Journal of Informetrics 11 (2017) 198–222

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Detecting rising stars in dynamic collaborative networks

George Panagopoulos[a,*], George Tsatsaronis[b], Iraklis Varlamis[c]

[a] Computational Physiology Lab, University of Houston, Houston, USA
[b] Content and Innovation Elsevier, Amsterdam (ELS-AMS), Netherlands
[c] Department of Informatics and Telematics, Harokopio University of Athens, Greece

ARTICLE INFO

Article history:
Received 13 March 2016
Received in revised form
16 November 2016
Accepted 17 November 2016

Keywords:
Data mining
Time-evolving graphs
Rising stars
Co-authorship graphs
Power graphs
Key performance indicators
Evaluation of scholars

ABSTRACT

In today's complex academic environment the process of performance evaluation of scholars is becoming increasingly difficult. Evaluation committees often need to search in several repositories in order to deliver their evaluation summary report for an individual. However, it is extremely difficult to infer performance indicators that pertain to the evolution and the dynamics of a scholar. In this paper we propose a novel computational methodology based on unsupervised machine learning that can act as an important tool at the hands of evaluation committees of individual scholars. The suggested methodology compiles a list of several key performance indicators (features) for each scholar and monitors them over time. All these indicators are used in a clustering framework which groups the scholars into categories by automatically discovering the optimal number of clusters using clustering validity metrics. A profile of each scholar can then be inferred through the labeling of the clusters with the used performance indicators. These labels can ultimately act as the main profile characteristics of the individuals that belong to that cluster. Our empirical analysis gives emphasis on the "rising stars" who demonstrate the biggest improvement over time across all of the key performance indicators (KPIs), and can also be employed for the profiling of scholar groups.

1. Introduction

Evaluation of faculties and research scholars in the modern academic world is becoming increasingly difficult. The main problems pertain to the different objectives that each faculty or department sets, but also to the constantly changing academic market. More precisely, there is a shift from the full-time, tenure-eligible faculty members that used to be the case in the past, to increasing numbers of part-time faculty members, non-tenure-track faculty members, and even online course instructors in colleges and universities around the globe. In addition, there is lack of widely accepted guidelines, but instead there is a plethora of best practices, often released by each faculty individually, on how the evaluation should take place (Buller, 2013).

As a result, the evaluation committees often need to provide evaluation summary reports of individuals that are aligned with the faculty's scope and objectives, and need to do so without having the ability or the time to meet the candidate in person. For this purpose, there is a large need of automated tools that can provide an analysis of the scholars' profiles based on a wide range of key performance indicators (KPIs) that may cover the majority of the faculty's needs or goals.

[*] Corresponding author.
E-mail addresses: gpanagopoulos@uh.edu (G. Panagopoulos), g.tsatsaronis@elsevier.com (G. Tsatsaronis), varlamis@hua.gr (I. Varlamis).

http://dx.doi.org/10.1016/j.joi.2016.11.003
1751-1577/Published by Elsevier Ltd.

There is certainly a large number of KPIs that may be considered for the evaluation of academics and faculties. Teaching and research performance, ability to raise research funds, as well as patents awarded or pending are only few of the aspects that may be considered for such an evaluation. In this work we are focusing on the evaluation of individual scholars from the perspective of their research activities performance. We argue that the scientific performance of individuals should take into consideration apart from bibliometric indicators such as the h-index (Hirsch, 2005) and journal impact factors, the social aspects of the researcher's contribution as well as its dynamics. The rationale behind this argument stems on the one hand from several past observations around the inconsistency of h-index (Waltman & Eck, 2012) and impact factor (Seglen, 1997) while, on the other hand, from the limitations that a single number evaluation model suffer from, like narrowing the margins of alternative interpretations and consequently, failing to highlight the advantages, disadvantages and future potential of a scholar. Such a view would also aid the evaluation committees to document in detail the rationale behind their decision and align it easier with the faculty's goals.

In order to address the aforementioned requirements, we present a computational method to analyze the profiles of individual scholars based on a number of performance indicators covering both scientific performance and social/collaboration related features, as well as their evolution over time. To measure an author's productivity and impact of his work we employ alternations of metrics such as the volume of publications and citations each work receives, emphasizing on their ephemeral character to distinguish between truly seminal works and temporal successes. Furthermore, to measure an author's sociability, which accounts for the number of co authorships he or she forms and their impact on his or her career, collaboration networks are exploited, i.e., co-authorship graphs, that are created from bibliographic data and analyzed using traditional graph mining (Cook & Holder, 2006) and Power Graph Analysis (Royer, Reimann, Andreopoulos, & Schroeder, 2008). We then proceed to measure the changes of each such indicator across consecutive time periods in order to capture the scholar's dynamics. As a final step, we cluster the scholars' profiles in each period using all of these indicators as features and perform a feature analysis to characterize the clusters and an evaluation using the clusters future values.

The evaluation of academic performance has several ethical implications, which have been extensively discussed in the related literature. For example, Sir Philip Campbell, the Editor-in-Chief of Nature journal, stated that the most effective and fair analysis of a person's contribution derives from a direct assessment of individual papers, and not of the venues they were published (Campbell, 2008). This reduced the importance of an impact factor and increased the significance of individual (per publication) citations (Cronin, 1984). Similarly, the widely known *h-index* received criticism since in some cases it may provide misleading information about a scientist's output or can be biased in various ways, such as using self-citations, publishing in different research domains, open access, the cumulative advantage of more senior researchers, etc. (Allison, Long, & Krauze, 1982; Antelman, 2004; Batista, Campiteli, & Kinouchi, 2006; Harzing & van der Wal, 2008; Hirsch, 2007; Wendl, 2007). The use of large bibliographic and citation databases (e.g. Scopus, Google Scholar, etc.) that provide the exact number of citations per publication solved the venue impact bias and also the cumulative bias of old papers, but there are still issues to be solved, such as the quantification of each author's contribution in a publication. A comprehensive survey on the pitfalls of research evaluation and a plan for objectivity in evaluation metrics is presented by Retzer and Jurasinski (2009). Another aspect that is also under extensive examination is how to evaluate the actual quality of a scientific work, which is an a rather multifaceted and complicated task (Andersen, 2013; Frey & Rost, 2010) that comprises more than simple publication and citation count (Ochsner, Hug, & Daniel, 2014). In this article, we propose a set of indexes that can be used to evaluate multiple facets of an author's research potential, we penalize the impact of a work as time passes and focus on the changes of these indexes across time in order to remove the cumulative bias and instead of ranking authors or classifying authors to good or bad ones, we cluster authors of similar potential into groups. This grouping can be more useful in a researcher selection process, since it can restrict the number of candidates to a smaller group of individuals with strong potential, leaving space for a selection process that takes into account all the aspects of each candidate without staying only on raw numbers and indexes.

We tilt our analysis over the cases of rising stars due to their significance and challenging nature. Li, Foo, Tew, and Ng (2009) describe rising stars are those who currently have relatively weak profiles but may eventually emerge as prominent researchers. The difficulty in identifying rising stars is that it is essentially a prediction task that spans across several years, rather than an instantaneous classification task. More specifically, it is necessary to collect data from when the author is still relatively inconspicuous and for a period of time in order to claim at an early stage that he or she will become a great scientist in the future. Thus, one has to look in specific details that reflect the potential of this author, rather than traditional metrics of author assessment. Furthermore, it is important to follow the author publication and citation record over time in order to validate that the prediction was correct. In our previous work (Tsatsaronis et al., 2011), attempting to model the "group leaders" profiles, we also defined rising stars as the authors who show high increase in the amount and impact of their published work over time. Daud, Abbasi, and Muhammad (2013) define rising stars as the authors which may become prominent contributors in the future, though are currently having a low research profile.

The novelty and the main contribution of the current work lies in the fact that for the first time to the best of our knowledge, quantity, impact and collaboration related features are combined together and monitored over time in order to create profiles of scholars and to highlight their strengths and weaknesses. Based on a set of time evolving features, a clustering algorithm and several cluster validity measures, authors are grouped into subsets of relevant performance. The clusters are consequently labeled based on their most representative features; for each cluster the features with the highest values across clusters are used. An additional contribution of this work is the methodology we use in the analysis of clusters, which highlights the topics of interest to the authors of each cluster, the fora in which they usually publish and the domains they belong to.

To evaluate the suggested methodology we performed experiments in a dataset created from the amalgamation of all the authors that have published a scientific work under a Greek affiliation according to the Scopus bibliographic database (a total of 43,567 scholars). The usage of Scopus ensures reliability in terms of bibliographic information gathered, and covers a broad range of data concerning publications and citations per year. It is vital to underline that any other group of researchers and any other bibliographical database could be used instead for the application of our methodology, as long as the database indexes information of publications and citations per article and per year.

The results of our analysis led us to three key findings: (1) scholars who share similar values of the bibliometric KPIs tend to share similar values in their collaboration KPIs as well, (2) monitoring the evolution of the KPIs over time is vital to identify scholars enduring the "test of time", e.g., scholars who interminably and consistently contribute high-impact works, and, (3) among the identified clusters we were able to distinguish the groups of scholars that have high dynamics and hence, the potential to reach high level in their career, referred in the bibliography as "rising stars". These findings, following the analysis of the 43,567 scholars, come as a proof of concept to the originally formulated argument that in the modern academic world the evaluation of scholars should not be based solely on bibliometric indicators, but also on their collaborations and dynamics.

2. Related work

In the recent past, several works have endeavored to identify "rising stars" in academic networks. The works either rank authors based on a score that indicates the potential of a "rising-star", or classify authors to "rising" and "non-rising" classes, or cluster together authors that have similar characteristics. Almost all of them, extract several features (indicators) for each author and evaluate their power in discriminating "rising stars". However, none of them so far has employed a combination of all three aspects, namely bibliometrical indicators, collaboration indicators, and monitoring over time.

More precisely, Li et al. (2009) propose the *PubRank* algorithm to identify "rising stars". In their work, the "rising stars" are nodes in the collaboration graph, who currently have relatively low profiles, but may eventually emerge as prominent researchers. They consider the mutual influence among researchers in the network using weighted and directed links for each collaboration (bi-directional graph), the average impact of the researcher's publication record using weighted nodes in the graph model, and chronological changes of these networks. In PubRank, influence is propagated over the network following a similar principle to that of the well known PageRank algorithm. PubRank is computed every year for each author and the trend is examined to reveal authors with a significantly larger increase than the average researcher, naming them rising stars. In their experimental evaluation, they used DBLP data from 1990 to 1995 to predict the rising stars. The cross validation was made in comparison to the year 2006, in order to quantify the strength of the predictions. They found out that their predicted rising stars PubRank score differed significantly from the scores of the average researcher. Subsequently they focused the method on authors from the Database domain in DBLP and using data from 1990 to 1994, identified rising stars and evaluated their findings on the top 20 of them manually, showing that all of them hold important positions in the industry or in the academia.

Daud et al. (2013) highlight two major limitations of *PubRank*: (a) that the author contribution in a publication is ignored, and, (b) that the venue rank cannot be considered static. They propose the *StarRank* algorithm, which overcomes the aforementioned limitations by considering the order of author names and by adjusting venue rank over time using entropy of topics presented in the venue. The first hypothesis does not apply in all disciplines; for example in some cases co-authors list their names in alphabetic order while in other cases the senior author always goes at the end of the author list. The second hypothesis is also weak, given that the terms in paper titles cannot always account for the covered domains of a venue. Moreover, the algorithm treats in an equal manner the publications in conferences and in journals. However publishing in an impactful journal is arguably harder than in a prestigious conference. Long, Lee, and Jaffar (1999) use a venue ranking scheme, which handles solely conference rankings, so it is obscure whether journal impact factors or another metric is used for journal ranking. Either way journals should weigh more than conferences in an author's career. The experimental evaluation is similar to that of Li et al. (2009), since Long et al. (1999) manually assess the top ten predicted rising stars in a future moment using the number of publications and the number of citations to prove that their method surpasses the aforementioned PubRank.

Arnetminer,[1] which is based on DBLP data, defines the rising stars using *uptrend*, a measure of the rising degree of a researcher. For a given author, it uses least mean squares to fit a curve the slope of which is based on the publications' impact factors and the time they were published. Using this curve, it predicts the author's uptrend, meaning the next year's score. Arnetminer also includes the category of "new stars" which are researchers that have less than 5 years of research longevity (based on the date of their first publication). This is similar to the definition of the rising star given by Daud et al. (2013), where rising stars are authors in the early years of their career and lack citations. However, the rising star definition should not be constrained to people with only few years of research, as many researchers can have a breakthrough in their career several years after their first publication, independently of their performance within this period.

[1] http://arnetminer.org.

Daud, Ahmad, Malik, and Che (2015) treated the problem of "rising-stars" detection as a binary classification problem. As a result, they used two labeled datasets and data from DBLP and Arnetminer database (1995–2000): one comprising 500 highly cited and 500 scantly cited authors, and a second comprising 500 authors with the highest Average Relative Increase in Citations (ARIC) and 500 authors with the lowest ARIC. They extracted eleven features based on the author and his/her co-authors that pertain to citations, influence and contribution and evaluated two generative classification models – the Bayes Network (BN) and Naïve Bayes (NB) – and two discriminating ones – the Maximum Entropy Markov Model (MEMM) and the Classification And Regression Tree (CART). For feature analysis, all four classifiers were trained with each of the eleven different features, with different subsets of the data and the best F1 scores[2] revealed that the most important features were the citations received in the published venue, the co-authors venue score, the co-authors citations and the venue specificity score, while the best classifiers were MEMM for the first dataset and CART for the second. For both datasets, venue-based features gave the best accuracy. In addition to the evaluation process, authors used a combination of authors' features to calculate a rising star score.

