This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

A Data-driven Unified Framework for Predicting Citation Dynamics

Antonia Gogoglou and Yannis Manolopoulos

Abstract—With the rising interest in predicting the scientific output, various efforts have been made to predict a scientist's h-index or the citation trajectory of a publication. In this work, we employ a dynamic categorization for scientists to ensure at each stage of their careers a comparison amongst their peers and combine this grouping with predictive models to estimate a scientist's future impact, as expressed by citation counts. Moreover, we investigate a wide range of factors identifying their importance in determining the future of science for different performance and academic levels with particular emphasis on features describing a scholar's position in multi-layered collaboration and citation networks. The robustness of the approach is examined on a longitudinal dataset centered around 700,302 data points representing Computer Scientists in various time periods with their complete networks of over 18 million collaboration links and 36 million citations. Our results indicate up to 30% improvement in prediction performance compared to baseline methods along with an average R^2=0.96 for short term and R^2=0.91 for long term predictions.

1 Introduction

In recent years the recording of research results and the communication amongst the research world has led to an abundance of bibliometric data and a growing interest in their use to quantify scientific impact, while extracting meaningful conclusions for decision making. The complexity of this quantification has become so high and the data so big in velocity, variety and volume that it exceeds individual understanding and judgment [1]. Therefore, the need for large scale analysis and automated decision support has risen in regards to scientific output and its evaluation. The results of such an analysis, in terms of measuring and comparing, are crucial, as scientific impact can shape the future of scientists' careers, institutions' reputation and publishing venues' impact. However, the evolution of scientific impact is a complex multi-criteria dynamic process with various factors at play that are either related with the scientists themselves or with their surroundings. Providing estimation on future research impact can potentially ensure competitive intelligence for strategy development and decision making. But how is future impact measured?

According to the seminal work of Eugene Garfield [2], citation analysis serves as the most fundamental proxy for the quantification of scientific output. On the other hand, until now citations can only measure current and past impact of a scientist, whereas in most decision making scenarios it is desirable to identify scientific potential and foresee its evolution. In addition, citation counts follow a heavy-tailed distribution in large datasets of scientists, thus increasing the difficulty in creating a unified but fair framework for estimating future citation counts for all scientists.

As the scientific community is a varied ecosystem of daunting size and rising complexity, the standards that constitute success may deviate significantly. In certain domains scientists tend to publish only a small number of nonetheless seminal articles, while in others scientists publish more frequently due to the fast paced nature of their fields. Differences in future career trajectories arise also due to academic age, i.e. mature scientists with a group of associated researchers publish more high impact publications as compared to new upcoming scientists. Depending on the task at hand, predictions for various time windows are also needed. As a result, the "one to fit all" approach in predictive modeling for citations will not yield the desirable results and will fail to uncover the factors that shape the evolution of different performance levels over time. Selecting a set of relevant factors can potentially favor certain groups or lack the adaptability for different use cases. For instance, the factors determining short-term future citation count may significantly deviate from respective ones for long-term impact. It has been stated that the correlation between citation counts in early years and future ones may be relatively low, since publishing patterns and performance change over time [3]. Also, Petersen et al. discovered based on empirical data that for younger scientists their reputation and the networks to which they belong heavily influence their future impact, as opposed to established academics [4].

According to Hirsch [5]: "A scientist has index h, if h of his papers have at least h citations each and the other papers have no more than h citations each". However, h-index describes only a fraction of a scientist's portfolio, i.e. the publications that have at least h citations each (known as h-core publications). Predicting a scientist's future h-index sheds light only on a subset of one's research activity, compared to the bigger picture provided by citation counts. Citations present a more skewed distribution within a wide range of values compared to the h-index. The h-index can also be heavily dependent on the citation accumulation of few relatively old publications [6], therefore its predictive power may be limited.

