A Scientist's Impact over Time: The Predictive Power of Clustering with Peers

Antonia Gogoglou
Department of Informatics
Aristotle University of
Thessaloniki, Greece
agogoglou@csd.auth.gr

Antonis Sidiropoulos
Department of Information
Technology
Alexander Technological
Educational Institute of
Thessaloniki, Greece
asidirop@csd.auth.gr

Dimitrios Katsaros
Department of Electrical &
Computer Engineering
University of Thessaly,
Greece
dkatsar@inf.uth.gr

Yannis Manolopoulos
Department of Informatics
Aristotle University of
Thessaloniki, Greece
manolopo@csd.auth.gr

ABSTRACT

The identification of latent patterns in big scholarly data that concern the performance of researchers is a significant task because it can potentially impact scientific careers since they are based in funding and promotion. This article investigates the temporal evolution of a scientist's impact. Instead of taking a detailed, microscopic view that examines the citation curves of every scientist's article, the article develops a scalable, macroscopic methodology that uses the articles' citation profiles to build a more abstract and high-level profile that characterizes a scientist. This profile is utilized to cluster scientists in a set of 'performance' clusters. To this end, established techniques such as Principal Component Analysis and Self-Organizing Map clustering are employed as well as a set of proposed heuristics. The effectiveness of the proposed methodology is examined by comparing the resulting rankings with the outcomes of the peer-review procedures that resulted in the E. F. Codd and the Turing awards. The good match between the outcomes of computerized and peer-review procedures provides solid evidence that the proposed techniques constitute a promising analysis method for big scholarly data.

IDEAS '16, July 11-13, 2016, Montreal, QC, Canada

DOI: http://dx.doi.org/10.1145/2938503.2938523

Keywords

h-index, perfectionism index, principal component analysis, self-organizing map, clustering, career path.

1. INTRODUCTION

The abundance of bibliometric data related to the citations amongst articles, which are now available by modern, online sources such as Microsoft's Academic Search[1], Google Scholar[2], Elsevier's Scopus[3], comprise a rich source of big data for analysis and modeling to detect interesting patterns concerning a scientist's performance, an article's citation curve, or an institution's rank. Among the variety of these tasks, the recognition and ranking of individuals has stimulated a lot of research in the field of scientometric analysis. Various indexes have been introduced in the past that attempt to quantify this performance. For instance, the h-index [4] is a proxy for productivity and impact, the e-index [15] complements the h-index for the ignored excess citations, the contemporary h-index [13] detects the young-promising star scientists by estimating the currency of articles that comprise the h-index, the f index [5] characterizes the interdisciplinary nature of a scientist's work, whereas the perfectionism index [14] finds those laconic and high impact scientists.

These bibliometric indices have been utilized in creating methodologies to evaluate individual scientists, journals and academic institutions. Two important aspects in the assessment of scientific output that have recently attracted significant attention are the element of time [2, 6, 11] and the predictive power of sciento-

[1] http://academic.research.microsoft.com/

[2] https://scholar.google.com/

[3] http://www.scopus.com/

metrics [1, 3, 9], i.e., the evolution of scientific output over time, and the estimation of future impact based on early information regarding scientific productivity.

The present article addresses the following questions: "How does the career of a scientist in terms of his/her impact on the community progress over time?" and "Are there early signs of scientific potential?"

Towards this goal, this article focuses on grouping and comparing scientists of similar academic age over a period of 30 years based on a set of representative bibliometric indexes. The key objective is to identify patterns in the evolution of scientific output and realize whether scientists progress or not with respect to the quality of their results, while at the same time identify scientists that demonstrated early signs of increased scientific impact by comparing them with their academic peers. This is a macroscopic methodology contrary to previous microscopic ones such as the one developed in [7]. In this context, the article makes the following contributions:

- It introduces the problem of consistently grouping scientists of similar performance and age over a long time period, while maintaining a common basis for comparisons,
- It develops a methodology for quantifying and visualizing the evolution of an individual's scientific career in terms of its impact over time,
- It evaluates the effectiveness of this computerized methodology by comparing it with a peer-review based method that results in well-known computer science awards.