In their work, Fu, Song, and Chiu (2014) introduce new aspects of the author's profile, namely *Influence*, *Connections* and *Exposure*, which can provide different rankings of authors and together with Citation Count can give a fuller picture about authors. The work builds on top of a complex network topology, which connects nodes of different types (authors, venues and papers) with different type of edges (directed and undirected, bi-partite and uni-partite) and provides different author rankings for each of the metrics. In their evaluation they study the correlation between the different metrics and the correlation with h-index. This work also ignores the evolution of metrics over time, and thus it is able to detect well-established authors but not authors with great potential.

In the opposite side of rising stars, Tagarelli and Interdonato (2015) search for inactive and non-productive users in social networks using time-aware analysis. The so-called "lurkers" or "silent users" are modeled using a set of productivity and content consumption properties, which are time-aware. The authors also perform clustering on the set of all users in order to detect the cluster of "silent users".

The methodology that we propose differs from the aforementioned approaches in several points:

- First, the impact of a paper is measured by the number of citations it receives and not by the impact factor of the venue it was published (Tsatsaronis et al., 2011). Although the venue where a work is presented can strongly affect the visibility of the paper and consequently the number of citations it receives, according to Seglen (1997) the use of impact factors conceals the difference in article citation rates. According to the same work, articles in the most cited half of articles in a journal are cited 10 times as often as the least cited half. As a result, the use of citations is a more effective and fine-grained measure of the impact of a paper, than the venue impact factor.
- Second, in this work we add the *decay factor*, which punishes a publications' success based on their oldness and the oldness of the citations they received. This makes our methodology more robust to the self-citation bias, whilst accepting authors to cite their work for communication purposes. According to the previous works of Wolfgang, Bart, and Balázs (2004) and Glänzel and Thijs (2004) the aging of self-citations is much faster than that of foreign citations, so since self-citations usually appear the next few years after a publication, their effect on the cumulative impact of the publication weakens fast. Our methodology describes the author's dynamics better than h-index (Hirsch, 2005) since it takes into account the annual change of citations for an author and not their cumulative number. Thus it better distinguishes between an author who keeps receiving the same amount of citations every year, a young "rising star" author whose annual number of citations grows every year and a well established author who receives fewer citations as years pass but already has a huge number of total citations and an h-index above 30 (Costas, Van Leeuwen, & Bordons, 2010).The use of clustering allows our methodology to first create groups of authors with similar characteristics and then to label these groups and detect the one comprising "rising-stars".
- Third, we identify an optimal number of clusters and perform clustering on the authors' dataset instead of binary classification (e.g., predict whether an author is a "rising star" or not) and this allows us to discover additional researcher types, similar to what Tsatsaronis et al. (2011) do.
- Fourth, we propose a more challenging qualitative evaluation, than that of Li et al. (2009), examining both, whether the identified clusters stay distinct in time and whether the clusters withhold their overall levels in the most crucial characteristics. The data we use to that endeavor comes from the same dataset in subsequent years, in other words we split the data in two parts, one for analysis and clustering and the other for evaluation of our algorithm.

3. Methodology

The proposed methodology analyzes the scientific performance of a scholar both in terms of quantity (volume of publications) and Impact (of publications). It also considers the sociability of an author and the ability to cooperate with well-known or well established scholars. These three criteria, are commonly used for assessing or reviewing research and they reach the consensus of various disciplines ranging from humanities (Hug, Ochsner, & Daniel, 2013) to engineering (Bornmann, Mutz, Neuhaus, & Daniel, 2008). In addition to this, they can be easily extracted from bibliographical and citation databases.

[2] F1 or F-measure is the harmonic mean of precision and recall and in essence is closer to the smallest of the two metrics.

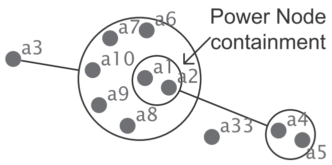


Fig. 1. An example of nested *Power Nodes*. The social impact of authors *a*1 and *a*2 is affected by the impact of their extended circle of co-authors in the PowerNode that contains the PowerNode of *a*1 and *a*2.

Our methodology first analyzes the publication and citation information database, which in our case is Scopus, and defines different collaboration graphs that show the frequency and impact of an author's collaborations. In a second step, we compute several metrics for each author based on the graphs. In a third step, we use the evolution of these metrics over time as the input to two clustering algorithms and find the different author types and their main features, which basically summarize their profile. A meta-analysis of the authors profiles within each cluster reveals the intrinsic features that contribute more to the clustering task, as well as a general sketch of each cluster's characteristics. In the following we describe the details of these steps.

3.1. Collaboration graphs, notations and definitions

Collaboration graphs (Odda, 1979) are widely used in the field of mathematics, social sciences and computer science. In such graphs the vertices correspond to individuals who collaborate with each other and the edges connecting them denote a collaboration. In co-authorship graphs the relationship depicted by an edge is the co-authorship of one or more scientific articles. In our methodology, we jointly use different types of collaboration graphs in order to capture the various aspects of researchers' publications (e.g., volume, impact, sociability). More specifically we employ the following types of graphs: Co-authorship graphs and Co-authorship Power Graphs.

Co-authorship graphs are a typical example of huge graphs with thousands of nodes and millions of edges. For example, Chakraborty, Patranabis, Goyal, and Mukherjee (2015) have crawled one of the largest publicly available datasets from Microsoft Academic Search (MAS), which houses over 4.1 million publications and 2.7 million authors. Their graph contained more than 8 hundred thousand authors, only from the computer science domain, who have written more than 2 million distinct papers. The graph contained a single edges between authors who have written at least one paper together. Given that a paper can have more than two co-authors, the number of edges was significantly larger. Our dataset comprises more than 40,000 authors who have co-authored more than 110,000 articles. As a result, there is a need for an efficient representation of these huge graphs that will further allow their analysis. In this direction, we have chosen *Power Graphs* and *Power Graph Analysis*, a methodology used in computational biology for processing huge and complex networks.

Power Graphs are special types of undirected weighted graphs, which offer a loss-less information representation of the original graphs, with reduced complexity. The three basic motifs recognized by *Power Graphs* are the *Star*, the *Clique* and the *Biclique*, and constitute the basic abstractions when transforming the original graph into a *Power Graph* with *Power Nodes*, i.e., sets of nodes, connected by *Power Edges*.

The first decision regarding the creation of co-authorship graphs is on the type and meaning of edges. When a paper has *k* authors, the two representation alternatives are either to add a *hyperedge* connecting the *k* author nodes, or to add simple edges connecting each pairwise combination of the *k* authors. Since *Power Graphs* do not support *hyperedges* we follow in this work the latter option.

In our case, *Power Graphs* are the abstraction of *co-authorship graphs* where a *Power Node* corresponds to a group of authors (e.g., *Power Node* that contains node *a*4 and *a*5 in Fig. 1) with dense collaboration and a *Power Edge* corresponds to a collaboration between an author and a group (e.g., node *a*3 and the big *Power Node* in the middle of Fig. 1) or between two author groups (e.g., *Power Node* that contains node *a*4 and *a*5 and the big *Power Node* in the middle of Fig. 1), depending on the motif. Finally, *Power Node* containment (e.g., the big *Power Node* in the middle of Fig. 1 that contains *Power Node* with nodes *a*1 and *a*2) denotes that the outer set of authors has dense collaboration with authors in the contained subset. The compression offered by *Power Graphs* allows us to process larger graphs with the same resources and the representation model that they use, with Power Nodes and Power Edges, fits perfectly to the co-authorship model, since Power Nodes comprise all authors that frequently collaborate and Power Edges depict the strength (frequency or impact, depending of the graph) of the collaboration. To provide an example of the compression we achieve, the power graph created for 2005 based on the publications of 43,567 authors comprises only 4699 Power Nodes.

In order to better explain the indicators we define in our methodology, we provide some definitions and notations in the list that follows. In our definitions we use the letters *x* or *y* for authors, *i* and *j* for publications and t_k for a time period or unit (e.g., year).

- C_x: set of author's *x* co-authors.
- P_x: set of author's *x* publications in all periods.
- $pub(x, t_k)$: number of publications of author *x* in period t_k.

- $cit(i, t_k)$: citations that paper i received at period t_k.
- $aut(i)$: number of authors of publication i.
- $per(i)$: the period of publication for article i.

In the definitions of *Power Graphs* we use χ or ψ to refer to a *Power Node* (i.e. an author or group of authors). We also employ:

- PN_χ: the set of *Power Nodes* connected to *Power Node* χ.
- PC_χ: the set of *Power Nodes* that contain *Power Node* χ.
- $WPN(\chi)$: the weight of *Power Node* χ.
- $WPE(\chi, \psi)$: weight of *Power Edge* connecting *Power Node* χ and *Power Node* ψ.

Since we process the collaboration graphs at different points in time, in order to examine the evolution of an author over time, we use the following notation for the periods that we examine:

- t_0: the period in time that the author published a research work for the first time.
- t_n: the period we examine, which must be after t_0.
- t_{end}: the most recent period in the publication database.

Our methodology employs both co-authorship graphs (namely the *Quantity Graph* and the *Impact Graph*) and Power Graphs (*Quantity Power Graph* and *Impact Power Graph*).

3.1.1. Graph types

The aim of *Co-authorship Graphs* is to measure the sociability of authors and the strength and impact of their collaborations. While the graph structure captures the ability of an author to collaborate with multiple other authors, the weight of the edges shows the actual strength or success of a collaboration.

The generic co-authorship graph of a certain period (e.g., year) t_n is formally modeled as: $QG_{t_n} = (V, WE)$, where V is the set of authors and a weighted edge $we = \{v_1, v_2, w_{v_1,v_2}\} \in WE$ represents that authors v_1 and v_2 have co-authored w_{v_1,v_2} papers from period t_0 until t_n. The weight contains the aggregated information of all papers in that period, either it is simply the number of publications or the cumulative number of citations, or any other time-penalized cumulative score. More specifically, in the *Quantity Graph* we define a quantity edge weight WE_{quan} as the amount of articles that authors x and y have co-authored in a certain period (1). On the other hand, in the *Impact Graph* we define an edge weight WE_{imp} as the impact of the collaboration of two authors (3). More specifically, when we create the *Quantity Graph* for a certain period t_n, we accumulate all articles authored by the same author (or co-authored by a pair of authors).

$$WE_{quan}(x, y) = \sum_{\forall i \in P_x \cap P_y} \cdot 1 \tag{1}$$

Generally, a co-authorship graph constructed at a certain point in time is able to contain information from the past until that point, however, traditional such graphs provide a mere snapshot information about publications ignoring useful information concerning the past changes in the graph (e.g., the point in time when an edge was first created, or when an author appeared in the graph). To overcome this hinder, we modify the edge weighting scheme in a way to make up for this otherwise lost information. In the *Impact Graph*, the citations received by an article, from the period it was published until time period t_n are not all equal. The most recent can be considered more important as they signify the paper's direct impact in current science. As a result, the definition of the edge weight in the *Impact Graph* is now given by:

$$WE_{imp}(x, y) = \sum_{\forall i \in P_x \cap P_y} \frac{\alpha \cdot \sum_{t_k=t_0}^{t_n} (w(t_k) \cdot cit(i, t_k)) + \beta}{aut(i) \cdot (1 + t_n - per(i))} \tag{2}$$

where

$$w(per(i)) = \frac{1}{1 + t_n - per(i)} \tag{3}$$

Normalizing citations based on time is a common principle in bibliometrics (Bornmann & Marx, 2015). It has also been applied in conjunction with bibliometric networks (Andersen et al., 2015), serving as an assessment measure for research groups derived from Louvain community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Depending on the dataset that is analyzed and the task, a normalization of citations by document type, publication channel type or research field can be useful and must be applied at this stage. Here WE_{imp} measures the aggregated impact of a list of papers co-written by authors x and y. In essence, the impact of an article at a given time t_n is proportional to the citation it received until that time, and inversely proportional to the number of authors that co-authored it and to the number of years that have passed until its publication. To give an example, if authors x and y have co-authored two articles i and j, with $per(i) = 2000$ and $per(j) = 2010$ then in the *Quantity Graph* of year 2015 the edge weight will be $TWE_{quan}(x, y) = w_{2000} \cdot 1 + w_{2010} \cdot 1 = \frac{1}{16} + \frac{1}{6}$.

Table 1
Graphs that are used in the proposed methodology.