In the present work we focus on citations as a "raw" and versatile indicator of scientific performance and attempt to tackle the following questions:

Antonia Gogoglou is with the Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece, Yannis Manolopoulos is with the Faculty of Pure and Applied Sciences, Open University of Cyprus, 2220 Latsia, Cyprus, e-mail: agogoglou@csd.auth.gr, yannis.manolopoulos@ouc.ac.cy

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

- *how can comparable groups of scientists be created and how do they evolve over time?*
- *what factors are at play in shaping future scientific impact and how can they be combined in a predictive data-driven unified framework?*
- *are different categories of scientists affected by the same factors regarding their evolution?*

To address the above issues, we introduce a two-phase framework to dynamically create groups of academic peers and predict future impact for scientists belonging to each group. To our knowledge, this is the only work using a non-proprietary longitudinal dataset to extract meaningful features for the characterization of science evolution from an in-depth multi-layered graph analysis and create a prediction scheme for scholars instead of publications. Building upon our previous work [7], we employ impact based clustering over academic age and time period, while evaluating the resulting grouping and its correlation with citation network segmentation (k-core decomposition). Peer grouping is further combined with predictive modeling and the produced framework's robustness is tested across various performance and age levels. Through our experimental process and sensitivity analysis, meaningful conclusions are extracted on the mechanisms with which different scientists evolve based on their impact level instead of assuming that all scientists are affected in the same way by research related factors. The novelty of our approach is twofold: firstly, the achieved performance of the proposed framework compared to existing models is high due to the peer-grouping phase, the representative yet adaptable combination of features that can characterize different scholarly levels effectively. Secondly, the graph mining approach at author level aggregates information to identify how a scientist's relative position in collaboration and citation networks affects their impact evolution.

2 Related Work

The "science of science" has recently attracted significant level of attention and rigorous efforts have been made to extract meaningful actionable information from the abundance of bibliographic data and produce rankings [8], aid decision making as well as peer review [9]. This is evident by the plethora of bibliometric indicators that have been proposed in the past decade. All these indicators attempt to quantify different aspects of scientific impact [10] and efforts have been made to allocate credit in a fair and representative manner, e.g. in multiple-author publications where impact is distributed amongst collaborators [11], [12]. The focus is now turning towards the quantification of future impact and rising influence, instead of measuring existing output in different ways. In his preliminary work, Price deduced that current visibility, publishing venue and age highly influence a publication's future outreach [13]. To this day, all stakeholders in the scientific community heavily depend on peer review, which is a pragmatic but often controversial way of assessing impact. Consequently, an extensive interest in using data intelligence to assist peer review becomes evident to overcome personalized criteria and produce valuable evidence-based insights.

Hitherto, existing approaches have focused on the prediction of future h-index values at author level [14], [15] or citation counts at publication level [16], [17]. Another categorization of existing approaches occurs with regards to their modeling methodology. A set of them treats the prediction of future impact as a *classification* problem [18]–[20], where a set of predefined categories has been constructed to characterize the current state of a scientific entity and measure the changes at a given time interval. The future behavior of a new scientific entity is approximated by that of the entire category in which it is placed. Additionally, classification based approaches use restrictive thresholds to define impact (e.g. a publication reaches the top 10% citations in a journal [18]) limiting their universality and usability. However, an entity's future state may significantly deviate from the group (e.g. sleeping beauties [21]). To rectify this shortcoming, *regression* based approaches have been introduced with the seminal work by Acuna et al. [14] and others [22]–[24]. This methodology has been criticized since predictability depends on the aggregation of career data across multiple age cohorts, leading to models unfair towards younger researchers [25].

Inspired by the evolution of social networks and Web modeling, various parameterized distribution models have also been introduced (*statistical modeling approaches*) [26], [27]. It is argued though that it constitutes an oversimplification to characterize the complex process of citation dynamics using a distribution model alone, even with a plethora of parameters [28], while the challenge is magnified at author level. In these approaches along with the spatio-temporal ones [29], [30], where citation accumulation is modeled as a time-series problem, there exists a heavy dataset bias. The need for extensive past data regarding a particular scientist or publication discourages efforts for early impact prediction of newcomers. In an analogous manner, one can consider citation acquisition as a network of connected nodes, ergo determining the future state of this network constitutes a *link prediction* problem [24], [31]. Due to the challenges that arise when estimating highly skewed quantities, such as citation counts, various researchers have altered the impact definition and proposed heuristics to approximate it. Datta et al. [32] provide a large scale empirical study of the factors that promote scientific influence, whereas [33] provides alternative metrics for current and future impact.