2. THE ANALYSIS TECHNIQUE

Since the heart of the investigated question is rather qualitative than quantitative, we have to devise a new methodology to address it. In this section we describe the proposed methodology and its associated strengths and limitations along with the metrics and heuristics used to validate it. Firstly, we provide details on the dataset used in the present work.

2.1 Data set description

For the purposes of this study, a set of scientists publishing in the field of computer science was collected along with their citation records, the publication year of their papers and the number of coauthors. The data were retrieved from the *Microsoft Academic Research* database, according to its field and domain categorization. The original set constitutes of $100,000$ scientists linked to the Computer Science field according to MAS and they were reduced to a set of $30,000$ scientists with an h-index higher than 8 to filter out low cited authors, since we are looking for distinguishing individuals. This resulted in a total of over 3 million papers with their associated information. To evaluate the temporal evolution of each author's scientific impact, we proceed to acquire "snapshots" of the publication and citation records of the authors in our dataset for 7 specific years with a 5-year time step [1983, 1988, 1993, 1998, 2003, 2008, 2013]. For each of these years, we divide the authors in subgroups according to their academic age, i.e., the number of years since they published their first paper. The intervals used to classify academic age were 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45 and 45-50 years (the right end of the intervals is open). The idea behind this categorization is to compare authors of similar academic age and monitor the evolution of their performance over time. Moreover, by dividing scientists into 10 groups according to their academic age for each one of the given years, it is ensured that a scientist who belongs to the first age group (0-5 years active) in 1983 will advance to the second age group (5-10) in the next year to be examined, that is 1988. Table 1 displays the data distribution over all years and age groups in absolute numbers. For evaluation purposes we have also identified a set of top scientists, consisting of those who have received the E. F. Codd[4]and the Turing award[5].

Age Groups	1983	1988	1993	1998	2003	2008	2013
0-5	1732	3264	6937	7330	5949	1670	7
5-10	1167	1814	3376	7094	7380	5972	1667
10-15	643	1169	1829	3400	7124	7385	5964
15-20	303	652	1178	1837	3413	7130	7388
20-25	134	306	654	1186	1835	3415	7133
25-30	38	134	311	657	1194	1835	3414
30-35	14	38	135	315	663	1199	1837
35-40	7	14	38	136	318	670	1200
40-45	2	7	14	39	136	317	670
45-50	7	2	7	14	40	139	321

Table 1: Cardinality of the 10 age groups.

2.2 Methodology

For every time interval and each age group, we deploy a clustering technique to group together scientists of analogous impact. The first step is to identify the appropriate features according to which we divide the authors into clusters, so that the resulting grouping is deemed representative of the quality and status of the scientists included. According to [12], bibliometric indices can be placed into three categories based on the part of the citation curve they emphasize on: i) indices focusing on the productive h-core, ii) indices focusing on citation count, and iii) indices taking into account the whole citation curve. Using one bibliometric index from each one of the above groups allows us to incorporate information on various aspects of a scientist's impact and quality of work. Therefore, to achieve a meaningful and representative segmentation of the scientists in our data set the h-index [4] was chosen as a representative metric from the first group, the total number of citations (C) from the second group, and the Perfectionism Index [14] (PI) from the third group.