Name	Process	Information
Quantity graph	Aggregate nodes' edges based on Eq. (1)	Authors' collaborations weighted by volume
Impact graph	Aggregate nodes' edges based on Eq. (3)	Authors' collaborations weighted by impact/success
Quantity *Power Graph*	Run *Power Graph Analysis* on the Quantity graph	Extended authors' egonet weighted by volume
Impact *Power Graph*	Run *Power Graph Analysis* on the Impact graph	Extended authors' egonet weighted by impact/success

Similarly, if the article *i* has initially received *cit*(*i*, 2000) citations in 2000, then received *cit*(*i*, 2005) citations in 2005 and another *cit*(*i*, 2010) in 2010, whereas the article *j* has received *cit*(*j*, 2010), then the impact of the collaboration will be:

$$WE_{imp}(x, y) = \frac{\alpha \cdot \left(\frac{cit(i,2000)}{16} + \frac{cit(i,2005)}{11} + \frac{cit(i,2010)}{6}\right) + \beta}{2 \cdot 16} + \frac{\alpha \cdot \left(\frac{cit(i,2010)}{6}\right) + \beta}{2 \cdot 6}$$

The choice of values α and β, denotes the interest on the impact of an author's work (α) or on the quantity of his/her publications (β). Several metrics and techniques have been proposed in the past for balancing between impact and quantity (Bergstrom, 2007), as well as for jointly examining the impact of authors, journals and publications separately or all of them in tandem (Bini, Del Corso, & Romani, 2010). For simplicity, in this work, we follow a weighting scheme that sets $\alpha = 0.7$ and $\beta = 0.3$ in an attempt to prioritize impact over quantity. Of course, further study and possibly a training scheme is necessary here, if we are interested to fine-tune these two weights or train a classification model that distinguishes among authors of different styles. In addition, the obvious difference in the two model's complexity stems from the fact that we aimed to inspect whether a traditional and simplistic approach could be surpassed by a more sophisticated one. The superiority of the latter is far from indisputable, since there are many cases where simple models can indeed capture the reality much better than more complex ones, as Occam's razor states. Our approach to this hypothesis relies on which of the two graphs will prove more vital in the clustering process.

3.1.2. Co-authorship Power Graphs

The most important contribution of the *Power Graph* model is its ability to group several nodes into *Power Nodes* and to aggregate edges into *Power Edges* based on three main *Power Graph* motifs, namely *star*, *clique*, and *bi-clique* may. In the *star* motif a *Power Edge* connects an author with a set of co-authors. The *clique* motif corresponds to a clique of authors that frequently publish papers together and the corresponding *Power Edge* is a loop to the *Power Node* itself. In the *bi-clique* motif, a *Power Node* groups two or more *stars* and as a result the *Power Edge* connects two distinct author sets whose members have published papers together (one author from each set). Finally, *Power Graphs* support another motif which can be very useful in our application. This is the *Power Node inclusion motif* (Royer, 2010), where a *Power Node* contains another *Power Node* and several more distinct authors.

Power Graphs define weights both for the *Power Nodes* and the *Power Edges*. The information that can be extracted from a power graph, pertains to an author's collaboration with strong individuals or groups and it is proportional to the weight of the collaboration itself. *Power Graphs* also provide information about the extended co-authorship society that the author belongs to, via the connected *Power Nodes*. So, it is important to define the *Power Node* and *Power Edge* weights properly in order to draw useful features from the *Power Graphs*. In our methodology *Power Graphs* result from the respective co-authorship graphs, using the same *Power Graph Analysis* technique.

In the case of the *Impact Power Graph* the weights measure the cumulative impact of a group of authors or their collaborations, whereas in the case of the *Quantity Power Graph* they measure the respective productivity of an author, a group of authors or a specific collaboration. Table 1 summarizes the types of the graphs used, describing the information in the originally constructed graphs, and in their *Power Graphs* transformation.

To illustrate better the notions of the aforementioned graphs, and what they represented, let us consider the example of an author *x*, who belongs in *Power Node* χ. In the *Power Graph* there will be a clique motif that connects *Power Node* χ with itself using a *Power Edge*. The weight of the *Power Edge* depicts the strength of the collaboration among authors in the clique, whereas the weight of the *Power Node* depicts the size of the clique. Similarly, there will be *Power Edges* that connect χ with other *Power Nodes*, e.g. ψ or ζ.

Given the *Quantity Power Graph*, we define for an author *x*, who belongs in *Power Node* χ the following features:

$$Sociability(x) = WPN(\chi) \cdot WPE(\chi, \chi) \tag{4}$$

where *Sociability* takes into account the size of the collaboration group that *x* belongs to (the weight of the *Power Node*) and the collaboration strength of authors within the clique. The collaboration strength in the *Quantity Power Graph* represents the number of coauthored articles.

$$PSoc(x|x \in \chi) = \sum_{\forall \zeta \in PN_{\chi}} WPN(\zeta) \cdot WPE(\chi, \zeta) \tag{5}$$

where *PSoc* is the Potential Sociability or the extended clique. This model aggregates the power of all potential collaborators, meaning the authors contained in the *Power Nodes* that connect to χ through a *Power Egde*, multiplied by that *Power Egde*'s weight, as the stronger the connection, the higher the probability the author will work with them eventually.

Given the *Impact Power Graph*, we define for an author *y*, who belongs in *Power Node* ψ the following features:

$$Impact(y) = WPN(\psi) \cdot WPE(\psi, \psi) \tag{6}$$

where *Impact* takes into account the size of the clique that *y* belongs to (the weight of the *Power Node*) and the cumulative impact of all the collaborations between authors within the clique. The collaboration impact between two authors in the case of the *Impact Power Graph* is based on the respective edge, hence it is the impact of their paper.

$$ComImpact(y|y \in \psi) = \sum_{\forall \zeta \in PN_{\psi}} WPN(\zeta) \cdot WPE(\psi, \zeta) \tag{7}$$

where *ComImpact* is the Community Impact, and acts in similar fashion to (5).

3.2. Key performance indicators for authors

Each researcher has a set of features that depict his/her success as an individual, without taking into consideration the co-authors' network. Such features generally account for the productivity, or else the volume and frequency of publications, and the impact they have, which is calculated using citations. Both productivity and impact have a strong temporal dimension, as highlighted in Section 3.1.1, which must be considered when building an author's profile.

Weighted cumulative productivity: The number of articles written by an author *x* from t_0 until a given time t_n, weighted by oldness (i.e. by the periods that have passed from the publication period of each article until t_n).

$$Productivity_{weighted\_cumulative}(x) = \sum_{\forall t \in t_0}^{t_n} \frac{1}{(t_n - t) + 1} \cdot pub(x, t) \tag{8}$$

Weighted cumulative impact: The number of citations made to the articles of author *x*, weighted by oldness (i.e. by the periods that have passed from the period of each citation until t_n).

$$Impact_{weighted\_cumulative}(x) = \sum_{\forall t \in t_0}^{t_n} \sum_{\forall j \in P_x} \frac{cit(t, j)}{(t - per(j)) + 1} \tag{9}$$

To give an example, assuming we are examining the Weighted Cumulative Impact of author *x* back in the year 2004 ($t_n = 2004$) and the author has written article *j* at 2001 ($per(j) = 2001$), which received the following citations: 2 in 2001, 20 in 2002, 30 in 2003 and 15 in 2004. The impact of article *j* is:

$$\frac{2}{2004 - 2001 + 1} + \frac{20}{2004 - 2002 + 1} + \frac{30}{2004 - 2003 + 1} + \frac{15}{2004 - 2004 + 1} = 37.16.$$

One year later ($t_n = 2005$) we examine again the Weighted Cumulative Impact for the same author, who has not published another article but has received 3 more citations in 2005. The new Impact will be:

$$\frac{2}{2005 - 2001 + 1} + \frac{20}{2005 - 2002 + 1} + \frac{30}{2005 - 2003 + 1} + \frac{15}{2005 - 2004 + 1} + \frac{3}{2005 - 2005 + 1} = 25.9.$$

As we can see, the weighted cumulative impact of a specific work decreases over time when the work stops receiving citations or when the volume of citations drops. The concept behind this definition is to penalize citations received many years ago, compared to citations received in the near past and implicitly measure the duration of a work's impact.

3.2.1. Collaborative features

All the features that follow are based on information from collaboration graphs (i.e. Quantity and Impact Graphs and Power Graphs). When measuring features that relate to the number and frequency of collaborations (e.g. sociability, strength) we can either use the impact or quantity graphs with the same final result. However, the features that measure the impact of an author's collaborations are solely based on the impact graphs.

Sociability: The sociability of an author *x* at period t_n accounts the number of co-authors of *x* until period t_n and is the degree of *x* (Everett & Borgatti, 1999) (i.e. the number of edges) in the Impact Graph (though it would be the same from Quantity).

$$Sociability(x) = |C_x| \tag{10}$$

Centrality: The centrality of an author *x* at period t_n measures the sociability of *x*, which is directly linked with that of his/her co-authors.

Table 2
Table of features.

Feat. No.	Name	Origin	Short explanation	Eq.
1	$Productivity_{cumulative}(x)$	Database (Scopus)	The amount of papers an author had written until the current year	
2	$Productivity_{weighted_{c}umulative}(x)$	Database (Scopus)	The amount of papers an author had written penalized by oldness	(8)
3	$Productivity_{current}(x)$	Database (Scopus)	The amount of papers an author wrote at the current year	
4	$Impact_{cumulative}(x)$	Database (Scopus)	The sum of the citations that the authors' papers had gotten until the current year	
5	$Impact_{weighted_{c}umulative}(x)$	Database (Scopus)	The sum of the citations that the authors' papers had gotten, penalized by oldness	(9)
6	$Impact_{current}(x)$	Database (Scopus)	The sum of the citations that the authors' papers got at current year	
7	$Sociability(x)$	Impact Graph	The degree of each author in the Impact graph of current year	(10)
8	$Centrality(x)$	Impact Graph	The eigenvector centrality of each author in the Impact graph of current year	(11)
9	$CLW(x)$	Impact Graph	The sum of the author's edges weight in the Impact graph of current year	(13)
10	$PN_w\{impact\}(x\|x \in \chi)$	Impact Powergraph	The weight of the power node the author belongs to, in the Impact power graph of current year	(14)
11	$PN_c\{impact\}(x\|x \in \chi)$	Impact Powergraph	The clique weight of the power nfode the authors belonged to, in the Impact power graph of current year	(13)
12	$PN_w\{quantity\}(x\|x \in \chi)$	Quantity Powergraph	The weight of the power node the author belongs to, in the quantity power graph of current year	(13)
13	$PN_c\{quantity\}(x\|x \in \chi)$	Quantity Powergraph	The clique weight of the power node the authors belongs to, in the quantity power graph of current year	(14)

Thus an author's weight is proportional to the sum of weights of his neighbors normalized by λ, which is the maximum weight in the graph.

$$Centrality(x) = \frac{1}{\lambda} \cdot \sum_{\forall n \in C_x} (Centrality(n)) \tag{11}$$

Weighted collaboration impact: An author's *x* Weighted Collaboration Impact measures the impact of the author's collaborations at a given time. It is the sum of weights of all edges that contain *x* in the Impact graph.

$$Impact_{weighted\_collab}(x) = \sum_{\forall y \in C_x} (WE_{qual}(x, y)) \tag{12}$$

3.2.2. *Power graph features*

The collaboration graphs give us an idea on the sociability of an author and the impact of his/her direct collaboration. Powergraphs generated from these collaboration graphs allow us to examine the potential of an author's collaborations, under the prism of the extended co-authorship society that the author belongs to. Thus, we assume an increased potential for authors who belong to an eminent co-authorship group or are part of a scientific (sub)network of high impact. Each of the following features are defined both for the impact and the quantity powergraph.

Author PowerNode weight: An author's PowerNode weight APN_w measures the volume or impact of papers published by author's *x* close community (authors that belong in the same power node χ with *x*.

$$APN_w(x|x \in \chi) = WPN(\chi) \tag{13}$$

Author PowerClique weight: An author's PowerClique weight APC_w measures the volume or impact of papers published by author's *x* wider community (including the co-authors of co-authors of *x*).

$$APC_w(x|x \in \chi) = \sum_{\forall \mu \in PN_\chi} WPN(\mu) \cdot WPE(\mu, \psi) + \sum_{\forall \nu \in PC_\chi} WPN(\nu) \tag{14}$$

3.3. *Clustering*

3.3.1. *Features that capture author dynamics*

All the features described in Section 3.2 and summarized in Table 2 measure the scholar's scientific performance either within a single period or in a set of periods. In order to capture the dynamics of an author's performance, we define an additional set of "change indicators". In the equations that follow, $f(t_i)$ is the value of a feature in a period t_i and t_n is the period we examine.

Min change: The minimum change of an author's feature value between two consecutive periods ($minChange_f$).

$$minChange_f = \min_{i \in \{1,n\}} (f(t_i) - f(t_{i-1})) \tag{15}$$

Max change: The maximum change of an author's feature value between two consecutive periods ($maxChange_f$).

$$maxChange_f = \max_{i \in \{1,n\}} (f(t_i) - f(t_{i-1})) \tag{16}$$

Last period change: The change of a feature *lastChange_f* from the last period depicts the author's dynamics at time t_n.

$$lastChange_f = f(t_n) - f(t_{n-1}) \tag{17}$$

Sum of changes: The sum of the features changes through the periods *totalChange_f*, portrays author's stability and shows the total change through time.

$$totalChange_f = \sum_{\forall i \in 1}^{n} ((f(t_i) - f(t_{i-1})) \tag{18}$$

Representative value: The four previous change indexes capture the change that a feature has undergone but in order to include the overall value of the feature, we create a last index that aspires to represent the level the feature oscillates in throughout time. The whole meaning of this is to refrain from categorizing the authors solely based on changes. A toy example of such mistake's results would be clustering an author with a sequence of 1000, 1001, 1002 citations in the same cluster with one that has 0, 1, 2 citations. The representative value derivation is strongly dependent in the initial's feature nature. The original feature set for an author comprises three different types of features: (a) cumulative features that sum the values of all periods, (b) penalized cumulative features that sum the values using decreasing weights for older periods' values, (c) current period features that have a value which corresponds to the performance of an author for a single period (the period examined each time).

