts from the total citation count to the papers formulating the productive h core. The first group above the horizontal axes focuses on the citation count, taking into account also excess citations (cluster2), whereas the group below the horizontal axes contains the h core related indices (cluster3). Cluster3 is the most highly populated group, since there exist many variations of the h-index expressing similar information about the core of the citation curve (Fig. 5).

The above grouping of the 38 bibliometric indices was further verified by subjecting our set of authors with the bibliometric indices as variables to k-means clustering with predefined number of clusters $k = 3$. The distance function used for the k-means was Spearman correlation, since we intend to group our variables based on how strongly correlated they are to achieve high intra-cluster correlation. The results of this clustering were significantly similar to our original grouping

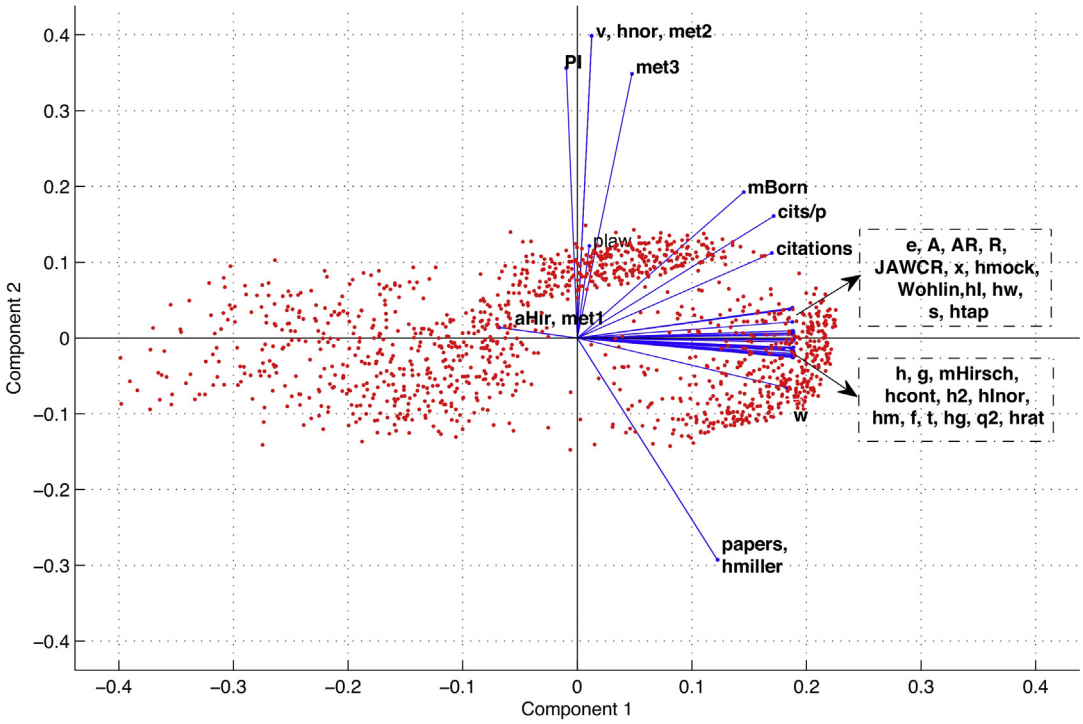

Fig. 4. Graphic representation of the Principal Component space based on the first 2 Principal Components. The blue vectors depict the 38 bibliometric indices and their position in the PC space, whereas the red dots depict the data points of our dataset projected on the PC space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

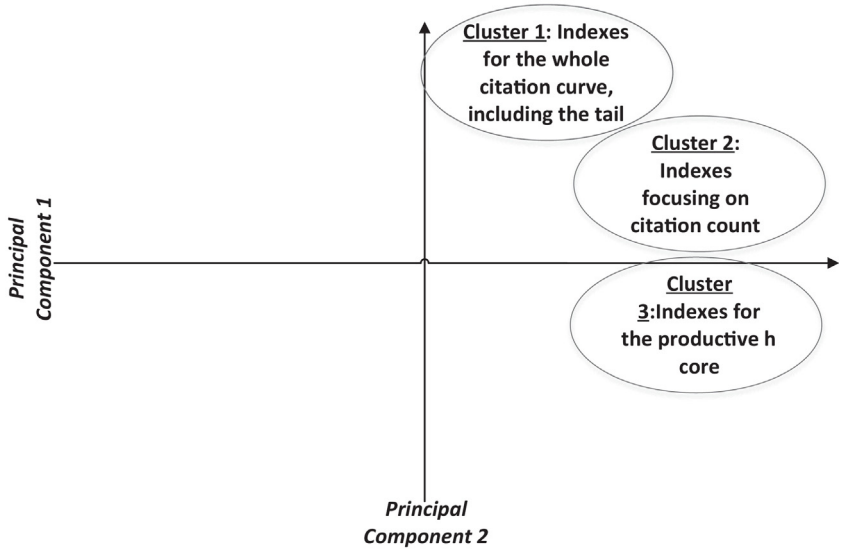

Fig. 5. Graphic representation of the 3 clusters of bibliometric indices on the 2 Principal Components coordinate system.

derived from PCA (see Table 3). A minor difference was that the *hw*-index was placed in cluster3 instead of 2 based on *k*-means, but we opted for our original placement of this index in cluster2, since it expresses a weighted *h*-index by citation impact meaning it gives more emphasis on the excess citation count rather than the *h*-core.

The next step is to specify the most representative indicator of each cluster. Considering for instance the cluster centroid as the most representative indicator is not appropriate, because the similarity of the indexes is based on the correlation of the ranking they assign to the author set and not the actual distance between the indices. Therefore we propose a simple technique that is to compute the ranks according to each indicator in a cluster and find a mean rank for every author per cluster. Then we compute Spearman correlation between this mean rank per cluster and the rank assigned by each

Table 4
Clusters' representative indicators.

Cluster	cluster1	cluster2	cluster3
Representative indice(s)	metric2	R	hrat (hg, h)

individual indicator of the cluster. The indicator whose ranking is found to be most correlated with the mean rank of the cluster it belongs to is declared as the most representative indicator of this cluster.

Table 4 contains the most representative index of every cluster according to the method described above. It is noteworthy that cluster3, being the most highly populated cluster has more than one indices with very similar Spearman correlation with the mean rank. All these indices could be considered representative of the cluster, since the differences in their correlation coefficient are minor. Table 4 depicts the representative metrics of each cluster, while the indicators in parentheses are the ones with a very similar correlation rank and could consequently be considered equally representative of each cluster.

With the number and identity of indicators at hand, we proceed to the calculation of the skyline set, which shall be a 3D skyline set.

5. Calculating the skyline set

In this section, we present the results of the experiments conducted with the three data sets (DB, NET, CS). For the experiments conducted, two different sets of bibliometric indices were chosen, so that one index from each of the clusters defined in Section 4 is selected to represent its respective cluster. By choosing one index from each cluster, we ensure that all different features of scientific performance are represented as attributes of the skyline operator. The first combination we contemplated is the set of the three cluster representatives (see Table 4) comprised by *R* (Jin et al., 2007), *h*-rationalized (Ruane & Tol, 2008) (*hrat*) and *metric*2 (Gogoglou et al., 2015). For the second set, we decided to incorporate a set of the most popular indices that belong to each cluster, namely *e*-index (Zhang, 2009) (excess index), the *h*-index and the *hnor* (Sidiropoulos et al., 2007). The second choice focuses more on popularity, meaning how well known and accepted the indices are, including indices that have been very well received by the scientific community and have existed in the literature for nearly a decade now.

However, experiments conducted with the set of popular indices resulted in skyline sets almost identical to the ones identified with the set of representative indices. That is to be explained, since *hnor* and *e*- index display a very high correlation (Spearman $\rho > 0.9$) with *metric*2 and *R*-index respectively. To display the flexibility of the skyline operator and indicate that we can have different sizes for the resulting skyline sets, dependent on the indices used, we have replaced the *hnor* with the more recently introduced *PI* (Sidiropoulos et al., 2015) (Perfectionism index), normalized to the number of papers. We use this newly formed set of indices to conduct the second set of experiments. The *PI* was chosen amongst the indices of cluster 1 for the purposes of the second experiment as it is very weakly correlated (Spearman $\rho < 0.3$) with the other two indices in the set (*e*- and *h*-index), thus resulting in more restrictive skyline sets including only the high impact scientists with overall selective publishing patterns. What is more, since the *PI* index takes into account the entire citation curve, when it is combined with the *e*-index that focuses on excess citations and the *h*-index that focuses on the *h*-core, the resulting skyline set constitutes of those scientists that have an overall high performance regarding all areas of the citation curve. On the other hand, the first set of indices (*metric*2, *R*, *hrat*) along with the original popular set of indices (*hnor*, *e*, *h*) result in a broader skyline set, where scientists of younger age or lower citation count can also be included as promising elite. By the end of this section, we will also investigate the adaptability and flexibility of the skyline operator by examining various other index combinations on our datasets and comparing the resulting skyline sets.