[4] `http://www.sigmod.org/sigmod-awards/`

[5] `http://amturing.acm.org/byyear.cfm`

As stated earlier, grouping scientists together across all 10 age groups over the selected 7 time intervals constitutes an unsupervised learning task. However, choosing the suitable clustering algorithm as well as the appropriate number of clusters to achieve a meaningful result that allows for comparisons over time and age, is a challenging procedure. Dynamic clustering algorithms, such as Learning Vector Quantization, ART model, Fuzzy C-means, DBSCAN, are able to find the optimal number of clusters, but this property contradicts the goals of our temporal clustering task. This is due to the fact that our data sets change for every given year and age group, for instance, for the age group 5-10 in year 1988 scientists could be divided into 3 clusters, while, for the same age group in year 1993, the optimal number of clusters could be 4. As a result, comparisons could not be performed over all years and age groups, as the connection between the top cluster of two different "snapshots" would be unclear. Moreover, if the number of clusters is dynamically defined, we would not be able to automatically rank the resulting clusters and detect the clusters containing the high impact scientists, and those with the lower impact scientists. Therefore, we have adapted a neural network clustering approach, the Self-Organizing Map (SOM) [8], that can be adjusted to produce a specified number of clusters based on the topology of neurons used and the learning procedure employed[6].

To address the issue of specifying the appropriate number of clusters, we opted for a two-phase approach. The first stage of our approach focuses on the automatic ranking of clusters, i.e., the interpretation of the level of scientific impact represented in each cluster. To achieve this ranking, we sum the maximum values of all 3 normalized features for each cluster. The same is done for the minimum and average values. Consequently, for each distinct clustering, a matrix is produced where the number of rows equals the number of clusters and 3 columns for the above mentioned summations ($sumMax$, $sumMin$, $sumMean$) and then a weighted combination of these summations is produced, by assigning equal weights to the maximum, minimum and average values. The top cluster is defined as the one with the biggest score, whereas the lowest ranked cluster has the smallest score; tie breaking is based on the summation of the maximum values of all features.

Now that a consistent cluster ranking has been set, the appropriateness of the number of clusters needs to be evaluated. To this end, a well-known metric for unsupervised learning was utilized, the *silhouette measure* [10], which provides an estimation of the similarity of a point with respect to the other members of the same cluster (cohesion), as well as its dissimilarity with respect to points belonging to the other clusters (separation). Although a high silhouette score indicates a well segmented grouping of the original data based on inter- and intra-cluster similarity, we also need an insightful result to interpret what the content of each cluster means for a scientist's impact. In this direction, we have employed two more measures, which are analogous to the traditional precision and recall. The set of top authors that we defined in Section 1 consisting of award winning scientists is employed for evaluation purposes, as these scientists need to be clustered to a high impact cluster in all our data sets. The following two measures were identified:

- *precision*: the fraction of the scientists that have received awards to the total number of scientists that have an average cluster membership classified as high,
- *recall*: the fraction of the scientists that achieved a high average cluster membership out of the ones that have been awarded.

It is expected that the values of precision and recall are going to be relatively low with respect to what values we usually encounter in information retrieval settings. Nevertheless, they can be utilized to provide an insight on how meaningful and effective our clustering approach is according to the chosen number of clusters.

# of clusters	precision	recall	silhouette
2	0.008	0.535	0.770
3	0.025	0.195	0.689
4	0.021	0.530	0.640
5	0.022	0.500	0.600

Table 2: Precision, recall and average silhouette for different numbers of clusters.

Table 2 displays the average silhouette scores as well as the precision and recall values for 4 different numbers of clusters: $2, 3, 4$ and 5 clusters respectively. As our grouping of scientists needs to reflect performance levels, we opted for a relatively small number of clusters. For any higher number of clusters the middle-rank clusters could not reflect a distinctly defined performance level. For each chosen number of clusters, the whole set of 70 clusterings over all time intervals and age groups were conducted. As depicted in Table 2, the best trade off between high silhouette and equally good precision and recall scores is achieved with the number of clusters equal to 4. In the next section, we proceed to validate the merits of our method and the insightful information it provides on our datasets.