In order to compute the representative value for a feature *f*, from period t_0 to the current period t_n, when $f(t_i)$ is the value of the feature in period t_i for the author, we use the following formulas:

- Cumulative features (cumulative productivity and impact): The representative value of this type of features for the whole period (t_0 to t_n) is actually equal to the value of the feature at t_n divided by the number of periods n, resembling an average-like estimation.

$$reprVal_{f_{cumulative}} = f(t_n)/n \tag{19}$$

- Penalized cumulative features (weighted cumulative productivity and impact, all the sociability related features): The sociability related features extracted from graphs and powergraphs consider the effect of time, when assigning edge weights and consequently fall into this category. These types of features are derived from the sum of their past values and divided by time, hence, their representative value equals to the feature's value in the last period, which already resembles an average.

$$reprVal_{f_{wcumulative}} = f(t_n) \tag{20}$$

- Current period features (current productivity and impact): For features of this type, and for a set of periods from t_0 to t_n, we have a set of n values, each of which corresponds to a period in time, and is independent to the previous values. The overall change is defined as the average value summed with the overall feature's trend.

$$reprVal_{f_{curr}} = \sum_{\forall i \in 1}^{n} \frac{f(t_i)}{n} + trend_{f_{curr}} \tag{21}$$

By definition, trend attempts to define the evolution of the feature in time so $trend_{f_{curr}}$ is defined as the gradient (inverse tangent of slope, in radians) of the linear regression that fits the feature's values across periods. Consequently, the change of trend is linear to the values of the series and summing it up gives a suitable boost/downfall to the sum. This is the reason why we add trend in Eq. (21) instead of multiplying it.

Using the aforementioned indicators on the 13 features of Table 2 we create an author's final feature vector, which comprises 65 features in total. This extended set of features aspires to capture in parallel the dynamics of an authors' performance in terms of productivity, impact and sociability. We will refer to the dataset made of these feature vectors as the evolution dataset.

3.3.2. Clustering algorithms

K-means Algorithm: Although more sophisticated (partitioning or agglomerative) clustering algorithms can be used, we decide to use K-means, because of its simplicity, speed, scaling and efficiency. K-means is an iterative partitioning approach that employs a given distance measure and the number of clusters k as input. The distance metric that we use is Euclidean distance, which is recommended for continuous values. In order to find the optimum number of clusters, k, we run the algorithm for different k values and keep the one that minimizes some internal clustering validity criteria. In the absence of a ground truth on the clusters that authors belong to, using internal validity indexes to evaluate our clustering schemes is the only option.

In order to use K-means, we map the set of 5 × 13 features resulting from the change indicators step to 13 new features, each one representing an initial annual feature while aggregating the information of its five indicators. This new type of feature is a linear combination of the first four change indexes multiplied by the representative value:

$$aggregated_f = (minChange_f + maxChange_f + lastChange_f + totalChange_f) * reprVal_f \tag{22}$$

This formula can be justified by taking into account that all four change indexes are equally important in modeling the feature's dynamic while the representative value is the most vital, as it distinguishes the feature's level.

3.3.3. Tuning K-means

Since we did not know a priori the number of clusters present in our data, we had to run a hyperparameter tuning-like process to reveal it. Thus, we run experiments with reasonable numbers of clusters and assessed them with a combination of Average Within Cluster Sum of Squares ($WCSS_{avg}$), Davies–Bouldin index and Average Distance Between Cluster Centroids ($DBCC_{avg}$) which are well known cluster validity indices.

3.4. Feature selection

Feature selection is a rudimentary mechanism in data mining and analytical statistics, which aims in producing a subset of the initial feature set, distinguishing the features of greater importance or influence to the model.

In our case, we want to identify the most important features for distinguishing the type (cluster or class) of an author, without knowing the actual classification. In this unsupervised case, we employ singular value decomposition (Golub & Reinsch, 1970) in order to extract the subset of the most informative features for our set of authors.

Singular value decomposition Singular value decomposition is a well known technique to decompose a dataset, identify the singular values explaining most of the variance and then transform it back to its original basis, achieving dimensionality reduction. The decomposition of a dataset X looks like this:

$$X = U * D * V^T \tag{23}$$

where the columns of the right matrix V are orthogonal to the columns of the left matrix U and D is a diagonal matrix with the singular values. Each singular value in D explains a percentage of variance in the dataset. Each of the columns of the right matrix is a right singular vector, with length equal to the number of features and intuitively shows which features contribute to the variance of the corresponding singular value. In this way, we can extract the columns that contribute the most to the overall variance of the dataset and, consequently, to the clustering. This can be achieved by keeping the first singular values that capture an adequate amount of variation and then finding the values of each column in the corresponding right singular vectors. Summing the right singular vector values, for each "strong" singular value, results in a vector with one value for each column, depicting the amount of contribution this feature has to the most of dataset's variance. Keeping the ones over a certain threshold, will reveal a set of characteristics that are important for clustering authors into groups.

4. Application, evaluation and analysis of results

Our methodology neither ranks authors based on a "rising-star" score nor classifies an author as "rising" or "non-rising" star, but rather clusters authors according to their performance indexes. As a consequence, a typical classification or information retrieval evaluation process that uses a set of authors for training and another set of authors for testing cannot be applied. Moreover, it is more important to see how a detected "rising-star" evolves over time, compared to any other author and this is what related works do (Daud et al., 2013; Li et al., 2009). According to our knowledge, there is not a golden standard collection that we could use as a benchmark for evaluating our methodology. So, we had to create a dataset that comprises all the publication and citation data for a cross-disciplinary set of authors. Since our method clusters authors, we had to evaluate how the different clusters detected at a certain point in time will stand after several years have passed. So, we split the dataset in two periods. We used the data from the first period for extracting authors' features and performing the clustering and data from the second period for comparatively evaluating the performance of predicted "rising stars" and "non-rising-stars".

For testing our methodology, we used the Scopus bibliographic database. The subset of scholars, where we applied our methodology was extracted from the publication and citation data that we kept for the years 1998–2005 and contains in total 43,567 scholars affiliated with 20 different universities or higher education institutions in Greece. Using this dataset we created the publication profiles, the co-authorship graphs and *Power Graphs*, which of course includes all their co-authors from universities all over the world, and clustered the authors as discussed in the previous sections. We evaluated our findings by comparing the overall behavior of authors in the different clusters in terms of collaborations, citations and publications using publication and citation data in the years that followed (2006–2013).

The main findings show that: (a) the authors can be coarsely divided into 7 well separated clusters, with clear differences in at least two of the three central dimensions, meaning their productivity, impact and sociability profiles, (b) the sociability features are of equal importance to the productivity and impact features in defining a "rising star", (c) the weighted cumulative features are the most informative features in detecting "rising stars", which means that such authors demonstrate a

Table 3
The distribution of authors in the dataset by affiliation. The first column shows the percentage of the authors in the dataset that are affiliated with the respective academic institution.

Authors' ratio	Name
21.5%	University of Athens
24.5%	Aristoteleion Panepistimion Thessalonikis
12.0%	Ethniko Metsovio Polytechnio
10.2%	Polytechnion Kritis
2.1%	Panepistimion Patron
4.5%	Panepistimio Kritis
4.6%	Panepistimion Ioanninon
4.6%	Dimokrition Panepistimion Thrakis
3.9%	Panepistimio Thesalias
2.2%	Panepistimion Aegaeou
3.2%	Geoponiko Panepistimion Athinon
1.4%	Panepistimion Pireos
1.3%	Ikonomikon Panepistimin Athinon
1.0%	Panepistimion Makedonias
0.5%	University of Peloponnese
1.1%	Harokopio Panepistimio
0.6%	Hellenic Open University
0.4%	Ionian Panepistimion
0.4%	Panteion Panepestimion Ikonomikon kai Politicon Epistimon
0.1%	University of Central Greece

Table 4
The distribution of venues (Conference Proceedings and Journals), from year 1981 until 2005.

Conference Proceedings	Journals	Year
0	472	<1981
1	272	1981
0	301	1982
2	328	1983
6	414	1984
9	543	1985
14	557	1986
12	669	1987
21	741	1988
23	848	1989
14	842	1990
31	1140	1991
18	1108	1992
58	1256	1993
65	1547	1994
62	1715	1995
156	2624	1996
145	2778	1997
115	2993	1998
145	2959	1999
166	3193	2000
222	3549	2001
210	3914	2002
248	4654	2003
497	5764	2004
772	6591	2005

continuous "above-average" academic performance throughout the years, (d) the clusters identified have consistent performance over the years, and (e) the "rising stars" can be clearly distinguished between the rest of the clusters, taking into account their dynamics.

4.1. Dataset

The dataset was created by collecting information about authors and their publications and citations from the digital library Scopus. The advantage of Scopus is that it provides the actual citations that a paper receives each year and not an indicator of impact, such as conference or journal impact factor. Our group of study are the authors that had a Greek University affiliation by 2014, and for this reason we selected the affiliations depicted in Table 3. The table also illustrates the percentage of scholars in the dataset that are affiliated with each of the academic institutions. The total number of papers the authors have written or co-written was 118, 605. Almost half of them (54,241) have been published before 2005. Table 4 presents the number of distinct journals and conferences, by year, in the dataset. The features extracted for each scientific article were: year of publication, title, author list, ISSN, journal of publication, citations per year for each year between 1998 and 2014, and sum of citations received before 1998 and after 2013. In the process of finding all the co-authors of these authors and collecting profile information from them, we resulted with a dataset of 132,031 identified authors, with affiliation, fields of study and the city they work in according to Scopus. The dataset is available upon request, as a csv file, and the SQL, Java and R code to preprocess and manipulate it in order to repeat the analysis, is available at GitHub.[3]

The core units of our methodology are the **articles** and the collaborations of Greek affiliated authors. Since an author may have co-authors from the same country but also international collaborations, measuring the impact of collaborations requires the analysis of the complete co-authorship graph and the publication and impact information for all authors in Scopus database. For simplicity, in our experiments, we retrieve the complete list of publications (and citations) for all Greek affiliated authors and their Greek co-authors, but no information for their non-Greek co-authors. As a result, when an author a collaborates once with a high-impact co-author a_{high} and once with a low impact co-author a_{low} both with a non-Greek affiliation, both collaborations contribute in an equal manner to the indicators of author a (based on the impact of each separate article). The proposed model can support this analysis, since it takes into account the accumulated impact of a co-author in a publication (from all his/her publications), as well as his/her accumulated sociability (from all collaborations). However, several access restrictions did not allow us to collect all the information contained in Scopus database. So the experimental conclusions must be generalized carefully.

4.2. Graph pre-processing and compression

The collaboration graphs for a specific year t_n are constructed by flattening the co-authorship **hypergraph**, where a paper is an edge connecting all authors, to a simple graph, where all pair-wise edges are drawn among the co-authors of a paper (avoiding self-edges). All such edges carry some extra information concerning the number of co-authors in the paper, the publication year and the total citations the paper has from t_0 until t_n.

When two authors collaborate more than once, multiple edges occur in the simple co-authorship graph (**multigraph** (Balakrishnan & Ranganathan, 2012)). As a result, the text files that contain some of the co-authorship graphs were more than 30 Gb, which rendered the run of our proposed algorithms unfeasible without big data technology. To this end, we endeavor a graph compression merging all multiple edges between author pairs to a single edge that aggregates the information for the collaboration between them in the fashion delineated in Section 3.1.1.

4.3. Clustering

4.3.1. Features

All of the monitored features were extracted per year, representing the information we wanted to capture for each of the authors' annual profiles. The individual features described in Section 3.2 are calculated directly from the database using java and SQL. The collaborative features are measured using the *R* library *igraph*.[4] The Impact graph of each year is used to calculate the 3 collaborative features. Finally, for the computation of the *Power Graph* features, we construct a quantity power graph for each year from the respective quantity graph, using the *Power Graph Analysis* library in Java.[5] The same process is executed for the creation of the Impact *Power Graph* from the Impact graph of each year.

To compute the specific *Power Graph* features, a graph structure parser based on the JUNG framework for graph processing[6] was created. The power nodes were handled as regular nodes, and a special class was defined for power edges. The weight of the power node was calculated as a function of its contained nodes (aggregation), which corresponded to authors. The power cliques weight was calculated by finding the neighbors of a power node and summing their weights, each multiplied with the respective edge's weight (edge connecting the power node and the neighbor). This number was then added to the sum of weights of the power nodes containing the examined power node, which were discovered recursively.

This procedure results to a dataset with 13 features for each year, which contains all authors that have published at least one paper until that year.

4.3.2. Author dynamics

In the context of calculating the change indicators for each author in the dataset and produce the 65 feature dataset from the original 13 features set, we examine the annual datasets in pairs of consecutive years. Java' *Math* library[7] was used in this part. If an author is encountered for the first time during the iteration, his/her values are stored as first change, like detracting from zero.