The results are presented in Tables 5–11. In Tables 7, 10 and 11 we also show the correspondence between skyline members and award winners for three awards, namely E.F. Codd, ACM SIGCOMM Lifetime Contribution and Turing award which are considered the top awards for the databases, networking and the whole computer science, respectively. We propose the skyline operator as a tool for recognizing star scientists for funds, promotions, etc. Since characterizing scientists as "star scientists" can be subjective and controversial, we have opted for identifying awarded scientists as a test case for validating the distinguishing power of the skyline operator and emphasizing on quality recognition.

5.1. Skyline sets in the databases field

Looking at Tables 5 and 6 which show the skyline members for the database field (for those having more than 50 publications), we observe first that the selection of dimension has significant impact upon the composition (size and contents) of the skyline set. The first set of dimensions generates a far larger skyline set than the second set of dimensions. The interesting thing though is that there is a small number of scientists which are present in both skyline sets, e.g., E. F. Codd, J. Ullman, Hector-Garcia-Molina, R. Fagin, M. Hammer a fact attributed to their scientific performance. The second interesting observation is that the content of both skyline sets change approximately every five years, which is attributed to the change and/or introduction of a new field of interest.

Looking at Table 6 we observe an increase in the skyline size (more than 100%) as time passes (from 1992 to 2013), and also a gradual change in the actual skyline set, i.e., new scientists enter the skyline set (e.g., R. Motwani), whereas some

Table 5
Members of Skyline set calculated from DB dataset with attributes (*metric2*, *R*, *hrat*). Every row represents the skyline set of a particular year. The first column displays the year and the second one contains the names of the scientists belonging to the skyline set that particular year.

Year	Skyline members								
1992	Beeri C., Lohman G.,	Bernstein P., Mendelzon A.,	Codd E., Papadimitriu C.,	DeWitt D., Sagiv Y.,	Domingue J., Shekita E.,	Fagin R., Stonebraker M.,	Hammer M., Tannen V.,	Haritsa J., Ullman J.	Koubarakis M.,
1993	Banchilhon F., Haritsa J., Banchilhon F.,	Beerie C., Koubarakis M., Bernstein P.,	Bernstein P., Lim E., Codd E.,	Codd E., Lohman G., Dewitt D.,	Dewitt D., Papadimitriu C., Fagin R.,	Fagin R., Stonebraker M., Gravano L.,	Goodman N., Swami A., Hammer M.,	Gravano L., Tannen V., Kossmann D.,	Hammer M., Ullman J. Papadimitriu C.
1994	Papak/ntinou Y., Worboys M.	Shekita E.,	Shim K.,	Stonebraker M.,	Suciu D.,	Swami A.,	Ullman J.,	Wei H.,	Wiener J.,
1995	Banchilhon F., Koubarakis M.,	Bernstein P., Narasayya V.,	Castellanos M., Papadimitriu C.,	Codd E., Shim K.,	Dewitt D., Stonebraker M.,	Fagin R., Ullman J.	Georgakopulos D.,	Gravano L.,	Hammer M.,
1996	Agrawal R., Koubarakis M., Agrawal R.,	Banchilhon F., Narasayya V., Banchilhon F.,	Beeire C., Papadimitriu C., Beeire C.,	Bernstein P., Shekita E., Bernstein P.,	Codd E., Shim K., Buneman P.,	Dewitt D., Stonebraker M., Casati F.,	Fagin R., Ullman J. Castellanos M.,	Georgakopoulos D., Chomicki J.,	Hammer M., Codd E.
1997	Dewitt D., Ullman J. Agiteboul S.,	Fagin R., Agrawal R.,	Haas P., Banchilhon F.,	Hammer M., Beeire C.,	Kuper G., Casati F.,	Maedche A. Castellanos M.,	Papadimitriu C., Chomicki J.,	Shekita E., Codd E.,	Stonebraker M., Dewitt D.
1998	Fagin R., Wei H., Agiteboul S.,	Gehrke J., Widom J. Agrawal R.,	Hammer M., Banchilhon F.,	Imilienski T., Beeire C.,	Papadimitriu C., Bernstein P.,	Papak/ntinou Y., Castellanos M.,	Shekita E., Chomicki J.,	Sistla A., Codd E.,	Ullman J., Dewitt D.
1999	Fagin R., Sarawagi S.,	Faloutsos C., Shekita E.,	Gehrke J., Sistla A.,	Hammer M., Ullman J.,	Livny M., Wei H.,	Maedche A., Worboys M.	Papadimitriu C.,	Papakonstantinou Y.,	Popa L.,
2000	Agiteboul S., Keogh E.,	Agrawal R., Papadimitriu C.,	Banchilhon F., Popa L.,	Beeire C., Shekita E.,	Castellanos M., Sistla A.,	Codd E., Stonebraker M.,	Fagin R., Ullman J.,	Hammer M., Wei H.,	Imilienski T., Worboys M.
2001	Agiteboul S., Hammer M.,	Agrawal R., Imilienski T.,	Banchilhon F., Keogh E.,	Beeire C., Shanmugasundaram J.,	Buneman P., Sistla A.,	Castellanos M., Stonebraker M.,	Codd E., Ullman J.,	Fagin R., Wei H.,	Florescu D., Widom J.

2002	Abiteboul S., Shim K., Abiteboul S.,	Agrawal R., Sistla A., Agrawal R.,	Castellanos M., Stonebraker M., Codd E.,	Codd E., Suciu D., Dewitt D.,	Fagin R., Tan W., Fagin R.,	Florescu D., Ullman J., Florescu D.,	Hadjieleftheriou M., Wei H., Hadjieleftheriou M.,	Hammer M., Widom J. Hammer M.,	Imielinski T., Imielinski T.
2003	Ives Z., Widom J.	Papak/ntinou Y.,	Popa L.,	Shanmugasundaram J.,	Shvaiko P.,	Sistla A.,	Suciu D.,	Tan W.,	Ullman J.,
2004	Abadi D., Keogh E.,	Agrawal R., Shekita E.,	Buneman P., Shvaiko P.,	Codd E., Sistla A.,	Fagin R., Suciu D.,	Florescu D., Swami A.,	Hammer M., Tan W.,	Hirschheim R., Ullman J.,	Imielinski T., Widom J.
2005	Abadi D., Imielinski T.,	Agrawal R., Motwani R.,	Buneman P., Shekita E.,	Chomicki J., Sistla A.,	Codd E., Swami A.,	Fagin R., Ullman J.,	Florescu D., Widom J.,	Hammer M., Wiener J.	Hirschheim R.,
2006	Agrawal R., Shekita E.,	Buneman P., Simeon J.,	Codd E., Sistla A.,	Fagin R., Swami A.,	Florescu D., Ullman J.,	Hammer M., Widom J.,	Hirschheim R., Wiener J.	Imielinski T.,	Motwani R.,
2007	Agrawal R., Shekita E.,	Buneman P., Sistla A.,	Codd E., Swami A.,	Fagin R., Ullman J.,	Florescu D., Widom J.,	Garcia-Molina H., Wiener J.	Hammer M.,	Hirschheim R.,	Motwani R.,
2008	Agrawal R., Motwani R.,	Buneman P., Shekita E.,	Codd E., Swami A.,	Fagin R., Ullman J.,	Florescu D., Widom J.,	Garcia-Molina H., Wiener J.	Hammer M.,	Levy A.,	Maedche A.,
2009	Agrawal R., Maedche A.,	Buneman P., Motwani R.,	Codd E., Shim K.,	Fagin R., Swami A.,	Florescu D., Ullman J.,	Garcia-Molina H., Wei H.,	Hammer M., Widom J.,	Imielinski T., Wiener J.	Ives Z.,
2010	Agrawal R., Madden S.,	Buneman P., Maedche A.,	Codd E., Motwani R.,	Fagin R., Shim K.,	Florescu D., Swami A.,	Garcia-Molina H., Ullman J.,	Hammer M., Widom J.	Imielinski T.,	Levy A.,
2011	Agrawal R., Motwani R.,	Buneman P., Swami A.,	Codd E., Ullman J.,	Fagin R., Widom J.	Florescu D.,	Garcia-Molina H.,	Hammer M.,	Imielinski T.,	Maedche A.,
2012	Agrawal R., Motwani R.,	Buneman P., Swami A.,	Codd E., Ullman J.,	Fagin R., Widom J.	Florescu D.,	Garcia-Molina H.,	Hammer M.,	Imielinski T.,	Maedche A.,
2013	Agrawal R., Motwani R.,	Buneman P., Swami A.,	Codd E., Ullman J.,	Fagin R., Widom J.	Florescu D.,	Garcia-Molina H.,	Hammer M.,	Imielinski T.,	Maedche A.,

Table 6
Members of Skyline set calculated from DB dataset with attributes (*PI*/*papers*, *e*, *h*). Every row represents the skyline set of a particular year. The first column displays the year and the second one contains the names of the scientists belonging to the skyline set that particular year.