3. EXPERIMENTS AND RESULTS

3.1 Clustering visualization

The first step in our analysis is to provide evidence that our clustering methodology with the chosen number of clusters produces meaningful segmentation of our

[6] The classic k-means, which is often used in cases with prespecified number of clusters, was utilized for additional experimentation, but led to low quality results. For brevity reasons we did not include the additional experiments in the present work.

data set. To this end, the scientists' membership to a cluster is plotted in the Principal Component space of the 3 clustering features for further evaluation. Figures 1 and 2 display the formed clusters for 2 age groups in the year 2013 and the respective positioning of the feature vectors. It can be seen that the scientists are depicted as points of different color according to the cluster they belong to; cluster 4 is represented with red dots, cluster 3 with green, cluster 2 with blue and cluster 1 with cyan. Authors to the far right (depicted with red) are the ones with the combination of the highest values in all three features.

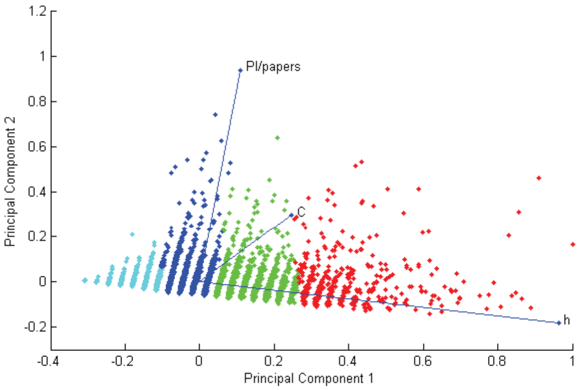

Figure 1: Authors with academic age 10-15 years in the year 2013 clustered in 4 groups projected on the Principal Component space.

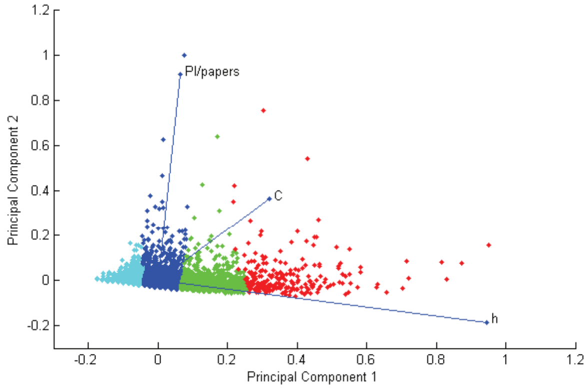

Figure 2: Authors with academic age 15-20 years in the year 2013 clustered in 4 groups projected on the Principal Component space.

Although our number of features is small, projecting them to the Principal Component space allows for their interrelations and correlations to be displayed. This plot allows for an assessment of the consistency of the clusters and the degree of separation between the 3 clustering features, which are plotted as vectors in the principal component coordinate system. As depicted in Figures 1 and 2, the PI index and the h-index are uncorrelated with each other, whereas the citation count C is connected to both features. The first Principal Component mostly focuses on citation count, whereas the second Principal Component expresses the publication count. As a result, the top cluster displayed to the far right of the figures includes the scientists with high citation count and small increase in publication count. As expected, the high impact cluster (cluster 4) is the smallest one in any given year, whereas the low impact and low-moderate cluster (clusters 1 and 2) are the most densely populated.

3.2 Scientists' impact over time

We now proceed to present the findings on the questions we initially set about the temporal evolution of scientists and the existence of early signs indicating increased academic impact. A more detailed picture of the evolution of cluster memberships and the trends that appear over the years is illustrated in Figures 3-5, which depict colormaps of the cluster memberships in the 7 examined time intervals. Each value on the y axis represents a specific scientist and different colors depict the cluster membership in the year indicated by the x axis value. Colder colors (starting from cyan) represent low impact clusters, while warmer colors indicate higher impact clusters. More specifically, cyan represents scientists that have been clustered in cluster 1 (low impact), blue for cluster 2 (low-moderate impact), green for cluster 3 (moderate-high impact) and finally red stands for cluster 4 (high impact). Intervals depicted with white represent scientists that have not yet published in the given time interval, i.e., they have not been clustered yet.