2.1 Challenges

As in various social networks (social media, World Wide Web, etc.), an actor's future impact can be determined by a wide range of factors, such as current reputation, collaborators' reputation, level of influence in one's neighborhood, etc. The complex interplay of these diverse factors is less than obvious across a large set of scientists with different behaviors and can confound attempts to produce accurate and meaningful predictions. To include the above complex network features, the computational cost increases drastically with the size of the network (i.e. number of scientists). For instance, if we consider two publications with m and n authors respectively, a citation link between the two publications creates $m \times n$ links at author level. Should one

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

consider how the status of linked scientists reflects upon their impact, the depth of the created graphs increases to recursively include the next level of the graphs, i.e. the collaborators and citers of the linked scientists, increasing complexity in the case of large scale analytics. When the analysis is spread along the time axis, different *academic ages* arise and the mechanisms of citation accumulation alter. Preferential attachment, the main mechanism of citation distribution [34], largely depends on the age of an actor. This, in particular, holds for fast growing networks where the attachment probability of new incoming entities varies based on the duration of one's existence in the network. For instance, older publications in a citation network are cited more often, whereas, on the other hand, younger scientists may change their position in the network more rapidly, thus influencing their future citation rates (i.e. citations per publication).

3 Dataset

For the purposes of our analysis, we used a real world dataset collected from Microsoft Academic Search (MAS)[1] containing 30,000 scientists (from here on referred as *core* scientists) publishing in various domains of Computer Science (CS). The complete publication and citation records for each core scientist were collected with the oldest publication in our set dating back to 1950. Our analysis takes place in timestamps within a time period of 34 years (1980-2013), where the data in MAS for the CS field are the densest. To ensure the temporal evaluation of our records, we also accessed DBLP[2] by using the XML search DBLP API and obtained the missing publication year for 6% of our publication records. Each scientist, with her/his portfolio and impact calculated in every year of her/his career, constitutes a "data point", leading to a total of more than 700,000 points. Moreover, the complete information regarding scientists that cite the *core* set (citation network), co-authors of the *core* (collaboration network) and their respective portfolios and associated citations were retrieved for every year in the selected time period. This way we can study the network topology of the resulting graphs and extract meaningful features for our prediction framework. As a result, a total of 5,001,130 authors were contemplated in our analysis. Regarding the initial sample or *core*, it includes scientists of the CS field from 25 domains (different publishing patterns), various performance levels ($1 < h$-index < 100) and productivity ($1 <$ portfolio size < 800). Finally, to ensure the completeness of records, extensive author name disambiguation was performed to merge multiple name variations of the same person. It is mentioned that existing large scale data sets are only broad but not deep in size, often containing a massive number of author names; however, the majority of them are "one-timers" in the dataset, meaning their entire portfolios and surrounding networks have not been retrieved. To our knowledge, this dataset is the only non-proprietary set of scientists covering in depth multi-layered graphs of bibliographic data over an extended time period around a *core* of scientists representing multiple performance levels, publishing patterns and age groups. Table 1 includes the statistics of our dataset[3].

TABLE 1
Statistics of the used dataset

Core scientists	30,000
Data points	700,302
Citing authors	3,391,288
Collaborators	1,609,842
Total authors	5,001,130
Authored publications (for core scientists)	2,260,397
Citing publications (for core scientists)	4,441,800
Total publications	6,672,197
Collaboration links (for core scientists)	1,343,817
Citation links (at author level)	36,101,404
Domains	25

4 Two-phase Unified Framework

In this section we will describe our approach for a two-phase unified framework to determine the mechanism with which scientific impact evolves over time and obtain an accurate prediction of a scientist's future status.

Problem Definition: Given a scientist s who at timestamp t has a portfolio p and c citations, s is assigned to an impact cluster (cl) with scientists of similar age and performance level based on a set of features $\mathcal{F}1$. The graphs representing the collaborators of s (G_{CoA}) and scientists citing s (G_c) before or at timestamp t are formulated. A set of features $\mathcal{F}2$ describing scientist s at timestamp t are calculated from the aggregated graphs and fed into a cluster-specific function φ_{cl} to predict citation count c of s at timestamp $t + \delta T$:

$$\{p, c, G_{CoA}, G_c\}_t \longrightarrow \mathcal{F}2_t \quad (1)$$

$$\varphi_{cl_t|\mathcal{F}1_t}(\mathcal{F}2_t, \delta T) \longrightarrow c_{t+\delta T} \quad (2)$$

In Figure 1 we provide an overview of the proposed framework to unify impact clustering over time with predictive models for future citation count estimation. The preprocessing stage of the framework for dividing bibliographic data into time intervals and scientists into age groups, along with the clustering phase to place scientists in cohorts corresponds to Phase 1 (Section 4.1). A prediction phase is added for acquiring k different predictive models equal to the number of clusters, so that fine-tuned models can be tailored to the different patterns that arise in citation acquisition. Additionally, we will explore how peer grouping enhances the performance of the predictive models, test the robustness of the produced unified framework and explore the factors that determine future impact for all levels and time periods (Section 4.2).