Year	Skyline members								
1992	Bernstein P.,	Codd E.,	Fagin R.,	Ullman J.,					
1993	Bancilhon F.,	Bernstein P.,	Codd E.,	Fagin R.,	Goodman N.,	Ullman J.,			
1994	Bancilhon F.,	Bernstein P.,	Codd E.,	Fagin R.,	Papakonstantinou Y.,	Ullman J.,			
1995	Bancilhon F.,	Bernstein P.,	Codd E.,	Hammer M.,	Fagin R.,	Goodman N.,	Ullman J.		
1996	Agrawal R.,	Bancilhon F.,	Bernstein P.,	Codd E.,	Fagin R.,	Hammer M.,	Ullman J.		
1997	Bancilhon F.,	Bernstein P.,	Codd E.,	Fagin R.,	Goodman N.,	Hammer M.,	Ullman J.		
1998	Agrawal R.,	Bancilhon F.,	Bernstein P.,	Codd E.,	Fagin R.,	Hammer M.,	Ullman J.		
1999	Agrawal R.,	Bancilhon F.,	Bernstein P.,	Codd E.,	Goodman N.,	Hammer M.,	Imielinski T.,	Sistla A.,	Ullman J.
2000	Agrawal R.,	Bancilhon F.,	Codd E.,	Goodman N.,	Hammer M.,	Imielinski T.,	Sistla A.,	Ullman J.	
2001	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Sistla A.,	Ullman J.,			
2002	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Sistla A.,	Ullman J.,			
2003	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Sistla A.,	Ullman J.,			
2004	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Ullman J.,				
2005	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Swami A.,	Ullman J.,			
2006	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Swami A.,	Ullman J.,			
2007	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Swami A.,	Ullman J.,			
2008	Agrawal R.,	Codd E.,	Hammer M.,	Imielinski T.,	Swami A.,	Ullman J.,			
2009	Agrawal R.,	Codd E.,	Garcia-Molina H.,	Hammer M.,	Imielinski T.,	Swami A.,	Ullman J.		
2010	Agrawal R.,	Codd E.,	Garcia-Molina H.,	Imielinski T.,	Motwani R.,	Swami A.,	Ullman J.		
2011	Agrawal R.,	Codd E.,	Garcia-Molina H.,	Imielinski T.,	Motwani R.,	Swami A.,	Ullman J.		
2012	Agrawal R.,	Codd E.,	Garcia-Molina H.,	Imielinski T.,	Motwani R.,	Swami A.,	Ullman J.		
2013	Agrawal R.,	Codd E.,	Garcia-Molina H.,	Imielinski T.,	Motwani R.,	Swami A.,	Ullman J.		

others depart (e.g., R. Fagin, A. Sistla). Even though the skyline set does not change dramatically from one year to the next, we can observe a substantial change from one decade to the next. In the first decade ('92-'03), the skyline set is dominated by persons that contributed to the theory and practice of relational databases (e.g., F. Banchilhon) whose work took place during the seventies and eighties; during the second decade ('05-'13) the skyline set is dominated by scientists that contributed to the area of data mining (e.g., R. Agrawal, A. Swami). Recall that the study in (Rahm & Thor, 2005) revealed that 10 out of the 20 most referenced papers for the decade '96-'04 belong to the data mining literature. Nevertheless, there are scientists with steady presence in the skyline set – we recognize E. Codd and J. Ullman as omnipresent in the skyline set.

Proceeding to examine whether there is any correlation among an "Edgar F. Codd Innovations" award-winning scientist and its presence in the skyline set, we merged Tables 5 and 6 into Table 7, which includes all the DB researchers that won the "Codd" award and/or appeared in our skyline sets at least once. Each one of the table columns corresponds to a year after 1980. The year of the first award is 1992. The symbol 's1' shows that the corresponding researcher is in the skyline set with attributes (*metric*2, *R*, *hrat*) and the symbol 's2' shows that the corresponding researcher is in the skyline set with dimensions (*PI*/*p*, *e*, *h*) during the specific year and the symbol *s*12 that s/he is in both. The symbol 'C' stands for gaining the "E.F. Codd" award at that year.

In general, we observe that there is a good match between the presence of a person in the skyline set and him/her gaining the "Codd" award, especially for those whose presence in the skyline set is significant. There are some notable exceptions to this pattern, concerning F. Bancilhon, N. Goodman, A. Halevy, W. Kim, R. Motwani, L. Raymond, and A. Swami who – according to this study – could be strong candidates for the next year's award. On the other hand, since the skyline set is a "multidimensional quantitative" assessment method, it is not able to capture cases such as R. Bayer, C. Mohan, B.G. Lindsay whose contributions in databases were not described in many articles so as to be caught by the selected indicators.

5.2. Skyline sets in the networking field

Here, we calculate the two skyline sets for the field of networking and show the skyline members in Tables 8 and 9. The first observation is that the two skyline sets differ significantly in size in their corresponding years, an observation that we made for the DB dataset as well. Similar to what we observed in the skyline set for the DB dataset, we see a changing skyline set as time passes, and new scientists e.g., H. Balakrishnan, S. Shenker enter the skyline set, and older ones e.g., L. Kleinrock depart from it. Again, some scientists are (almost) constantly present in it (e.g., S. Floyd). Therefore, the generic pattern that we encountered in the analysis of the skyline sets for the DB dataset is also found in the NET dataset, which means that the findings were not a result of biases present in our first dataset.

Proceeding to examine whether there is any correlation among an "ACM SIGCOMM Lifetime Contribution" award-winning scientist and its presence in the skyline set, we merged Tables 8 and 9 into Table 10, which includes all the networking researchers that won the "Lifetime Contribution" award and/or appeared in our skyline sets at least once. Each one of the table columns corresponds to a year after 1980. The year of the first award is 1989. The symbol 's1' shows that the corresponding researcher is in the skyline set with dimensions (*metric*2, *R*, *hrat*) and the symbol 's2' shows that the corresponding researcher is in the skyline set with dimensions (*PI*/*p*, *e*, *h*) during the specific year and the symbol *s*12 that s/he is in both. The symbol 'L' stands for gaining the "SIGCOMM Lifetime Contribution" award at that year. We observe a strong match between award

Table 7
s1:(*metric*2, *R*, *hrat*) and s2(*PI*/*p*, *e*, *h*) skyline members vs. "E.F. Codd Innovations" (C) award winners.

Scientists	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Abiteboul Serge																			s1,C	s1	s1	s1	s1	s1										
Agrawal Rakesh																	s12	s12	s12	s12	s12,C	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Babcock Brian																							s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Bancihlon Francois														s12	s2	s2	s12	s12	s12	s12	s12													
Bayer Rudolf																						C												
Beeri Catriel											s1	s1	s1	s1			s1	s1	s1	s1	s1	s1												
Bernstein Philip				s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12,C	s12	s12	s12	s2	s12														
Carey Michael																										C								
Ceri Stefano																																		C
Chamberlin Don	s12	s12	s12	s12	s12	s12	s12	s12																C										
Chaudhuri Surajit																															C			
Codd E.F.	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Datar Mayur																					s1	s1					s1	s1	s1	s1	s1	s1	s1	s1
Dayal Umeshwar																s1															C			
DeWitt David											s1		s1	s1	s1	s1,C	s1	s1	s1	s1				s1										
Fagin Ronald			s1	s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s1	s1	s1	s1,C	s1	s1	s1	s12	s12	s12	s12	s12	s12
Faloutsos Christos						s1														s1														
Garcia-Molina Hector																				C								s1	s1	s12	s12	s12	s12	s12
Goodman Nathan														s12		s2		s2		s2	s2													
Gray Jim														C																				
Halevy Alon																													s1		s1			
Hammer Michael											s1	s1	s1	s1	s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s1	s1	s1	s1
Imilienski Tomasz																			s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Kim Won																																		
Kitsuregawa Masaru																														C				
Lindsay Bruce																																		
Lorie Raymond		s1	s1						s1	s1	s1	s12				s1	s1	s12	s12	s12	s12	s12	s12	s12	s12	s2	s2	s2	s2	s2			C	
Maier David																		C																
Mohan C.																	C																	
Motwani Rajeev																										s1	s1	s1	s1	s1	s12	s12	s12	s12
Papadimitriou Chr.									s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1													
Papakonstantinou Y.															s12				s1	s1				s1										
Quass Dallan																		s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1				
Selinger Patricia										s1	s1	s1	s1										C											
Stonebraker Michael									s1	s1	s1	s1	s1,C	s1		s1		s1			s1	s1	s1											
Swami Arun														s1	s1										s1	s12	s12	s12	s12	s12	s12	s12	s12	s12
Ullman Jeffrey	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12,C	s12	s12	s12	s12	s12	s12	s12
Vardi Moshe																													C					
Widom Jennifer											s1								s1			s1	s1	s1	s1	s1	s1	s1,C	s1	s1	s1	s1	s1	s1

Table 8
Members of Skyline set calculated from NET dataset with attributes (*metric*2, *R*, *hrat*). Every row represents the skyline set of a particular year. The first column displays the year and the second one contains the names of the scientists belonging to the skyline set that particular year.