[3] https://github.com/GiorgosPanagopoulos/Detecting-Rising-Stars-in-Dynamic-Collaborative-Networks.
[4] http://igraph.org/r/.
[5] http://www.biotec.tu-dresden.de/research/schroeder/powergraphs/download-command-line-tool.html.
[6] http://jung.sourceforge.net/.
[7] http://commons.apache.org/proper/commons-math/.

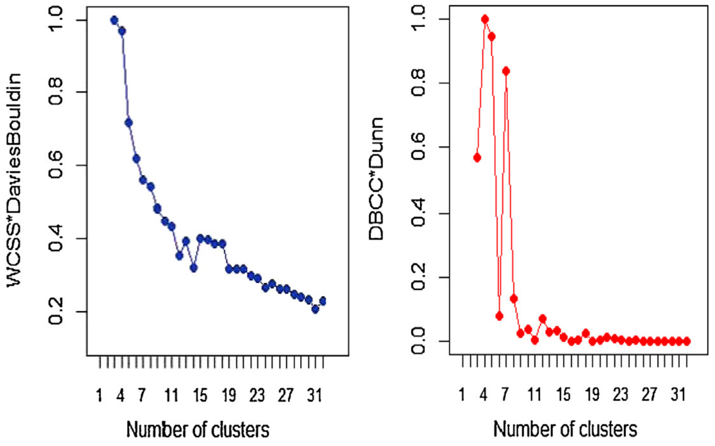


Fig. 2. The plot of the clustering validity metrics $WCSS_{avg} * DB$ and $DBCC_{avg} * DU$ over the different number of clusters. The best cases in $DBCC_{avg} * Dunn$ metric is 4, 5 and 7 clusters. From these three cases, the best in terms of $WCSS_{avg} * DB$ is 7 (the lowest), hence this is the optimum number of clusters, combining the best possible similarity inside the clusters and dissimilarity between them.

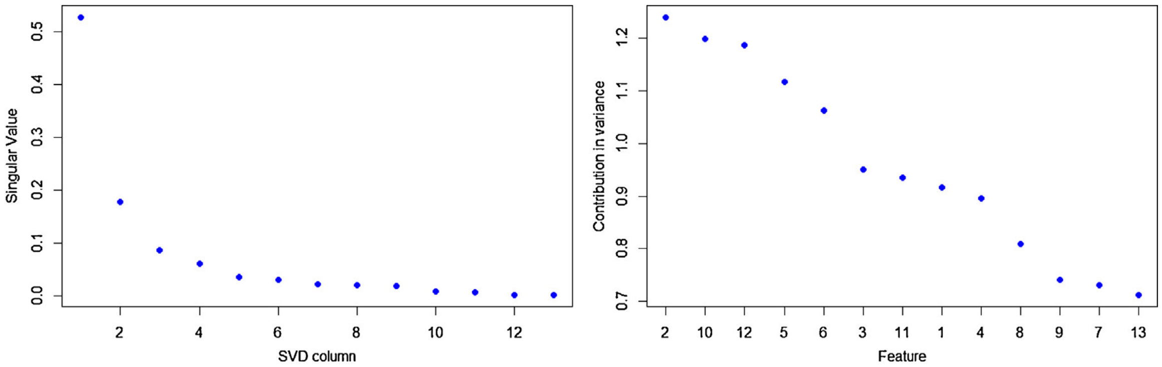


Fig. 3. On the left is the singular value decomposition of the dataset, which shows that the first 4 eigenvalues describe more than 90% of the dataset with small variance. This is an indication that with a limited amount of our selected features we can properly describe and represent the dataset. On the right is the contribution of the individual features in the dataset's variance. The figure presents the feature, with their numbers as defined in Table 2, in decreasing order of contribution to the underlying concepts of the dataset and hence the analogous impact on the clustering.

4.3.3. *Finding the optimal number of clusters*

In order to determine the optimal number of clusters, K-means was executed for $k \in 2 \rightarrow 100$ and calculated the cluster validity scores of Section 3.3.3. The *clusterSim* library in *R*, and more specifically the *index.DB* function were used to calculate the four validity scores (i.e., Average Within Cluster Sum of Square-WCSS, Davies–Bouldin index, Average Distance Between Cluster Centroids-DBCS, and Dunn index). Fig. 2 shows and explains the data based on which the 7 was decided as the optimal number of clusters.[8] This case of clustering results in the best Inter-Intra cluster distance combination and in the most cohesive and discrete clusters.

4.3.4. *Selection of the best features*

In order to distinguish the most influential columns as described in Section 3.4 and detailed in Wall, Rechtsteiner, and Rocha (2003), we applied singular value decomposition. Fig. 3 depicts the singular values extracted for the dataset. From this plot we can see the first 4 singular values enclosing the majority of dataset's variance.

This means that we must consider the first 4 singular values and keep the features having the biggest impact on them, based on the corresponding right singular vectors. Fig. 3 shows the contribution of each of the 13 features to the first four eigenvalues. According to this plot, the number of papers an author has written penalized by their oldness is the most informative characteristic, followed by the impact of the author's clique, the quantitative power, and the penalized impact due to citations received. Our results disagree with the analysis of Fiala, Šubelj, Žitnik, and Bajec (2015) where page-rank based methods, relative to our graph-based ones, were found inferior to citation count.

[8] Results for 30 to 100 clusters are omitted since the metric values do not change much.

4.3.5. Cluster labeling

The last step was the labeling of the clusters based on their principle characteristics. To accomplish this, we first created two thresholds for each of the 65 features in the dataset. For each feature, we extract its 0.85 and 0.15 percentile values among the authors in all the 7 clusters. The *n*th percentile value for an observation variable is considered the value that is higher than the *n* percent of all the observed data values. These values for a selected feature can quickly show whether the cluster's centroid value for this particular feature can be considered high or low. Having two vectors of high and low thresholds, consisting of 65 values each, we examined all clusters' centroids to find out the features distinguishing them. We then characterized each cluster based on the marked features, as follows:

- Cluster 1 (394 authors): Highest number of average papers and citations with increasing rate but low minimum changes. Also highest clique's impact and quantity. Powerful authors belonging to strong communities.
- Cluster 2 (33,532 authors): Low maximum and high minimum changes in social indicators. In other words steady, minuscule increase in the social aspect. Scant dynamics in paper fertility. It seems to be the typical, most common profile of a scholar, i.e., the profile of the average author/scholar.
- Cluster 3 (1,161 authors): Steady positive change indicators in productivity and high values in changes of citations, especially in last change. Cluster of increasingly successful authors, but still socially moderate, as they lack an extended coauthorship community.
- Cluster 4 (2846 authors): Low last but high maximum change in social metrics. Otherwise similar behavior to cluster 2. Group of average scholars with fairly active collaborations in the past but currently withering social.
- Cluster 5 (544 authors): Substantial average value and rate in clique's impact and quantity, with high dynamic in citations. Rising citation receivers, associated to groups with high impact, influence and potential.
- Cluster 6 (3958 authors):The smallest sum of changes in all social metrics, depicting a steady behavior in the social aspect. Otherwise similar behavior to cluster 2. Group of socially stable authors and moderate producers.
- Cluster 7 (1132 authors): Considerable sum and last change in degree and power node weight, showing increasing dynamic in the quantity and impact of the clique. Limited number of papers and citations with low increase rate. Lower profile authors belonging to upcoming co-authorship groups.

The above cluster descriptions (cluster labels) are the results of an analysis of the feature values of each cluster that we performed using a Shiny[9] visualization[10] that we developed for this purpose. The visualization is interactive and allows to examine the feature values for each cluster centroid, to compare them with the respective values of the high and low percentiles for each feature in the whole set and to clarify which are the strengths and weaknesses of that cluster.

4.4. Evaluation

Since our data and consequently our methodology, do not include the information of journal impact, it cannot be directly compared with the ones mentioned in the Related Work, because they cannot be applied to our data. Moreover, our method is fundamentally built on citations and their chronology, rendering it unable to run to datasets like DBLP. Therefore we endeavor a qualitative evaluation of the clusters in terms of coherence, meaning whether each cluster withholds its characteristics through the time, as well as in terms of validity, showing that the distinguished clusters maintain substantial differences with each other in key factors.

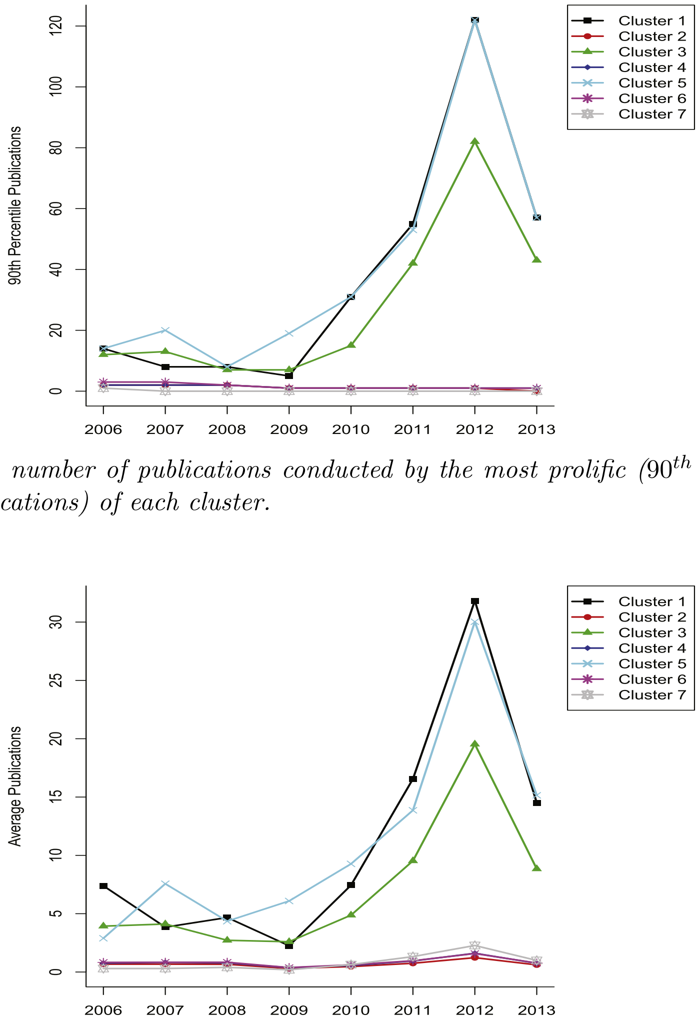
4.4.1. Evaluation of coherence

In the context of evaluating the cluster coherence, an exploratory analysis was conducted examining the three most central dimensions of our analysis, the number of citations, coauthors and publications. To avoid bias from outliers and focus on the general future performance, We kept track of the average value of these metrics in each cluster, over the years 2006 to 2013, to assess whether their overall behavior stays consistent in time as well as if it agrees with the labeling we derived and investigate any potential diverges from it. In addition to the average, we employed the 90th percentile, which corresponds to the value below which stands the 90% of the authors in the particular cluster based on the metric under examination. We preferred presenting a high percentile than a lower one because we want to underline the most prominent performance of each cluster, while the general image is conceived through the average. To be more specific, in Fig. 4 one can see that cluster 1 starts stronger in 2006, but cluster 5 quickly catches up and sometimes surpasses it, with cluster 3 being inferior by a moderate amount of publications. The weak clusters (2, 4, 6, 7) have limited amount of publications, confirming our labeling. The steep increase in publications after 2010, corresponds to the increase of data for that period, which as mentioned above, prohibited us from using these years for training purposes.

In addition, in Fig. 5 the inter cluster differences become clearer, together with the immense divergence between the week and strong clusters. The reason behind this disparity is that in general, the strong researches tend to have much more citations than the weak ones, forming a big gap. Another interesting phenomenon in this case, is that while the behavior of

[9] http://shiny.rstudio.com/.

[10] https://georgepanagopoulos.shinyapps.io/ClusterLabeling/.