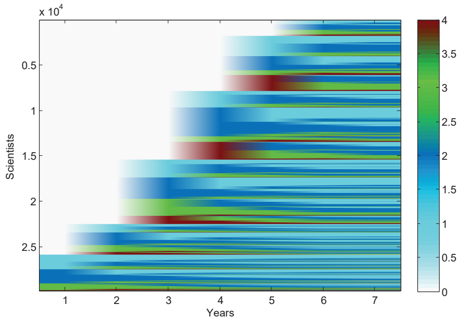

Figure 3: Cluster membership of all authors in the set since they first appear in a cluster.

In Figure 3 we can see that many scientists start from the first two clusters and then either remain on that impact level or progress to higher impact clusters. However, there is also a group of scientists with declining impact as time progresses, that end up into lower impact clusters than when they started. A small percentage of scientists are classified as high impact ones from the beginning of their academic career. Figures 4-5 provide an insight in the groups of scientists that present zero declines in impact level (i.e., cluster memberships) and the ones that have been classified more than 4 times

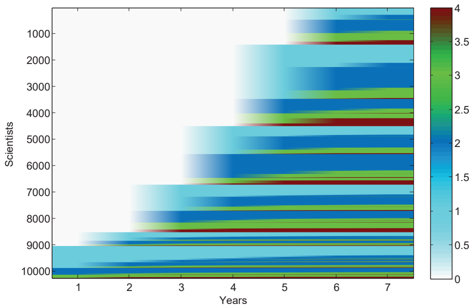

Figure 4: Cluster membership of scientists that managed to progressively increase their score in all 3 bibliometric indices (C, h, PI) thus improving their cluster memberships.

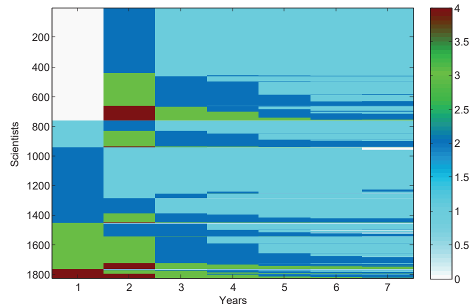

Figure 5: Cluster membership of authors who have been clustered more than 4 times in a lower impact cluster than the best cluster membership they have ever scored.

in lower impact clusters than the top cluster they were grouped into respectively. In particular, Figure 4 displays the scientists that have managed to progressively improve their impact as measured by the 3 bibliometric indices used in this study, while Figure 5 demonstrates the ones that failed to maintain their highest achieved status. This decline is mostly captured by the Perfectionism Index that becomes smaller with the addition of new low or zero cited papers. Moreover, a scientist's academic age increases gradually; therefore, they are being compared with a different set of more mature scientists, and if they fail to analogously raise the quality of their research, they end up in a lower impact cluster.

As already mentioned, this clustering approach can yield useful results regarding identification of distinguished scientists early in their career as well as the timely assessment of their academic career as characterized by bibliometric indices. Table 3 contains[7] a set of well renowned scientists publishing in the field of Databases that have won a E. F. Codd award and their membership to the 4 given clusters. It can be observed that (almost) all of them have been grouped in the top cluster (cluster 4) from the beginning of their career. This observation provides a confirmation of the validity of our methodology and indicates that award winning scientists often have achieved high scores according to bibliometric indexes early in their careers. Even though their scores may not be significantly high as absolute values, they can prove distinguishing when compared to the analogous scores of their academic peers belonging to the same age group. Nevertheless, exceptions to this strong pattern can be seen in the cases of J. Gray, S. Chaudhuri who have an 'industrial' rather than 'academic' profile.