Year	Skyline members								
1992	Chandrakasan A., Willinger W.,	Eager D., Zhang L.	Gallager R.,	Guibas L.,	Hluchyj M.,	Karger D.,	Kleinrock L.,	Obraczka K.,	Tennenhouse D.,
1993	Badrinath B., Karger D.,	Bhagwat P., Kleinrock L.,	Culler D., Obraczka K.,	Eager D., Zhang L.,	Floyd S., Ramaswami R.	Gallager R.,	Guibas L.,	Hluchyj M.,	Jamin S.,
1994	Anderson T., Jamin S.,	Badrinath B., Kleinrock L.,	Bhagwat P., Maltz D.,	Culler D., Obraczka K.,	Floyd S., Ramaswami R.,	Gallager R., Ramjee R.,	Guibas L., Zahorjan J.,	Heidemann J., Zhang L.	Hluchyj M.,
1995	Floyd S., Shenker S., Anderson T.,	Gallager R., Stankovic J., Bhagwat P.,	Guibas L., Stoica I., Belding E.,	Hluchyj M., Viterbi A., Davie B.,	Imielinski T., Zahorjan J., Ferrari D.,	Jamin S., Zhang L., Floyd S.,	Kaashoek F., Want R. Gallager R.,	Kleinrock L., Guibas L.,	Ramaswami R., Hluchyj M.,
1996	Imielinski T., Tennenhouse D., Anderson T,.	Jacobson V., Turletti T., Berry R.,	Katz R., Viterbi A., Bhagwat P.,	Kleinrock L., Want R., Deering S.,	Ramaswami R., Zahorjan J., Floyd S.,	Ramjee R., Zhang L. Guibas L.,	Satyanarayanan M., Hluchyj M.,	Shenker S., Hui Z.,	Stankovic J., Imielinski T.,
1997	Jacobson V., Zahorjan J., Abdelzaher T.,	Kaashoek F., Zhang L. Anderson T.,	Katz R., Balakrishnan H.,	Kleinrock L., Barford P.,	McCanne S., Berry R.,	Ramjee R., Bhagwat P.,	Shenker S., Culler D.,	Stankovic J., Deering S.,	Starobinski D., Floyd S.,
1998	Gallager R., Kaashoek F., Want R., Anderson T.,	Grossglauser M., Katz R., Wetherall D., Balakrishnan H.,	Guibas L., Kleinrock L., Willinger W., Barford P.,	Hluchyj M., McCanne S., Zahorjan J., Belding E.,	Hu Y., Padmanabhan V., Zhang L. Culler D.,	Hui Z., Satyanarayanan M., Deering S.,	Imielinski T., Shenker S., Dovrolis C.,	Jacobson V., Stankovic J., Floyd S.,	Jamin S., Tennenhouse D., Grossglauser M.,
1999	Guibas L., Shenker S., Anderson T.,	Hluchyj M., Stankovic J., Balakrishnan H.,	Jacobson V., Wetherall D., Barford P.,	Katz R., Wroclawski J., Belding E.,	Kleinrock L., Zhang L. Bolot J.,	McCanne S., Crovella M.,	Padhye J., Culler D.,	Psounis K., Deering S.,	Seshan S., Dovrolis C.,
2000	Floyd S., Ramaswami R., Anderson T.,	Guibas L., Shenker S., Balakrishnan H.,	Hu Y., Taft N., Belding E.,	Hui Z., Viterbi A., Chandra R.,	Jacobson V., Wetherall D., Crovella M.,	Katz R., Zhang L. Culler D.,	Lee S., Fall K.,	McCanne S., Floyd S.,	Paxson V., Hui Z.,
2001	Jacobson V., Sanadidi M., Balakrishnan H.,	Katz R., Seshan S., Belding E.,	Lee S., Shenker S., Bolot J.,	Padhye J., Viterbi A., Culler D.,	Padmanabhan V., Zhang L. Deering S.,	Paxson V., Floyd S.,	Ramaswami R., Gribble S.,	Ratnasamy S., Hui Z.,	Rodriguez P., Jacobson V.,

2002	Jamin S., Ratnasamy S., Balakrishnan H.,	Katz R., Shenker S., Culler D.,	Maltz D., Viterbi A., Deering S.,	McCanne S., Wetherall D., Floyd S.,	Padhye J., Zhang L. Gribble S.,	Papadimitratos P., Hu Y.,	Papagiannaki K., Jacobson V.,	Paxson V., Maltz D.,	Ramaswami R., McCanne S.,
2003	Padhye J., Subramanian L.	Padmanabhan V.,	Papadimitratos P.,	Paxson V.,	Proutiere A.,	Ramaswami R.,	Ratnasamy S.,	Rodriguez P.,	Shenker S.,
2004	Balakrishnan H., Maltz D., Balakrishnan H.,	Culler D., McCanne S., Culler D.,	Deering S., Padhye J., Deering S.,	Floyd S., Papadimitratos P., Druschel P.,	Gribble S., Perrig A., Floyd S.,	Hu Y., Ramaswami R., Hu Y.,	Jacobson V., Ratnasamy S., Jacobson V.,	Jamin S., Shenker S. Jamin S.,	Levis P., Levis P.,
2005	Maltz D., Viterbi A.,	McCanne S., Wetherall D.	Padmanabhan V.,	Papadimitratos P.,	Paxson V.,	Perrig A.,	Ramaswami R.,	Ratnasamy S.,	Shenker S.,
2006	Culler D., Padmanabhan V.,	Deering S., Paxson V.,	Floyd S., Perrig A.,	Foster I., Ramaswami R.,	Hluchyj M., Ratnasamy S.,	Hu Y., Shenker S.,	Kelly F., Viterbi A.	Maltz D.,	McCanne S.,
2007	Culler D., Lu S., Balakrishnan H.,	Deering S., Maltz D., Culler D.,	Floyd S., McCanne S., Erceg V.,	Foster I., Padmanabhan V., Floyd S.,	Gallager R., Perrig A., Foster I.,	Gribble S., Ramaswami R., Gallager R.,	Hluchyj M., Ratnasamy S., Gribble S.,	Hu Y., Shenker S. Hluchyj M.,	Jamin S., Hu Y.,
2008	Jamin S., Shenker S.	Kelly F.,	Lu S.,	Maltz D.,	McCanne S.,	Padmanabhan V.,	Perrig A.,	Ramaswami R.,	Ratnasamy S.,
2009	Balakrishnan H., Hu Y.,	Bhagwat P., Jamin S.,	Culler D., Lu S.,	Deering S., Maltz D.,	Erceg V., McCanne S.,	Estrin D., Padmanabhan V.,	Floyd S., Ramaswami R.,	Foster I., Ratnasamy S.,	Gribble S., Shenker S.
2010	Balakrishnan H., Jamin S.,	Bhagwat P., Lu S.,	Culler D., Maltz D.,	Erceg V., McCanne S.,	Estrin D., Padmanabhan V.,	Floyd S., Ramaswami R.,	Foster I., Ratnasamy S.,	Gallager R., Shenker S.	Gribble S.,
2011	Balakrishnan H., Jamin S.,	Bhagwat P., Lu S.,	Culler D., Maltz D.,	Erceg V., McCanne S.,	Estrin D., Padmanabhan V.,	Floyd S., Ramaswami R.,	Foster I., Ratnasamy S.,	Gallager R., Shenker S.	Gribble S.,
2012	Balakrishnan H., Jamin S.,	Bhagwat P., Lu S.,	Culler D., Maltz D.,	Erceg V., McCanne S.,	Estrin D., Padmanabhan V.,	Floyd S. Ramaswami R.,	Foster I., Ratnasamy S.,	Gallager R., Shenker S.	Gribble S.,
2013	Balakrishnan H., Jamin S.,	Bhagwat P., Lu S.,	Culler D., Maltz D.,	Erceg V., McCanne S.,	Estrin D., Padmanabhan V.,	Floyd S., Ramaswami R.,	Foster I., Ratnasamy S.,	Gallager R., Shenker S.	Gribble S.,

Table 9
Members of Skyline set calculated from NET dataset with attributes (*PI*/*papers*, *e*, *h*). Every row represents the skyline set of a particular year. The first column displays the year and the second one contains the names of the scientists belonging to the skyline set that particular year.

Year	Skyline members								
1992	Cheriton D., Gallager R.,	Deering S.,	Guibas L.,	Hluchyj M.,	Karger D.,	Keshav S.,	Kleinrock L.,	Tennenhouse D.,	Want R.,
1993	Cheriton D.,	Floyd S.,	Gallager R.,	Guibas L.,	Kleinrock D.				
1994	Bhagwat P.,	Cheriton D.,	Floyd S.,	Gallager R.,	Guibas L.,	Kleinrock L.			
1995	Bhagwat P.,	Cheriton D.,	Floyd S.,	Gallager R.,	Guibas L.,	Hluchyj M.,	Kleinrock L.,	Shenker S.	
1996	Floyd S.,	Gallager R.,	Guibas L.,	Jacobson V.,	Katz R.,	Kleinrock L.,	Satyanarayanan M.,	Shenker S.	
1997	Floyd S.,	Guibas L.,	Jacobson V.,	Katz R.,	Kleinrock L.,	Shenker S.,	Stankovic J.,	Zhang L.	
1998	Floyd S.,	Guibas L.,	Hu Y.,	Jacobson V.,	Katz R.,	Kleinrock L.,	Shenker S.,	Stankovic J.,	Zhang L.
1999	Belding E.,	Deering S.,	Floyd S.,	Guibas L.,	Hu Y.,	Jacobson V.,	Katz R.,	Shenker S.,	Zhang L.
2000	Belding E., Shenker S.,	Deering S., Zhang L.	Floyd S.,	Guibas L.,	Hu Y.,	Jacobson V.,	Katz R.,	Ratnasamy S.,	
2001	Belding E.,	Floyd S.,	Jacobson V.,	Katz R.,	Ratnasamy S.,	Shenker S.			
2002	Floyd S.,	Jacobson V.,	Katz R.,	Ratnasamy S.,	Shenker S.				
2003	Floyd S.,	Maltz R.,	Ratnasamy S.,	Shenker S.					
2004	Floyd S.,	Maltz R.,	Ratnasamy S.,	Shenker S.					
2005	Floyd S.,	Maltz R.,	Ratnasamy S.,	Shenker S.					
2006	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Maltz D.,	Shenker S.			
2007	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Maltz D.,	Shenker S.			
2008	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Maltz D.,	Shenker S.		
2009	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Maltz D.,	Ratnasamy S.,	Shenker S.	
2010	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Ratnasamy S.,	Shenker S.		
2011	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Shenker S.			
2012	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Shenker S.			
2013	Balakrishnan H.,	Culler D.,	Floyd S.,	Foster I.,	Jacobson V.,	Shenker S.			