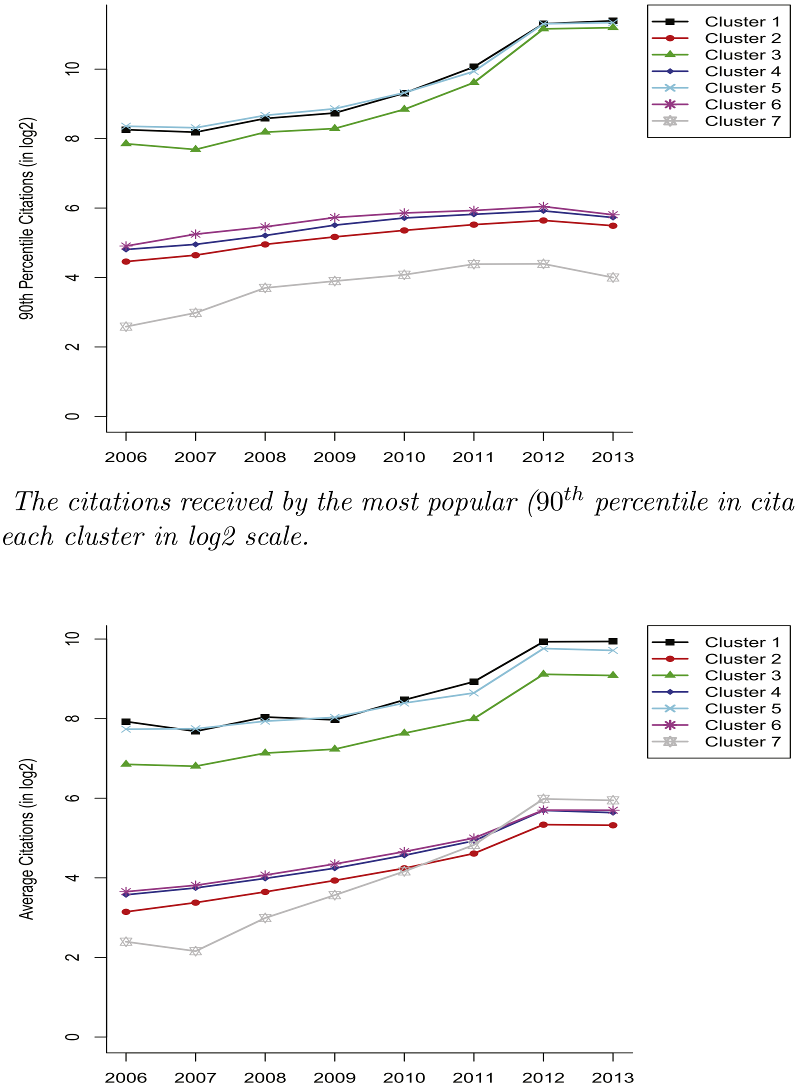
(a) *The number of publications conducted by the most prolific (*90^{th} *percentile in publications) of each cluster.*

(b) *The number of publications conducted by the average author in each cluster.*

Fig. 4. The number of publications of two representative values for each cluster. The pattern is rather similar in both cases, showing the achieved accuracy in the entirety of each cluster for this metric. (a) The number of publications conducted by the most prolific (90th percentile in publications) of each cluster. (b) The number of publications conducted by the average author in each cluster.

the strong and the average agree for the strong clusters, the weak ones undergo a contrast. Namely, the average authors tend to get more citations, while the stronger ones are bounded, which proves the correspondence to authors of limited success.

Furthermore, the behavior of the clusters in terms of number of coauthors seems to be rather interesting. In Fig. 6, the first thing one may observe, is that for weak clusters the average value surpasses the 90th percentile. This signals the existence of very strong outliers in that dimension, but being very few, the 90th percentile fails to reach them. A very representative cluster in this case is 7, which was labeled initially as languorous in all aspects except their community dynamic. Through detailed investigation we found out this cluster was solely based on very weak authors, with an exception of <1% (57 out of 1132) being exceptionally strong in the social aspect. In Fig. 5(b), cluster 7 though pure in 2006, ultimately surpasses in average citations the rest of weak clusters. The pattern does not agree with Fig. 5(a), for the aforementioned reason regarding the few outliers that shape the average value. However this reveals that these outliers were greatly benefited from their extended collaborative environment in the long run, possibly showing a latent correlation between its dynamic and future citations. Overall the behavior of the strong clusters seems stable and coherent in time, with the average value oscillating a little more than the 90th percentile, something quite intuitive, since generally the stronger the author the most permanent and prolific their collaborations are. A possible argument is that some clusters could be merged in one, specifically clusters 1 and 5 as well as 6 and 4. For the former couple, which is most pivotal since the one part corresponds to rising stars, in terms of citations and publications this might pose a valid argument, however cluster 5 lacks the coauthors of cluster 1, as is prevalent in Fig. 6b, and though the difference does not seem so extended in first sight, given that it is in log scale and that



(a) *The citations received by the most popular* (90^{th} *percentile in citations) of each cluster in log2 scale.*

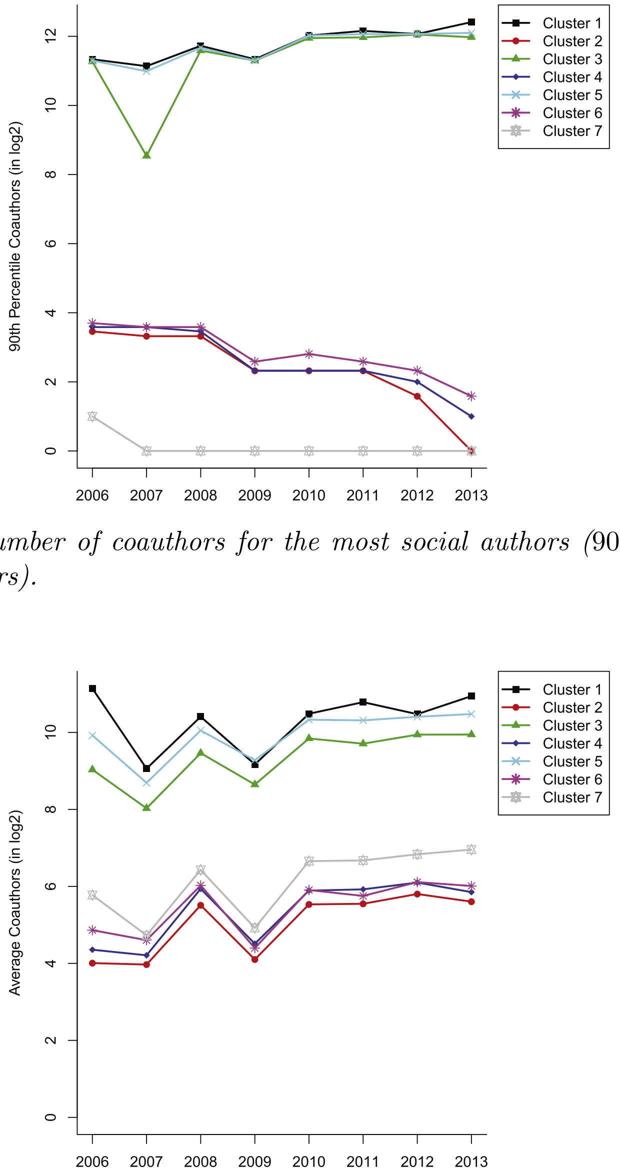
(b) *The citations received by the average author in each cluster in log2 scale.*

Fig. 5. The number of citations of two representative values for each cluster in log2 because of the considerable differences between weak and strong clusters. The pattern in time is prevalent in both plots and the difference between each cluster is distinct.

our line of reasoning is based on the crucial role sociability plays in an author's success, these two clusters should indeed be separate, although it is food for thought and further research. As for clusters 6 and 4, these could indeed be considered one.

4.4.2. Evaluation of validity

The experiments were executed on a desktop PC with 16GB RAM and 2CPUs, which did not manage to process the co-authorship graphs generated for the years after 2008. The number of collaborations in each paper exploded after 2005 since the average number of authors in a publication before 2005 is 7.3, while after 2005 and until 2013 it explodes to 32. As a result, the co-authorship graphs size exceeded 40GB and it was impossible to run Power Graph Analysis in order to compress the graph and get the power graph features, or compute the spectrum of the graph, to calculate the eigenvector centrality as required for extracting the necessary social features. Hence, we unavoidably restraint our analysis in features that do not require the processing of the complete social graph. Fortunately, the most pivotal characteristics according to our analysis in Section 4.3.4 could be derived straight from our database. These features are: penalized productivity, current citations, number of coauthors and productivity. In order to visualize these dimensions in a way to facilitate a thorough comparison between the clusters, we employ radar plots. Penalized productivity is the most important feature and has values in [0, 1], in contrast to the other three features, so we draw a radar plot exclusively for it in Fig. 7. For the other three features, Fig. 8



(a) *The number of coauthors for the most social authors* (90^{th} *percentile in coauthors).*

(b) *The number of coauthors collaborating with the average author in each cluster.*

Fig. 6. The number of coauthors of two representative values for each cluster. The weak clusters exhibit a difference between the 90th percentile and the average, while the strong ones seem to relatively agree in a uniform image, with 90th percentile being a little more stable than the average.

displays where each clusters' values oscillate, in a log10 scale to facilitate a clear depiction of all clusters in one plot. For penalized productivity, we plot the 25th, 50th, 75th and 90th percentile values and the average, whereas for all the other features we plot the 50th, 90th percentile and the average value for each.

Since there is no standard way to decide, whether a predicted "rising star" indeed became a star, there is not a straightforward way to evaluate the validity of our clustering scheme. However, if the prediction algorithm works well, it is expected that the majority of predicted "rising stars" will have a really good record and impact and will clearly distinguish from the respective best authors of other clusters. For this reason, we choose different *n* percentile values for our features (authors that rank above the n% of all authors in the cluster in each specific feature) in order to provide evidence on the difference between the evolution of "rising stars" and "non-rising stars".

Fig. 7 displays several percentiles of the most important feature, according to our analysis. The values of each percentile generally agree with our intuition, i.e. 25th is smaller than average. The weak clusters present a similar behavior with the exception of the 90th percentile were some outliers may draw the cluster higher. For the strong clusters, the average seems

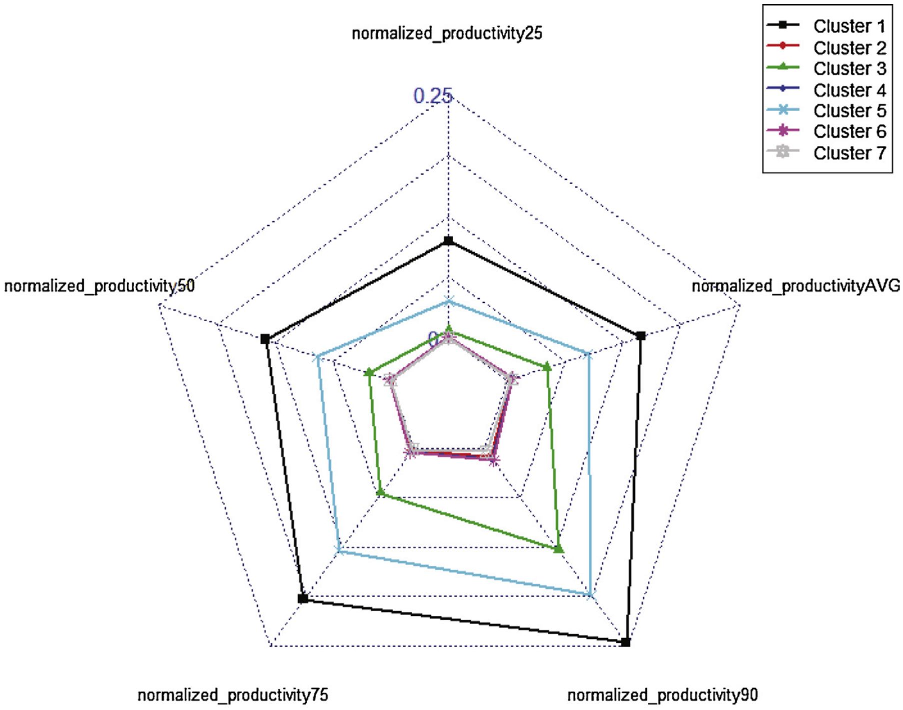


Fig. 7. The clusters' penalized productivity (normal scale) in several percentiles to underline the difference in the entirety of each cluster.

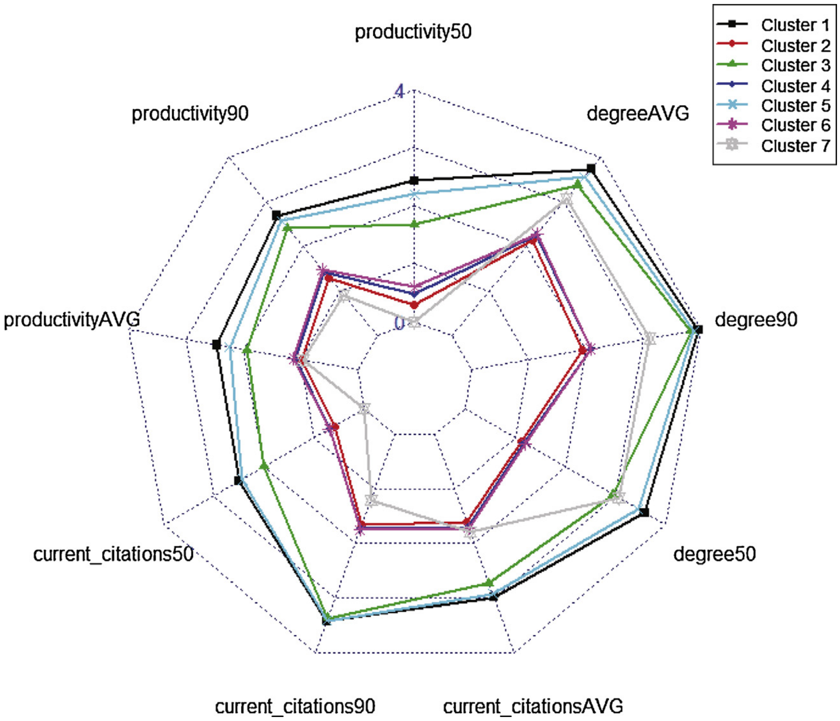


Fig. 8. The average, 50th and 90th percentile of degree, productivity and current citations, in Log10 scale to alleviate the disparity between clusters.

close to the 50th percentile, abating the possibility of outliers effecting the average in a strong cluster. The hierarchy follows the labeling, showing that rising stars stay indeed distinguished in the future.