	1983	1988	1993	1998	2003	2008	2013	award on
Ceri S.	3	4	4	4	4	4	4	2013
Lindsay B.	3	4	4	4	3	3	3	2012
Chaudhuri S.			1	3	4	3	3	2011
Dayal U.	3	4	4	4	4	4	3	2010
Kitsuregawa M.		2	2	2	2	2	2	2009
Vardi M.	3	4	4	4	4	4	4	2008
Widom J.		2	4	4	4	4	4	2007
Ullman J.	4	4	4	4	4	4		2006
Carey M.	1	2	4	4	4	4	4	2005
Fagin R.	4	4	4	4	3	4	4	2004
Chamberlin D.	4	4	4	3	3	3	2	2003
Selinger P.	3	4	3	3	2	2	2	2002
Agrawal R.		3	4	4	4	4	4	2000
Garcia-Molina H.	3	4	4	4	4	4	4	1999
Abiteboul S.		3	4	4	4	4	4	1998
Maier D.	4	4	4	4	4	3	3	1997
Mohan C.	3	4	4	4	3	3	3	1996
Dewitt D.	2	4	4	4	4	4	4	1995
Bernstein P.	4	4	4	4	4	4	4	1994
Gray J.	3	3	3	3				1993
Stonbraker M.	4	4	4	4	4	4	4	1992

Table 3: Cluster membership for scientists that have won the ACM SIGMOD's E.F. Codd award.

In Table 4, we repeat the same analysis but for those who have won a Turing award. The generic pattern is similar to that observed for Table 3, but now the case of seeing less scientists being constantly grouped in cluster 4 appears more often. This is mainly due to the fact that our data cover a specified period of time, while a number of Turing Award winners have reached their academic peak before that time.

As it is often the case, many indexes used to assess scientific impact (like the h-index) are cumulative in nature, meaning that they never decrease. As a result, using them as standalone metrics without an added time window leads to poor conclusions about the real scientific impact of authors, thus eliminating their predictive power. A unified framework such as the one proposed in this work that incorporates combination of features and the time parameter as well as the concept of peer comparison can lead to valuable characterization of scientific output and reveal early signs of increased scientific potential.

4. CONCLUSIONS

The present article developed an unsupervised methodology for analyzing and categorizing scientific careers over time aiming to be used as a – complementary to other approaches – tool for promotions and funding by recognizing individuals of high scientific impact. The developed methodology is appropriate for big data anal-

[7] The empty cells in Table 3 and 4 indicate that a scientist was not part of the examined age groups during the specific year.