winners and presence in the skyline set(s) having for instance S. Floyd, V. Jacobson, L. Kleinrock, D. Cheriton, J. Postel to be award winners with significant (or omni) presence in the skyline set(s). At a first glance it may seem that D. Culler, L. Guibas and H. Balakrishnan are notable exceptions to this 'rule', but this is valid only for D. Culler since L. Guibas is a Computational Geometry person and not a Networking/Communications person even though his field has found broad applications in problems related to ad hoc networks, and H. Balakrishnan is a young scientist. On the other hand, a solely quantitative method such as the skyline operator based on the particular dimensions selected in this study was not able to identify cases such as P. Baran whose work on packet switching dates back to 1960, whereas our analysis focuses on the years 1980–2013 and it could be possible that scientists who peaked before this time period may have been surpassed by younger ones after 1980, thus not managing to enter our skyline set. There also the cases of as V. Cerf, P. Mockapetris who did not follow academic careers with extensive paper writing. All in all, for the networking data, the skyline set is significantly consistent with the award decisions.

5.3. Skyline sets and the Turing award winners

Finally, we examined the overlap between the set of skyline members and the set of those who have won a Turing award and display the results in Table 11. We have not included some winners for whom sufficient data do not exist. We observe that scientists such as L. Lamport, D. Rivest, S. Floyd, B. Liskov, R. Tarjan, L. Valiant, B. Lampson, R. Milner, A. Shamir, R. Simon and others with significant presence in the skyline set(s) have won the Turing award. Other scientists such L. Zadeh, J. Ullman, M. Garey, S. Ramakrishnan even though they appear many times in the skyline set(s) they still are not Turing winners. Moreover, there are cases such as A. Yao (theory), K. Thompson and D. Ritchie (UNIX, C), S. Micali (cryptography), J. Hartmanis (computational complexity) who are Turing winners without presence in the skyline set(s); this is partially attributed to the nature of the achievement that led to the award (e.g., development of UNIX/C by Thompson and Ritchie), or to the extent of our data to the past (e.g., J. Hartmanis), without of course excluding the case that the award committee has considered purely qualitative criteria in the decision.

Table 10
s1:(*metric*2, *R*, *hrat*) and s2:(*PI*/*p*,*e*,*h*) skyline members vs. "ACM SIGCOMM Lifetime Contribution" (L) award winners.

Scientists	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Anderson Thomas															s1		s1	s1	s1	s1	s1	s1												
Balakrishan Hari																			s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s12	s12	s12	s12	s12	s12
Baran Paul										L																								
Cerf Vint																	L																	
Cheriton David											s2	s2	s2	s2	s2	s2								L										
Clark David											L																							
Crowcroft Jon																														L				
Culler David														s1	s1				s1	s1	s1	s1	s1	s1	s1	s1	s12	s12	s12	s12	s12	s12	s12	s12	s12
Danthine Andre																					L													
Doyle John			s1	s1	s1	s1	s1																											
Eager Derek			s1				s1	s1	s1	s1	s1	s1	s1	s1																				
Estrin Deborah																														s1	s1			
Farber David																L																		
Ferrari Domenico																	s1										L							
Floyd Sally														s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12,L	s12	s12	s12	s12	s12	s12
Foster Ian																											s12	s12	s12	s12	s12	s12	s12	s12
Fraser Sandy													L																					
Gallager Robert	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12		s1									s1	s1	s1	s1	s1	s1	s1
Green Paul															L																			
Guibas Leonidas		s1	s1					s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12													
Jacobson Van									s12	s2							s12	s12	s12	s12	s12	s12,L	s2	s2	s2	s2	s2	s2	s2	s2	s2	s2	s2	s2
Kleinrock Leonard	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12L	s12	s12	s12	s12	s12	s12	s12	s12	s12														
Kashoek Frans																s1		s1	s1															
Kahn Robert														L																				
Katz Randy																	s12	s12	s12	s12	s12	s12	s12											
Kirstein Peter																				L														
Lam Simon																									L									
McKeown Nick																																	L	
Mockapetris Paul																										L								
Paxson Vern																				s1	s1	s1	s1		s1	s1						L		
Perlman Radia																															L			
Peterson Larry																																		L
Postel Jon	s1	s12	s12	s12	s1	s12	s1			s1		s1						L																
Pouzin Luis																		L																
Ratnasamy Sylvia																				s1	s1	s12	s12	s1	s1	s1	s1	s1	s1	s12	s12	s1	s1	s1
Roberts Larry																			L															
Shenker Scott																s12	s12	s12	s12	s12	s12	s12	s12,L	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Stankovic Jack								s1	s1	s1	s1	s1				s1	s1	s12	s12	s1,s3														
Towsley Donald				s1		s1																							L					
Vetterli Martin						s1																												
Viterbi Andrew							s1									s1	s1				s1	s1	s1			s1	s1							
Zahorjan John							s1					s12			s1	s1	s1	s1	s1															
Zhang Lixia										s1	s1	s1	s1	s1	s1	s1	s12	s12	s12	s12	s12	s1	s1											
Zimmerman Hubert												L																						

Table 11
s1:(*metric*2, *R*, *hrat*) and s2(*PI*/*p*, *e*, *h*) skyline members vs. Turing (T) award winners.

Scientists	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Agrawal Rakesh																					s1		s12		s2	s12	s12	s12	s12	s12	s12	s2	s2	s2
Aho Alfred							s1	s12	s12	s12	s12	s12	s12	s1	s1	s12	s12	s12	s12	s1	s1													
Allen Fred																											T							
Balakrishnan Hari																									s1				s12	s12	s12	s12	s12	s12
Banerjee Uptal															s1	s1	s1	s1			s1	s1												
Cardwell Neal																						s1	s1								s1	s1	s1	s1
Cocke John								T																										
Codd Edgar	s12	s12,T	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s2																
Corbato Fernado											T																							
Feigenbaum Edward															T																			
Floyd Sally															s1,T								s1	s1	s1	s12	s1	s12	s12	s12	s12	s12	s12	s12
Foster Ian																												s12	s12	s12	s12	s12	s12	s12
Garey Michael	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s1	s1	s1	s1	s1	s1	s1
Gean Stuart					s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s2	s2	s2	s2	s2	s2		s2	s2	s2								
Goldberg Daniel										s12	s2					s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s1	s1	s1	s1
Goldwasser Shafi																																	T	
Hartmanis Juris														T																				
Haussler David																							s1	s1	s1	s1	s12	s12	s12	s12	s12	s12	s12	s12
Higgins Desmond																								s2	s2	s2	s2	s2	s2	s12	s12	s12	s12	s12
Hinton Geoffrey							s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s1	s12	s12	s12	s1	s1	s1						
Hoare Ron	T					s2	s2	s12	s12	s12	s12	s12	s12	s12	s12																				
Jacobson Van																						s2	s12	s12	s12	s12								
Kahan William										T																								
Kahn Robert																									T									
Karp Richard	s1	s1	s1	s1	s1	s1,T	s1	s1								s1	s1	s1	s1	s1	s1													
Kennedy James																s1	s12	s1	s1											s12	s2	s12	s12	s12
Knuth D. (awarded '74)						s2	s2	s2			s2	s2	s2	s2																				
Kohonen Teuvo										s1												s12	s1		s1	s12	s12	s12	s12			s12	s12	s12
Lamport Leslie			s1	s12	s12	s1	s1	s1	s1	s1	s1	s1	s1						s12	s12	s12	s12	s12	s1										T
Lampson Butler							s1	s1					T																					
Lewis Philip	s1	s1	s1			s1	s1	s1																										
Liskov Barbara							s1	s1	s1	s1																			T					
Masinter L.						s1	s1	s1		s1		s1		s1			s1																	
McCreight Edward			s1	s1	s1	s1	s1	s1	s1	s1			s1	s1	s1	s1	s1	s1	s1				s1											
Micali Silvio																																	T	