From Fig. 8 the difference between the strong and the week clusters seems profound in almost every dimension. The only exception is the 50th percentile of degree for cluster 7 which surpasses that of cluster 3, confirming our aforementioned claim about the high collaborations of cluster 7, while cluster 3 clearly surpasses it in average and 90th percentile degree, displaying the superiority of the cluster as a whole. Apart from that, the hierarchy of the clusters is in agreement with the labeling.

Someone may argue that true rising stars should have surpassed the established stars by that time (year 2013) in most features, but there are two main arguments undermining this perspective. First one is that 7 years are not enough for most scholars to reach a star level, which generally addresses a career of decades. Secondly, there are specific features were cluster 5 and 3 approximate the values of cluster 1, namely in the 90th percentile of citations received in current year and the number of collaborations, which intuitively seems right, because the best of the rising stars, reach a star level citation harvest relatively quickly due to popularity, novel ideas and trend. However the productivity and normalized productivity are issues addressing the number of papers a scholar has authored up to that point, the most reasonably difficult aspect for a new star to surpass an established one, because of time limitations.

The clustering scheme comprises 7 clusters, which are compared against several dimensions. Each line in the radar plots corresponds to a cluster and shows how the authors in the same percentile across different clusters compare. Both Figs. 7 and 8 show a clear difference between clusters 1, 3, 5 (strong authors) and the remaining four clusters (weak authors) at year 2013.

From the plot in Fig. 8, and the current_citations50 axis it is evident that half the authors of cluster 5 ("rising stars" cluster) at 2013 have almost the same number of citations (around $10^{2.5}$ citations) with authors in cluster 1 whereas authors in cluster 3 have significantly fewer citations (10^2). Similar results hold for productivity and sociability features among the three clusters. In conclusion, the analysis of the 7 clusters shows increased average numbers of productivity, impact and sociability through the years for authors in clusters 1, 3, 5 and especially for clusters 1 and 5. Additionally, authors in cluster 5 have the most increasing trend among the three clusters. Focusing on the top-50% of the authors in each cluster, it becomes obvious that authors in cluster 5 have managed to compete the best authors of cluster 1, which validates the labeling of our clusters: Authors in cluster 5 where the "rising stars" of 2005 whereas authors in cluster 1 were already well-established.

In Fig. A.10 and Table A.6 in the Appendix, we give the detailed record of publications and citations for the top-4 authors in cluster 5 (in total number of citations and publications until 2006). For comparison we also provide the numbers for the first author in cluster 1. What is worth mentioning for the 4 authors, is that although in the monitoring period (red zone, until 2006) they have a good but stable impact and a raising productivity, they all show an impressive improvement during the evaluation period (green zone, 2006–2013), reaching their maximum yearly citations at 2013.

4.5. *Interpretation of the results*

The main contribution of this work, relies on the methodology it introduces, which examines in tandem productivity, impact and collaboration of authors and how these features evolve in time and generates clusters of authors with similar features instead of producing an overall ranking of authors. In order to test our methodology, we used a specific dataset comprising researchers with Greek affiliation, belonging to various scientific disciplines. This introduces an additional difficulty when attempting to compare between authors or provide a universal ranking, since academics in fine arts and humanities are compared to researchers from natural sciences and engineering, who significantly differ in the way they publish, collaborate or cite their works. Probably, the conclusions drawn from the current clustering will be different from the ones drawn from a dataset complied solely by scientists of a specific field. However, our methodology can be applied in both cases, and produce meaningful results. More specifically, restraining the application of the algorithm in scientists of a specific kind, will arguably produce a fairer clustering, as the noise inherent in our dataset due to the multiple fields will no longer exist. In this section, we provide further details on the clusters and analyze the main venues their authors publish in as well as their most prevalent fields, in order to argue that this approach can also be applied in this kind of mixed data, producing logical results and examining data from a different perspective.

First of all, since we employed a multi-disciplinary author (and publication) corpus, it would be interesting to see what are the main venues, where the authors of each cluster publish and in what degree. For this reason, we present in Table 5 the 5 most preferred venues per cluster i.e. the ones with the most publications from the cluster's authors in the period 1998–2005, the number of publications per cluster i.e. the publications of the authors that belong to the cluster for that period and the number of distinct venues per cluster. For each of the top venues we give the number of publications in parenthesis and the number of distinct authors that took part in these publications in square brackets. Secondly, in order to have a better view of the scientific fields of the authors that populate each cluster, we summarize them per cluster in Fig. 9. The three clusters with the stronger author profiles (i.e. 1, 3 and 5) are in the first row, whereas the other three (2, 4 and 6) are in the middle row and cluster 7 in the bottom row. The height of each bar corresponds to the number of authors in that cluster belonging to that particular field. Here it should be noted that each author belongs to multiple fields. This property, which was inherent in our dataset, was the biggest constraint in tallying a publication to a specific field, as its authors belonged to multiple. However, this might be a more realistic approach, considering that nowadays scientists can blend their areas of interest easier than ever, or collaborate with scientists from other disciplines to produce something meaningful. For example, a computer scientist may assist psychologists to analyze psychometric data and publish in a journal of psychology, or a mathematician can assist computer scientists on a machine learning task and publish in the field of computer science. In addition, there are scientists who turned their careers from one field to another. All these cases contend classifying a

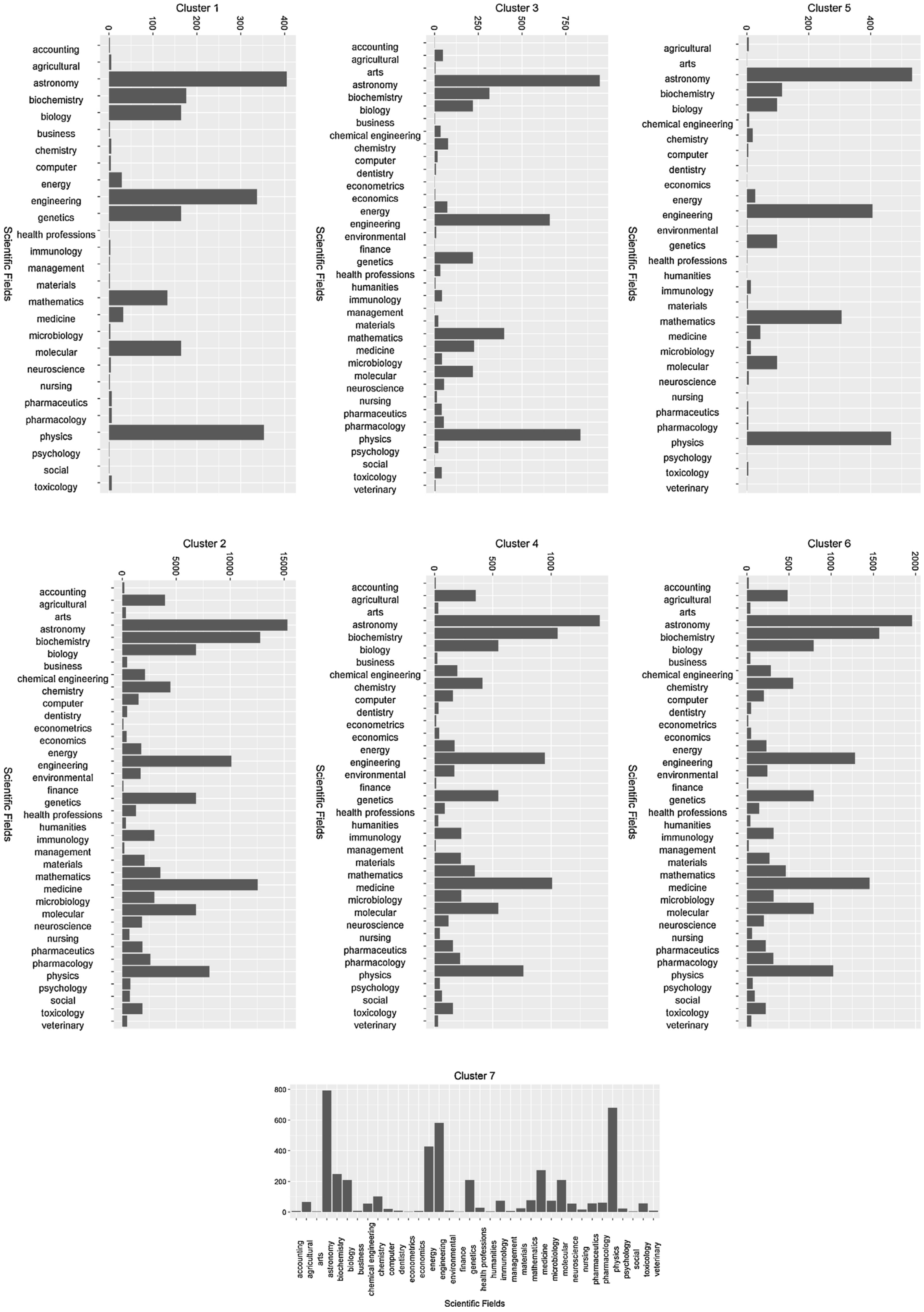


Fig. 9. The fields where the authors of each cluster work on (clusters 1, 3, 5 at the top row, 2, 4, 6 at the middle row and 7 at the bottom row). The horizontal axis counts the number each field appears for an author. An author can be associated with more than one fields.

Table 5
The venues that received the most publications from the authors of each cluster (the number of publications in parenthesis, the number of distinct authors in curly brackets).

Cluster	Distinct venues	Total publications	Top 5 most preferred venues
1	18	37	European Physical Journal C (16){199}, Physics Letters, Section B (6){59}, Lecture Notes in Computer Science and LNAI(1){3}, AIP Conference Proceedings (1){2}, Journal of Mathematical Physics (1){1}
2	2605	6257	Lecture Notes in Computer Science and LNAI (140){273}, Journal of Physics: Conference Series (42){121}, Physical Review B (27){78}, Astronomy and Astrophysics (26){108}, Cement and Concrete Composites (26){43}
3	294	437	European Physical Journal C (16){477}, Macromolecules (12){5}, AIP Conference Proceedings (9){6}, Annals of Oncology (9){5}, Physics Letters, Section B (7){438}
4	994	1529	Lecture Notes in Computer Science and LNAI (23){20}, European Physical Journal C (16){16}, Journal of Physics: Conference Series (14){13}, World Journal of Gastroenterology (11){10}, Journal of Food Engineering (11){7}
5	57	112	European Physical Journal C (16){239}, Journal of Polymer Science, Part A (9){2}, Macromolecules (6){2}, Physics Letters, Section B (5){206}, Thin Solid Films (5){1}
6	1244	2119	Lecture Notes in Computer Science and LNAI (36){21}, European Physical Journal C (16){17}, World Journal of Gastroenterology (14){20}, Physical Review B (12){8},AIP Conference Proceedings (11){13}
7	142	187	European Physical Journal C (16){9}, AIP Conference Proceedings (8){4}, Journal of Materials Processing Technology (4){1}, Physics Letters, Section B (3){9}, Nuclear Physics A (3){3}

scientist in only one specific field a complex and risky assignment. That is why we choose to apply our method in this setting and examine the results.

Looking at Table 5, the first simple observation is that clusters 1, 3, 5 and 7 comprise a limited number of publications when compared to the rest. The majority of publications of cluster 1 is in two distinguished Physics journals and are the result of the collaboration of large groups of researchers. This is typical for publications concerning large experiments in Physics. The results are similar in cluster 5 concerning these two journals, with the main difference lying in the fact that the cluster also comprises publications in high esteemed venues in other disciplines, like chemistry. Venues from multiple disciplines appear in cluster 3 (e.g. Physics, Molecular Science, Medicine). The results in Fig. 9 are in general agreement to those of Table 5. We see that the clusters with “strong” profiles are mostly dominated by authors from the Physics, Engineering and Astronomy domain, whereas the clusters with “weaker” profiles are more multidisciplinary. A noted difference is that although medicine is indeed more prevalent in Cluster 3 than in 1 and 5, as is suggested by the venue analysis, chemists and molecular scientists are not substantially increased in clusters 3 and 5 compared to cluster 1, as would be expected by the venue analysis. This suggests that chemistry and molecular science venues possibly accept publications from authors belonging to multiple fields, like mathematicians, which are analogically more in cluster 3 and 5 than in cluster 1. In addition, cluster 7 seems to have similar behavior with the “strong” ones, in terms of preferred venues and distribution of fields. The main difference when compared with other clusters, is that the number of distinct authors that participated in publications is significantly smaller when compared with clusters 1, 3 or 5, which clearly shows that cluster 7 comprises authors with fewer collaborators, as does Fig. 6. The examination of cluster 7, in comparison with clusters 1, 3 and 5, at a later point in time (i.e. 2013), shows that the increase of productivity and impact its authors is also limited. This validates the role of sociability in the definition of a “rising-star” author, in the sense that authors with many collaborations from a cluster with similar field distribution, achieved greater success than the less social ones in the passage of time.