4.1 Phase 1: Impact based peer grouping

4.1.1 Defining clusters of peers

At any given timestamp t (e.g. a specific year) scientists of different academic age, performance level and across vari-

1. We appreciate Microsoft for providing gratis their database API. The version of the API used in this work has been discontinued by Microsoft since the summer of 2016.

2. http://dblp.uni-trier.de/

3. https://github.com/AntoniaGogoglou/Dataset-of-bibliographic-networks

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

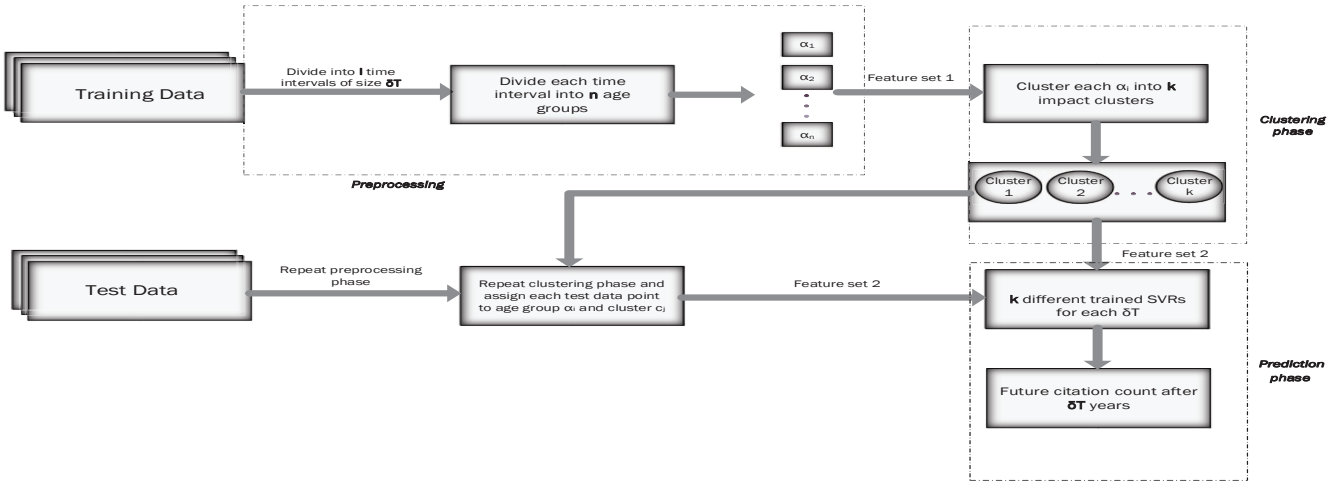

Fig. 1. The flow of our proposed data-driven unified framework for future impact prediction

ous domains need to be compared based on their impact. The question arises: *"how does the career of a scientist evolve in terms of her/his impact over time?"*. To achieve fair and meaningful comparisons, a systematic grouping methodology needs to be employed. Defining hard set threshold-based performance levels can be controversial due to different practices across fields and time-varying publishing behaviors [35]. To address this issue we deploy a dynamic clustering technique and a set of heuristics to group together authors with comparable impact. For each time interval of size δT (e.g. δT=5 years) scientists are divided in n subgroups according to their academic age α, i.e. the number of years since their first publication. For each subgroup clusters of scientists are created according to their impact. The idea behind this segmentation is to compare scientists that have existed in the scientific network for similar time.

Given the high correlation amongst bibliometric performance metrics and based on proposed categorizations of those metrics in [10], [36], we opted for three straightforward "raw" metrics to represent both impact and productivity (feature set $\mathcal{F}1$): *citation count* (c) to represent cumulative impact, h-index (h) for ranked output, and *citation rate* (c/p, where p is the number of publications) to account for impact normalized over productivity. Figure 2 depicts the coefficient of variation (βCV) [37] for various contemplated combinations of features and number of clusters k. When βCV curve stabilizes, it is expected that the variabilities in the intra and inter-cluster distances remain stable, implying that adding more clusters should be of little help to understand the dataset variability. We notice that the best performing combination is dividing scientists into four clusters based on c, h-index and c/p. Opting for four clusters provides enough separation in a dataset with varying publishing patterns. For denser datasets with scholars of similar age and performance, a tighter segmentation with less clusters could also be employed. Our choice for the number of clusters, though, constitutes a fair guideline for highly diverse scholarly groups, irrespective of dataset size.