Milner Robin						s1						T		s1	s1	s1	s1	s1	s1	s12	s12	s12	s12	s12	s12											
Nygaard Kristen																						T													
Oppen Derek																									s1		s1			s1	s1	s1	s1	s1	s1
Parekh Abhay																			s1	s1	s1		s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Pearl Judea																																	T			
Pnueli Amir	s1	s1												s1	s1	s1	s1,T	s1	s1	s1	s1	s1	s1												
Quass Dallan																						s1	s1	s1	s1	s1	s1								
Quinlan Jason														s1	s1	s12	s1	s12	s12	s12	s12	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Rabin M.(awarded '76)						s1																														
Ritchie Dennis				T																															
Rivest Donald	s1	s1	s1	s1	s1	s1	s1				s2								s12			s12	s12,T	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Sleator Daniel				s1					s1	s1	s1	s1		s1	s1	s1	s1	s1	s1						s1	s1	s1					s1	s1	s1	
Schapire Robert																				s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Schunck Brian		s12	s12	s12	s12	s12			s2																											
Shamir Adi	s2	s2								s2													T													
Shenker Scott																								s1	s1	s1	s1	s12	s12	s12	s12	s12	s12	s12	s12
Simon R.(awarded '75)	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12														s2	s2	s2	s2	
Sinha Saurabh	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12																									
Srikant Ramakrishnan															s1	s1	s12	s12	s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12		
Stearns Richard		s1	s1	s1										T																					
Sugeno M.																				s1	s1		s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Sutherland Ivan																																				
Tarjan R.	s1	s1	s1	s12	s12	s12	s1,T	s1	s1	s1	s1	s12	s12	s12	s12	s12	s12	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1					
Thomspon Ken				T																															
Traiger Irving		s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1																		
Ullman Jeffrey	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s2	s2	s2	s2	
Valiant Leslie			s1	s1		s1	s1	s1	s1	s1	s1	s12	s1	s1	s1	s1	s1	s1	s12	s1	s1	s1	s1								T				
Vapnik Vladimir	s1	s1	s1	s1	s1																			s1	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	
Vlissides John															s12	s12	s12	s12	s12																
Wegner Emanual											s12	s12	s12	s12	s12	s12	s2	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12
Wirth Niklaus					T		s1																												
Witkin Andrew									s1	s1	s1	s1	s1	s1	s1	s1	s1				s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1	s1
Wolper Pierre			s1						s1								s1							s1	s1	s1									
Yao Andrew																				T															
Zadeh Lotfi	s12	s12	s12	s12	s12	s12	s12	s12	s12	s12	s2	s2	s2	s2	s2	s12	s12	s12	s1	s12	s12	s1	s1												
Zamir Oren																													s1	s1	s1	s1	s1	s1	s1

Table 12
The indices constituting each of the 8 combinations used in our experiments.

	Indices
Combination 1(C1)	mBor-hw-g
Combination 2(C2)	mBor-x-hg
Combination 3(C3)	hnor-AR-hm
Combination 4(C4)	hnor-R-g
Combination 5(C5)	v-C-h2
Combination 6(C6)	v-hI-f
Combination 7(C7)	PI/p-C-hrat
Combination 8(C8)	PI/p-s-p

Table 13
Percentage differences in skyline memberships averaged over all years for each pair of combinations.

	C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C6 (%)	C7 (%)	C8 (%)
C1	0	29	10	25	32	16	43	23
C2	43	33	13	24	23	20	0	24
C3	25	25	15	0	18	20	24	13
C4	29	0	18	25	27	31	33	41
C5	32	27	17	18	0	17	23	21
C6	16	31	10	20	24	0	20	10
C7	10	18	0	15	17	10	13	19
C8	23	41	19	13	21	10	24	0

5.4. Skyline sets with different combinations of indices

In the above subsections we have presented the analytical results for the experiments conducted with two chosen combinations of indices, (*metric*2, *R*, *hrat*) and (*PI*/*p*, *e*, *h*). To investigate the value of the skyline operator, we conducted several more experiments with different index combinations identifying their similarities and deviations. Since all the possible combinations given the 3 clusters of indices are numerous, we present for brevity reasons the comparisons between 8 different combinations choosing again one index from each cluster. The CS dataset was chosen for these additional experiments, as it is the largest one and the most diverse one. All the tested combinations ended up producing a distinguishable skyline set, but differences in the size of the resulting skyline sets were identified. Table 12 contains the indexes corresponding to each combination, while Fig. 6 presents the number of skyline members for each of the 8 combinations over all given years. We notice that all combinations produce a skyline set consisting of 20–60 scientists depending on the particular characteristics of the chosen indices and the size of the dataset for each given year. As the years go by, the datasets become richer and more scientists manage to increase their impact thus ending up belonging to the distinguished set. Therefore, in recent years we ended up having more densely populated skyline sets.

Next, we proceed to further investigate the different memberships in the skyline sets produced by different combinations of indices. To this end, we compare the aforementioned 8 combinations of indices in pairs. Given *combination1* and *combination2* we calculate how many differences occur between the skyline sets produced by the two combinations for each year. The differences are subsequently normalized by dividing them with the size of the largest of the two skyline sets (formulated using *combination1* and *combination2* respectively). From these pairwise comparisons, a set of 34 normalized differences is produced corresponding to the number of years we contemplated. These differences are averaged over all years and the resulting scores are displayed as percentages in Table 13.

As can be seen in Table 13, most combinations of indices produce skyline sets that differ about 25% from the rest of the combinations. That is to be expected since, even though the indices come from the same cluster, they tend to emphasize on slightly different author characteristics. As a result, they end up distinguishing a common core of scientists as skyline members, but each different index adds to this core a number of also distinguished scientists according to the particular criteria each index introduces. Some higher percentages (i.e., 43%) are explained based on the lower correlation of those indices (i.e., mBor) with the rest of the indices and their particular nature which is to be complementary metrics and not standalone indices.

Now that we have investigated the skyline operator's distinguishing power in identifying award winning scientists under various combinations of indices, we proceed to explore its performance in data sets with scientists of average scientific impact and more similar scientometric data.

5.5. Skyline sets in moderate citation count groups

In the aforementioned experiments we employed groups of award winning scientists as a validation set to compare with the scientists distinguished by the skyline operator. Our findings lead us to conclude that the skyline operator could be of assistance to award giving committees. Other cases however, where the skyline operator could provide useful insights as

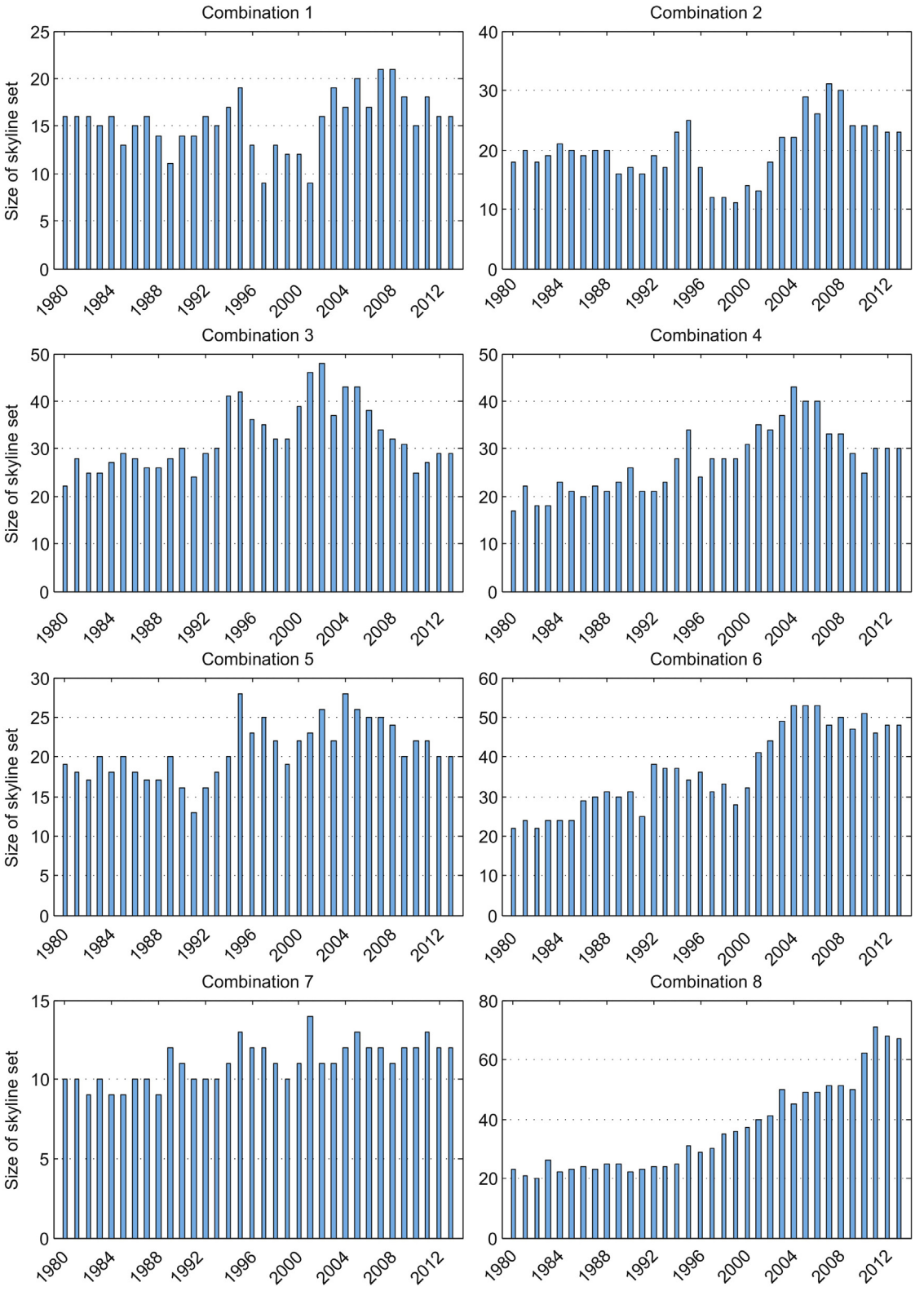

Fig. 6. Number of scientists belonging to the skyline sets for each of the 8 combinations of indices over years 1980–2013.