The weaker clusters exhibit closely homogeneous field distributions and are still overwhelmed by astronomy and engineering, but one can distinguish some basic differences, like medicine surpassing physics. Medicine being so prevalent in cluster 2, 4 and 6 may agree with intuition, as these clusters make up for more than 80% of our authors and altogether represent an average scientist as explained in Section 4.3.5. This finding possibly reveals that many medical doctors publish something, because medicine is a highly academic field, but generally not all of them pursue a scientific career. On the other hand, the physicists pursuing academic achievements, may be few, but have substantial scientific prestige, as shown by their presence in the “strong” clusters. Moreover, the venue analysis of clusters 2, 4 and 6, shows that physical venues receive more publications than medical ones. However the physical venues present in the top five are very general, including possibly astronomy or engineering scientists too, while the only medical venue is the journal of Gastroenterology. This, together with the increased presence of medical doctors in these clusters, suggests that they prefer publishing in journals specific to their medical specialty, resulting in numerous medical venues but with fewer publications each, compared to the physical venues which are limited but receive publications by almost all the spectrum of physics.

Another clear observation is that computer science publications, more specifically in the LNCS and LNAI series dominate clusters 2, 4 and 6. Computer scientists are analogously and literally more in the weaker clusters than in the strong ones, based on Fig. 9. The high number of publications in LNCS, attest to the fact that it includes the LNAI and LNBI subseries. Moreover, one can conclude that computer scientists do not tend to collaborate so much compared to other fields, like physics, due to the substantiate difference in the ratio between publications and authors in the most preferred venues, e.g.

comparing LNAI/LNCS with European Physical Journal. This might be one of the reasons why few computer scientists make it to the strong clusters.

Finally, biochemistry and biology have also more significant presence in the “weaker” clusters. One would expect that fields such as these, would tend towards the stronger clusters, as they generally have high citation and collaboration rates, but as it seems the algorithm did not get heavily biased in this case.

In conclusion, what is evident from this analysis is that researchers in natural sciences are prevalent throughout the dataset. They also achieve positions in the stronger clusters due to their large sociability, since they form large groups in their experiments, and their high citation rates. Although there is a difference between clusters that comprise mainly researchers from this domain (e.g. 1, 3, 5) and clusters that comprise researchers from all domains (e.g. clusters 2, 4, 6), the overall distributions of fields throughout clusters do not differ massively. This on the one hand provides with some interesting conclusions as the ones defined above, while on the other, justifies the need to examine our methods on authors of a specific discipline too, as future work.

5. Conclusions

In this work we presented a novel methodology for extracting and analyzing profiles of scholars based on bibliometric and collaborative information between the scholars. The methodology exploits several aspects of the information we can derive from an authors’ works, and applies a novel pipeline to cluster the authors into categories, generating authors’ profiles and determining automatically the optimal number of clusters. Applying the suggested methodology in a real dataset helped us categorize authors in seven categories, including the rising stars, who stand out for their dynamic and potential. For the implementation of the suggested approach and its evaluation, we crawled bibliographic data with time span, from the Scopus digital library and covered any inconsistencies by applying a technique similar to collaborative filtering. Graph representation and mining techniques allowed us to capture the social, individual and time related facets of an author’s impact, while in tandem extracting various pieces of information regarding the rate of publications and emerging authors, the top authors and their co-authorships’ endurance in time, the impact a successful co-authoring community can have over someone’s career, etc. The same methodology can be applied on a different dataset, e.g. on authors from a single field, and produce a different number of clusters, among which it is expected to find the “rising stars” cluster. It can also be employed to find groups of authors with similar productivity, impact and sociability profiles and label them accordingly.

During this process new metrics characterizing an author evolved, introducing time penalization in bibliometrical, *Power Graph* and social network analysis, to capture the temporal character of a collaboration’s success. These features were deployed into building annual datasets and then capturing the evolution of each author’s feature, with certain indicators. These indicators comprised the features based on which we applied the K-means clustering algorithm and grouped the authors. The number of the clusters was defined by experimenting and using well established clustering validity measures. It is on our intention to apply with more advanced clustering algorithms (density-based, connectivity based, probabilistic) that incorporate cluster coherence metrics and automatically find the best number of clusters. In this way, we will be able to overcome the limitations of K-means (takes the number of cluster as input, produces spherical clusters etc.). Cluster labeling was conducted depending on the feature characteristics of each cluster.

Furthermore, an attempt to expose the most influential features to the clustering, was executed, by applying singular value decomposition. The results showed that the number of papers penalized by time is the most influential feature for an author’s career.

Our plans for the future work focus on constructing a classification mechanism, which can classify an author to a respective group, given the required features, i.e., embedding to the approach the learning of a classifier based on the resulting clusters which will be seen as classes/categories. The analysis performed in the selected dataset of Greek affiliated researchers verified that there are differences between domains such as natural, formal, social or applied sciences. Since the proposed methodology can be applied on any set of authors, it is part of our next plans to focus on the researchers of a specific domain.

Moreover, it is important to work with the whole crawled dataset and experiment with diverse clustering approaches. To this end we aim to employ a big data framework, like Apache Spark (Zaharia, Chowdhury, Franklin, Shenker, & Stoica, 2010), to endeavor initially the use of all years’ graphs and secondarily an examination of the graphs’ spectrum behavior in time and a spectral clustering approach, which was rendered impossible due to the graphs’ volume and the eigendecomposition complexity.

Author contributions

Conceived and designed the analysis: George Panagopoulos, George Tsatsaronis and Iraklis Varlamis
Collected the data: George Panagopoulos
Contributed data or analysis tools: George Tsatsaronis and Iraklis Varlamis
Performed the analysis: George Panagopoulos
Wrote the paper: George Panagopoulos, George Tsatsaronis and Iraklis Varlamis
Other contribution: George Tsatsaronis

Appendix A. Examples of rising-star authors

Table A.6
The evolution of 4 rising stars and an already well-established researcher (as detected from the analysis of the 1998–2006 period).

Author	Foudas Costas	Maltezos Stavros	Sampsonidis Dimitrios	Vardas Panos	Chrousos George
University	University of Ioannina	NTU Athens	Aristotle University of Thessaloniki	University of Crete	University of Athens
Cluster	5	5	5	5	1
Discipline	Physics	Physics	Physics	Medicine	Medicine
Citations until 1998	2032	491	27	176	8432
Citations until 2006	6205	1514	1246	2552	33,309
Citations until 2012	11,691	6671	5319	12,231	58,408
Citations until 2016	25,195	20,329	18,079	27,038	74,198
Publications until 1998	95	22	26	49	479
Publications until 2006	174	128	92	205	812
Publications until 2012	412	366	286	446	1046
Publications until 2016	693	695	635	563	1224
H-index at 2016	73	59	57	59	132

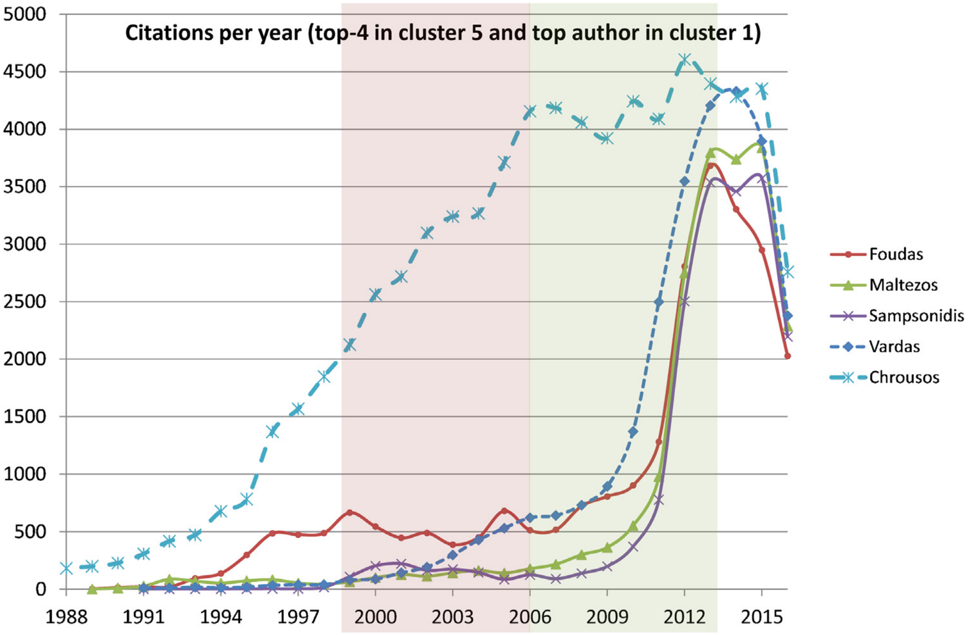


Fig. A.10. The top-4 authors of cluster 5 (rising-stars) and the top author of cluster 1 (well-established researcher). The rise in number of citations for the authors in cluster 5 is during the evaluation period.

References

Allison, P. D., Long, J. S., & Krauze, T. K. (1982). Cumulative advantage and inequality in science. *American Sociological Review*, 615–625.

Andersen, J. P. (2013). *Conceptualising research quality in medicine for evaluative bibliometrics*. Det Informationsvidenskabelige AkademiDet Informationsvidenskabelige Akademi.

Andersen, J. P., Bøgsted, M., Dybkær, K., Mellqvist, U.-H., Morgan, G. J., Goldschmidt, H., et al. (2015). Global myeloma research clusters, output, and citations: A bibliometric mapping and clustering analysis. *PLoS ONE*, *10*(1), e0116966.

Antelman, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries*, *65*(5), 372–382.

Balakrishnan, R., & Ranganathan, K. (2012). *A textbook of graph theory*. Springer Science & Business Media.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*(1), 179–189.

Bergstrom, C. (2007). Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, *68*(5), 314–316.

Bini, D. A., Del Corso, G. M., & Romani, F. (2010). A combined approach for evaluating papers, authors and scientific journals. *Journal of Computational and Applied Mathematics*, *234*(11), 3104–3121.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*(1), 93–102.

Buller, J. (2013). *Best practices in faculty evaluation: A practical guide for academic leaders*. Wiley.

Campbell, P. (2008). Escape from the impact factor. *Ethics in Science and Environmental Politics*, *8*(1), 5–6.

Chakraborty, T., Patranabis, S., Goyal, P., & Mukherjee, A. (2015). On the formation of circles in co-authorship networks. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining, KDD '15* (pp. 109–118). New York, NY, USA: ACM. http://dx.doi.org/10.1145/2783258.2783292. ISBN 978-1-4503-3664-2

Cook, D. J., & Holder, L. B. (2006). *Mining graph data*. John Wiley & Sons.

Costas, R., Van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, *61*(8), 1564–1581.

Cronin, B. (1984). . *The citation process. The role and significance of citations in scientific communication* (Vol. 1) London: Taylor Graham.

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. In *Database systems for advanced applications*. pp. 13–24. Springer.

Daud, A., Ahmad, M., Malik, M., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, *102*(2), 1687–1711.

Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *Journal of Mathematical Sociology*, *23*(3), 181–201.

Fiala, D., Šubelj, L., Žitnik, S., & Bajec, M. (2015). Do pagerank-based author rankings outperform simple citation counts? *Journal of Informetrics*, *9*(2), 334–348.

Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics*, *13*(1), 1–38.

Fu, T. Z., Song, Q., & Chiu, D. M. (2014). The academic social network. *Scientometrics*, *101*(1), 203–239.

Glänzel, W., & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. *Scientometrics*, *59*(3), 281–310.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, *14*(5), 403–420.

Harzing, A.-W., & van der Wal, R. (2008). Google scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*, 61–73.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation*, rvt008.

Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). Searching for rising stars in bibliography networks. In *Database systems for advanced applications*. pp. 288–292. Springer.

Long, P., Lee, T. K., & Jaffar, J. (1999). *Benchmarking research performance in Department of Computer Science, School of Computing, National University of Singapore*. http://www.comp.nus.edu.sg/tankl/bench.html

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2014). Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies. *Zeitschrift für Erziehungswissenschaft*, *17*(6), 111–132.

Odda, T. (1979). On properties of a well-known graph or what is your Ramsey number? *Annals of the New York Academy of Sciences*, *328*(1), 166–172.

Retzer, V., & Jurasinski, G. (2009). Towards objectivity in research evaluation using bibliometric indicators—A protocol for incorporating complexity. *Basic and Applied Ecology*, *10*(5), 393–400.

Royer, L. (2010). *Unraveling the structure and assessing the quality of protein interaction networks with analysis (Ph.D. thesis)*. Dresden, Germany: Technical University of Dresden.

Royer, L., Reimann, M., Andreopoulos, B., & Schroeder, M. (2008). Unraveling protein networks with power graph analysis. *PLoS Computational Biology*, *4*(7), e1000108.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497.

Tagarelli, A., & Interdonato, R. (2015). Time-aware analysis and ranking of lurkers in social networks. *Social Network Analysis and Mining*, *5*(1), 1–23.

Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Nørvåg, K., Schroeder, M., et al. (2011). How to become a group leader? Or modeling author types based on graph mining. In *Research and advanced technology for digital libraries*. pp. 15–26. Springer.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*. pp. 91–109. Springer.

Waltman, L., & Eck, N. J. V. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406–415.

Wendl, M. C. (2007). H-index: However ranked, citations need context. *Nature*, *449*(7161), 403.

Wolfgang, G., Bart, T., & Balázs, S. (2004). A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, *59*(1), 63–77.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (Vol. 10)* (p. 10).