	1983	1988	1993	1998	2003	2008	2013	award on
Stonebraker Michael	4	4	4	4	4	4	4	2014
Lamport Leslie	4	4	4	4	4	4	4	2013
Goldwasser Shafi	3	4	4	4	4	4	4	2012
Micali Silvio	3	4	4	4	4	4	4	2012
Pearl Judea	3	4	4	4	4	4	4	2011
Valiant Leslie	4	4	4	4	4	3	3	2010
Thacker Charles	3	3	3	2	1	1	1	2009
Liskov Barbara	4	4	4	4	4	4	4	2008
Clarke Edmund	2	3	4	4	4	4	4	2007
Emerson E. Allen	3	4	4	4	4	4	4	2007
Sifakis Joseph	3	3	3	4	3	3	3	2007
Allen Frances	3	3	3	2	2	1	1	2006
Naur Peter	1	2	1	1	1	1		2005
Kahn Robert	3	3	3	2	1	1	1	2004
Cerf Vint	2	2	2	3	3	2	2	2004
Kay Allan	2	3	3	2	2	2	2	2003
Adleman Leonard	4	4	4	3	3	3	3	2002
Rivest Ronald	4	4	4	4	4	4	4	2002
Shamir Adi	4	4	4	4	4	4	4	2002
Nygaard Kristen	3	2	2	2	2	2	2	2001
Dahl Ole-Johan	2	2	2	2	1	1	1	2001
Yao Andrew	4	4	4	4	4	3	3	2000
Brooks Frederick Phill.	2	2	3	2	3			1999
Gray Jim	3	3	3	3	3			1998
Engelbart Douglas	2	2	3	2	2	2		1997
Pnueli Amir	4	4	4	4	4	4	4	1996
Blum Manuel	3	3	3	3	3			1995
Feigenbaum Edward	2	3	2	3	2	2		1994
Reddy Raj	3	3	3	3	2	2	2	1994
Hartmanis Juris	2	3	2	3	2	2		1993
Stearns Richard	4	3	3	3	2	2		1993
Lampson Butler	4	4	4	4	3	3	3	1992
Milner Robin	4	4	4	4	4	4	4	1991
Corbato Fernando	3	2	1	1	1	1		1990
Kahan William	3	3	3	2	2	2	2	1989
Sutherland Ivan	3	3	3	2	2	2	2	1988
Cocke John	3	3	3	2	2	2	2	1987
Hopcroft John	4	4	4	4	4	4	4	1986
Tarjan Robert	4	4	4	4	4	4		1986
Karp Richard	4	3	4	4	4	4		1985
Wirth Niklaus	4	4	4	4	3	3	3	1984
Thompson Ken	2	3	3	2	2	2	1	1983
Ritchie Dennis	4	4	3	3	2	2	2	1983
Cook Steve	3	4	4	4	3	3	3	1982
Codd Edgar	4	4	4	3	3	2	2	1981
Hoare Tony	4	4	4	4	4	4		1980
Floyd Robert	4	3	3	3	2	2		1978
Backus John	3	3	3	2	2	2		1977
Rabin Michael	3	3	3	3	3	3		1976
Scott David	3	3	2	3	2	2		1976
Newell Allen	4	4	4	4	4			1975
Simon Herbert	4	4	4	4				1975
Knuth Donald	4	4	4	4	4	4		1974
Backman Charles	3	2	2	1	1	1	1	1973
Dijkstra Edsger	4	4	4	4	4			1972
McCarthy John	3	3	3	3	3	3		1971
Wilkinson James	3	2	3	2	2	2		1970
Minsky Marvin	3	3	3	3	3	2		1969
Hamming Richard	2	2	2	2				1968
Wilkes Maurice	4							1967
Perlis Alan	3	2	3	2	2	2		1966

Table 4: Cluster membership for scientists that have won the Turing award.

ysis purposes and it is based on established dimensionality reduction and clustering algorithms with the addition of proposed heuristics and metrics to allow for an automated and unified over time ranking to be achieved. The identified as top scientists based on their cluster memberships through the years are in accordance to the ones who have won discipline-specific (i.e., E. F. Codd) and generic awards (i.e., Turing).

Acknowledgments. The offer of Microsoft to provide gratis their database API is appreciated.

5. REFERENCES

[1] E. Bruna. On identifying rising stars in ecology. *BioScience*, 64(3), 2015.

[2] P. della Briotta Parolo, R. Pan, R. Ghosh, B. Huberman, K. Kaski, and S. Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734–745, 2015.

[3] P. Erdi, K. Makovi, Z. Somogyvari, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalanyi. Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics*, 95(1):225–242, 2013.

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

[5] D. Katsaros, L. Akritidis, and P. Bozanis. The f index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science and Technology*, 60(5):1051–1056, 2009.

[6] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[7] N. Kejzar, S. Korenjak-Cerne, and V. Batagelj. Clustering of distributions: A case of patent citations. *Journal of Classification*, 28(2):156–183, 2011.

[8] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer, 2001.

[9] P. Z. Revesz. A method for predicting citations to the scientific publications of individual researchers. In *Proceedings of the IDEAS*, 2014.

[10] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[11] M. Schreiber. Restricting the h-index to a publication and citation time window: A case study of a timed Hirsch index. *Journal of Informetrics*, 9(1):150–155, 2015.

[12] A. Sidiropoulos, A. Gogoglou, D. Katsaros, and Y. Manolopoulos. Gazing at the skyline for star scientists. *Journal of Informetrics*, 2016. in press.

[13] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.

[14] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Identification of influential scientists vs. mass producers by the Perfectionism index. *Scientometrics*, 103(1):1–31, 2015.

[15] C.-T. Zhang. The e-index, complementing the h-index for excess citations. *PLoS One*, 4(5):e5429, 2009.