Table 14
Size of datasets for years 2003–2013 sampled from the CS dataset under the constraint of citation count being between 1500 and 2000.

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Size of dataset	1397	1638	1851	2048	2238	2517	2777	2947	3040	3056	3065

stated in Section 5, include grant allocation, tenure committees, etc. In these scenarios, the datasets are usually smaller in size than the ones utilized in our previous experiments and contain scientists of moderate citation count with analogous scientometric data. However, in such cases it is difficult to evaluate the resulting skyline sets without a priori knowledge of the particular features of the dataset and the specific purposes of the skyline operator's application. Therefore, we explore the skyline operator's ability to identify promising scientists from set of mediocre size and moderate citation count based on selected features. We chose a sample of our CS dataset for years 2003–2013 based on citation range, since citation count is a "raw" scientometric index, meaning it introduces limited bias. For each of these years the scientists whose citations range between 1500 and 2000 were selected. The sizes of the resulting samples for each year are displayed in Table 14. This range of citations was chosen because it results in relatively small datasets (1000–3000 scientists) as is often the case in the above mentioned scenarios and also represents a moderate performance level.

The skyline operator was applied on all 11 datasets of Table 14 for 10 different combinations of indices: the 8 combinations (C1–C8) from Section 5.4 and the 2 combinations from Section 5 (*metric*2, *R*, *hrat* and *PI*/*p*, *e*, *h* denoted as C9 and C10 respectively). In Fig. 7 the resulting sizes of the skyline sets are displayed for all 10 combinations. It can be observed that the sizes of the skyline sets are significantly smaller than the ones in the previous experiments which is to be expected, since these datasets are significantly smaller. However, through this additional experimentation it is indicated that a distinguishable skyline set can always be identified from groups of scientists of various performance levels using different combinations of indices dependent on the specific task. Even when the scientists present similar scientometric data, appropriate indexes can be chosen to produce the desirable set of distinguished scientists, who need not necessarily be of extraordinary impact (i.e., Turing award winners).

Therefore, the skyline operator can be considered a flexible tool able to assist various decision making processes. In any case, based on the task, we trust that the decision makers will be able to choose an appropriate set of indices to represent the qualities they want to focus on. Our proposed combinations (most representative indices, most popular indices, etc.) can offer broad and diverse criteria for scientific assessment at all levels.

6. Discussion

The evaluation of an individual's scientific work is a complex, multi-parametric task that involves also qualities which are not measurable by numbers. Nevertheless, a careful analysis of the scientometric data – as this is done via the rich ecosystem of the scientometric indicators – offers significant insights into a scientist's work. In this article, we attempted to group various indices available in literature into a small number of meaningful clusters; highly correlated indices have been placed in the same cluster. Choosing at least one representative index from each cluster allows us to characterize scientific productivity and impact based on broad and objective criteria. Numerous efforts have been made to achieve a universal grouping of indexes (Schreiber, Malesios, & Psarakis, 2012; Wildgaard et al., 2014) and discover their underlying correlations (Bornmann et al., 2011). However, in the present work, inspired by the areas of the citation curve as defined in Ye and Rousseau (2009), we investigated a potential clustering of author-level scientometric indexes with respect to their association with the areas and axes of the citation curve. Following an extensive literature review a set of 38 scientometric indices were chosen focusing on publication count, citation count as well as normalization over number of authors and time. Utilizing Principal Component Analysis these indices were grouped in 3 clusters that have not appeared in any previous studies. Each cluster accounts for different features of a scientist's impact as expressed by their citation curves.

Since the number of clusters is small and the inter-cluster correlations strong (Spearman $\rho > 0.6$), choosing one index from each of those groups provides a representative set of author characteristics, which we employed as features in the *skyline operator*. The use of the *skyline operator* allows for the distinction of scientists whose performance cannot be surpassed by others' with respect to the indexes selected as features. Through the experiments conducted, the operator's effectiveness was tested using various index combinations and concluded that a distinguishable skyline set can be produced for different time periods, disciplines and combinations of indices. Efforts to distinguish scientists from a large set of scientometric data have been made before either using a geometric combination of indices (Revesz, 2014), adding the element of time in the scientometric indicators (Wolcott et al., 2015) or utilizing probability models (Vieira, Cabral, & Gomes, 2014). However, these models often present limitations, meaning they cannot be automatically applied to a variety of datasets under a customizable number of criteria. The proposed methodology of the *skyline operator* offers a straightforward, easily applicable and versatile methodology that can also be extended to publication, publishing venues and even institution level to distinguish a set of highly qualified points in a dataset based on a multitude of criteria.

The skyline members identified in our study proved to be award-winning scientists, a fact implying that the skyline operator could potentially be used by award committees to assess scientific excellence. Similarly, the proposed methodology can be utilized for grant allocation, tenure positions and other purposes requiring the distinction according to a set of numerical criteria. As was indicated by our experiments, different combinations of indices produce either a broader or a

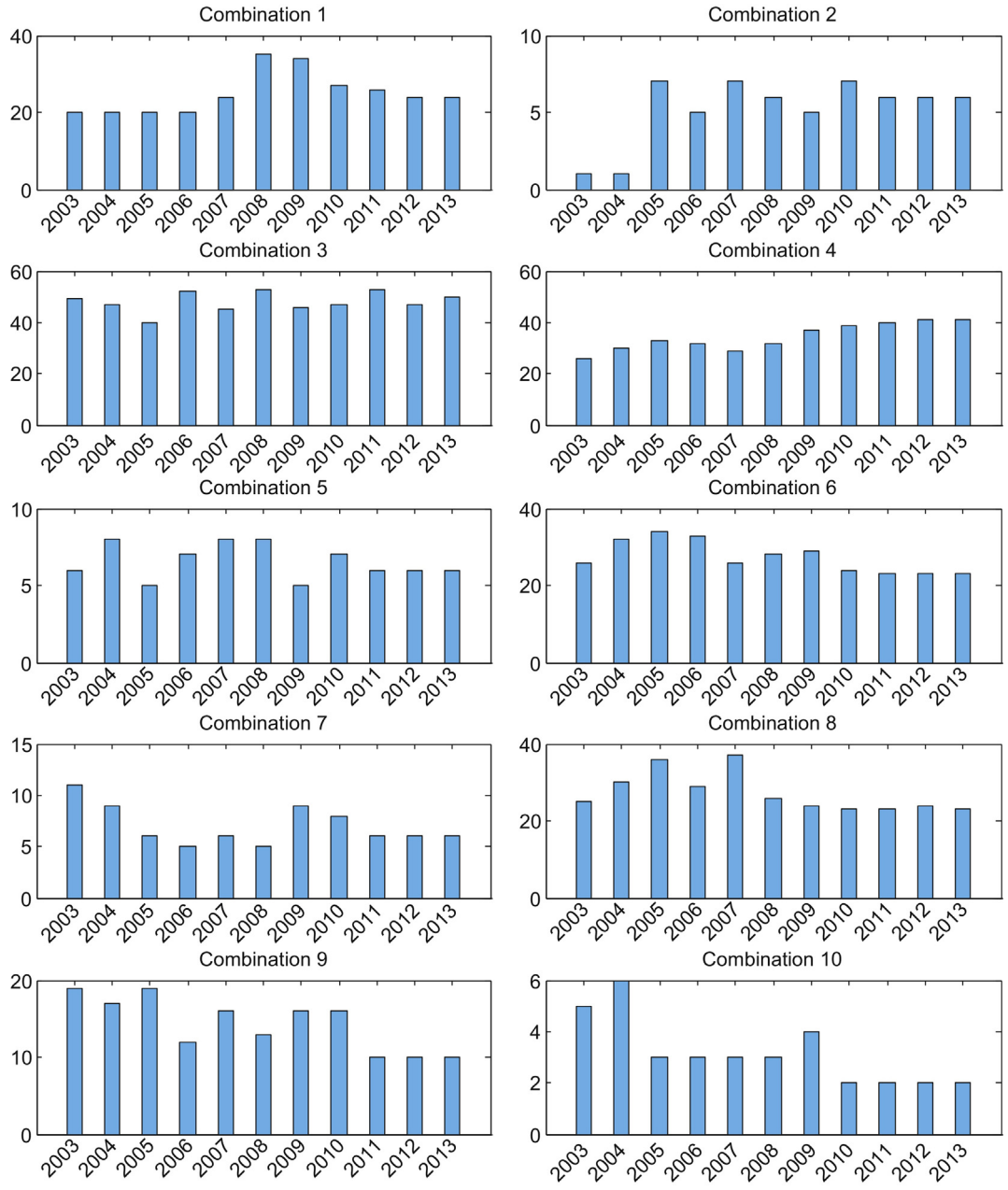

Fig. 7. Number of scientists belonging to the skyline sets for each of the 10 combinations of indices over years 2003–2013 for the moderate citation group.

narrower set of top scientists. The number of authors distinguished and their level of scientific impact varies with respect to the characteristics of the data set. However, due to brevity reasons the investigation of all possible combinations of indices and all available variations of the basic *skyline operator* were not explored under the scope of the present work. In this article, we outlined the usefulness and adaptability of the *skyline operator*, when used together with a set of appropriately selected features, to automatically perform multi-criteria distinction for the evaluation of scientific impact.

All in all, the methodology proposed utilizes the clusters of indices together with the skyline operator in order to distinguish highly accomplished scientists. This approach ensures two basic properties; firstly, that a pool of highly accomplished individuals can be automatically extracted from a large set of data and secondly, that these individuals will be judged based on various aspects of their publishing behavior and impact due to the use of indices from different clusters focusing on particular features of scientific impact.

Authors' contribution

Conceived and designed the analysis: A. Sidiropoulos, A. Gogoglou, D. Katsaros and Y. Manolopoulos.
Collected the data: A. Sidiropoulos, A. Gogoglou, D. Katsaros and Y. Manolopoulos.
Contributed data or analysis tools: A. Sidiropoulos, A. Gogoglou, D. Katsaros and Y. Manolopoulos.
Performed the analysis: A. Sidiropoulos, A. Gogoglou, D. Katsaros and Y. Manolopoulos.
Wrote the paper: A. Sidiropoulos, A. Gogoglou, D. Katsaros and Y. Manolopoulos.

References

Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). *h*-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.
Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2010). *hg*-Index: A new index to characterize the scientific output of researchers based on the *h*- and *g*-indices. *Scientometrics*, *82*(2), 391–400.
Anderson, T. R., Hankin, R. K. S., & Killworth, P. D. (2008). Beyond the Durfee square: Enhancing the *h*-index to score total publication output. *Scientometrics*, *76*(3), 577–588.
Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*(1), 179–189.
Bollen, J., van de Sompel, H. A. H., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS One*, *4*(6), e6022.
Bornmann, L., Mutz, R., & Daniel, H. D. (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.
Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-P. (2011). A multilevel meta-analysis of studies reporting correlations between the *h*-index and 37 different *h*-index variants. *Journal of Informetrics*, *5*(3), 346–359.
Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-P. (2014). *h*-Index research in scientometrics: A summary. *Journal of Informetrics*, *8*(3), 749–750.
Börzsönyi, S., Kossmann, D., & Stocker, K. (2001). The skyline operator. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)* (pp. 421–430).
Cabrerizo, F., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). q^2-Index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, *4*(1), 23–28.
Chan, C.-Y., Jagadish, H. V., Tan, K.-L., Tung, A. K. H., & Zhang, Z. (2006). Finding *k*-dominant skylines in high dimensional space. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)* (pp. 503–514).
Chomicki, J., Ciaccia, P., & Meneghetti, N. (2013). Skyline queries, front and back. *ACM SIGMOD Record*, *42*(3), 6–18.
Chomicki, J., Godfrey, P., Gryz, J., & Liang, D. (2003). Skyline with presorting. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)* (pp. 717–719).
Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics*, *69*(1), 131–152.
Egghe, L., & Rousseau, R. (2008). An *h*-index weighted by citation impact. *Information Processing and Management*, *44*(2), 770–780.
Eom, H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE*, *6*(9), e24926.
Fagin, R., Lotem, A., & Naor, M. (2001). Optimal aggregation algorithms for middleware. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)* (pp. 102–113).
Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111.
Godfrey, P., Shipley, R., & Gryz, J. (2007). Algorithms and analyses for maximal vector computation. *The VLDB Journal*, *16*(1), 5–28.
Gogoglou, A., Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2015). Bibliometric indices for the assessment of the citation curve tail. In *Proceedings of the Panhellenic Conference on Informatics (PCI)* (pp. 305–310).
Gupta, H. M., Campanha, J. R., & Pesce, R. A. G. (2005). Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, *35*(12), 981–986.
Harzing, A. W. (2007). *Reflections on norms for the h-index and related indices*.. Available at: http://www.harzing.com/pop_norm.htm
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.
Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The *r*- and *ar*-indices: Complementing the *h*-index. *Chinese Science Bulletin*, *52*(6), 855–863.
Jin, W., Han, J., & Ester, M. (2004). Mining thick skylines over large databases. In *Proceedings of the International Conference on Knowledge Discovery in Databases (PKDD)* (pp. 255–266).
Katsaros, D., Akritidis, L., & Bozanis, P. (2009). The *f* index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science & Technology*, *60*(5), 1051–1056.
Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original *h*-index. *ISSI Newsletter*, *2*(3), 4–6.
Kosmulski, M. (2007). MAXPROD – A new index for assessment of the scientific output of an individual, and a comparison with the *h*-index. *Cybernetics*, *11*(1).
Kung, H. T., Luccio, F., & Preparata, F. P. (1975). On finding the maxima of a set of vectors. *Journal of the ACM*, *22*(4), 469–476.
Langville, A. N., & Meyer, C. D. (2014). *Who's #1?: The science of rating and ranking*. Princeton University Press.
Lee, J., You, G.-W., & Hwang, S.-W. (2009). Personalized top-*k* skyline queries in high-dimensional space. *Information Systems*, *34*(1), 45–61.
Magnani, M., Assent, I., & Mortensen, M. L. (2014). Taking the big picture: representative skylines based on significance and diversity. *The VLDB Journal*, *23*(5), 795–815.
Miller, C. W. (2006). *Superiority of the h-index over the Impact Factor for physics*.. Available at: http://arxiv.org/pdf/physics/0608183.pdf
Prathap, G. (2010). Is there a place for a mock *h*-index? *Scientometrics*, *84*(1), 153–165.
Rahm, E., & Thor, A. (2005). Citation analysis of database publications. *ACM SIGMOD Record*, *34*(4), 48–53.
Revesz, P. Z. (2014). Data mining citation databases: A new index measure that predicts nobel prizewinners. In *Proceedings of the International Database Engineering & Applications Symposium (IDEAS)* (pp. 1–9).
Riikonen, P., & Vihinen, M. (2008). National research contributions: A case study on Finnish biomedical research. *Scientometrics*, *77*(2), 207–222.
Rousseau, S., & Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics*, *40*(1), 45–56.
Ruane, F., & Tol, R. (2008). Rational (successive) *h*-indices: An application to economics in the republic of Ireland. *Scientometrics*, *75*(2), 395–405.
Schreiber, M. (2008). To share the fame in a fair way, h_m modifies *h* for multi-authored manuscripts. *New Journal of Physics*, *10*(4).
Schreiber, M., Malesios, C., & Psarakis, S. (2012). Exploratory factor analysis for the Hirsch index, 17 *h*-type variants, and some traditional bibliometric indicators. *Journal of Informetrics*, *6*(3), 347–358.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch *h*-index for disclosing latent facts in citation networks. *Scientometrics*, *72*(2), 253–280.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2015). Identification of influential scientists vs. mass producers by the Perfectionism index. *Scientometrics*, *103*(1), 1–31.
Silagadze, Z. K. (2010). Citation entropy and research impact estimation. *Acta Physica Polonica B*, *41*(11), 2325–2333.
Tiakas, E., Papadopoulos, A., & Manolopoulos, Y. (2015). Skyline queries: An introduction. In *Proceedings of the 6th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1–6).

Tol, R. (2009). The *h*-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, *80*(2), 317–324.
Tsai, C.-F. (2014). Citation impact analysis of top ranked computer science journals and their rankings. *Journal of Informetrics*, *8*(2), 318–328.
Vieira, E. S., Cabral, J. A., & Gomes, J. A. (2014). How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *Journal of Informetrics*, *8*(2), 390–405.
Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations. *Journal of the American Society for Information Science & Technology*, *62*(10), 1963–1978.
Voorneveld, M. (2003). Characterization of Pareto dominance. *Operations Research Letters*, *31*(1), 7–11.
Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, *101*(1), 125–158.
Wohlin, C. (2009). A new index for the citation curve of researchers. *Scientometrics*, *81*(2), 521–533.
Wolcott, H. N., Fouch, M. J., Hsu, E., Bernaciak, C., Corrigan, J., & Williams, D. (2015). Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers. In *Proceedings of the International Society of Scientometrics & Informetrics Conference (ISSI)* (pp. 403–408).
Wu, Q. (2010). The *w*-index: A measure to assess scientific impact by focusing on widely cited papers. *Journal of the American Society for Information Science & Technology*, *61*(3), 609–614.
Ye, F. Y., & Rousseau, R. (2009). Probing the *h*-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics*, *84*(2), 431–439.
Zhang, C.-T. (2009). The *e*-index, complementing the *h*-index for excess citations. *PLoS One*, *4*(5), e5429.