To facilitate a common basis for comparisons across time intervals and age groups, the number of clusters needs to be predefined instead of dynamically calculated. The

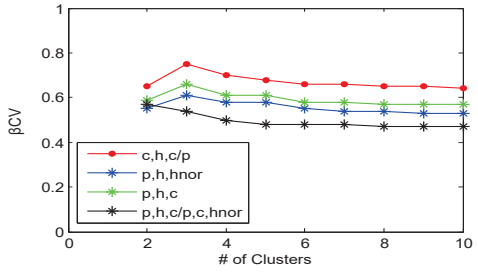

Fig. 2. Coefficient of variation (βCV) for different feature combinations and number of clusters

clusters need also be automatically ranked to represent impact levels. Therefore, an adaptation of the Self-Organizing Map (SOM) [38] was used with a 2×2 topology. Essentially, the resulting segmentation of SOM is a projection of the three dimensions/features on a $2D$ space, where clusters represented by nodes (centroids) that are topologically close are more similar compared to the ones placed further away. In other words, the topology of the input space is preserved and an interpretable grouping is provided [39]. In this sense, items placed in cluster 1 are more similar to items placed in cluster 2, as compared to the ones of cluster 4. For instance, one cluster can be comprised of scientists with high values in all three features, whereas another one places together the ones with high h and c/p values, but lower citation count c. Therefore, we propose a heuristic based on the summation of feature values for each cluster to match the SOM mapping to meaningful impact levels.

As indicated in the pseudocode of Figure 4.1.1, given a set of scientists S and a time period T, scientists are assigned to an age group α_i based on their academic age α_s (lines 3-11). An updated citation count, h-index and citation rate are calculated for each scientist in every α_i including only publications published before time t (lines 12-15). In the next step, we create a mapping of the scientists belonging to each age group α_i into k=4 clusters represented by the array $Clustered$ (line 17). The clusters are then ranked based on a weighted combination of maximum, minimum and

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

Require: A set of scientists S with full publication and citation records in a period of time T
Ensure: An array Clustered[clusterId, list of scientists] of scientists grouped in $k = 4$ clusters
1: Divide T in timestamps with the interval between two consecutive timestamps defined as δT
2: Define n age groups $[\alpha_1, \alpha_2, ..., \alpha_n]$ where each α_i covers a range of years (in our case 5 years)
3: **for** each timestamp t **do**
4: **for** each scientist s **do**
5: $s.p$:= set of publications of s
6: p_t := array of zeros
7: **for** pub in $s.p$ **do**
8: **if** $pub.year < t.end \,\&\, pub.year > t.begin$ **then**
9: add pub to array p_t
10: Academic age of scientist s at timestamp t is $\alpha_s := t.end - min(p_t.year)$
11: Assign scientist with α_s to appropriate age group α_{it} for the given timestamp t
12: c_t := calculate citation count for p_t
13: h_t := calculate h-index from c_t and p_t
14: $(c/p)_t$:= calculate citation rate as sum(c_t)/size(p_t)
15: Add scientist's s tuple of features $\{c_t, h_t, (c/p)_t\}$ to his age group α_{it}
16: Standardize values of features $\{c_t, h_t, (c/p)_t\}$ over all age groups of timestamp t
17: **for** each age group α_{it} at timestamp t **do**
18: Clustered_α_{it}[clusterId, list of standardized $\{c_t, h_t, (c/p)_t\}$] := SOM(α_{it}, k)
19: **for** each clusterId cl **do**
20: sumMax[cl] := max(c_t) + max(h_t) + max($(c/p)_t$) where $clusterId = cl$
21: sumMin[cl] := min(c_t) + min(h_t) + min($(c/p)_t$) where $clusterId = cl$
22: sumMean[cl] := mean(c_t) + mean(h_t) + mean($(c/p)_t$) where $clusterId = cl$
23: weightedScore[cl] := $\frac{1}{3}sumMax[cl] + \frac{1}{3}sumMin[cl] + \frac{1}{3}sumMean[cl]$
24: Rank clusterIds based on their weightedScore
25: Update Clustered_α_{it} with ranked clusterIds
26: Clustered.add(Clustered_α_{it})
27: Return Clustered[clusterId, list of scientists] at the end of timestamps

Fig. 3. Pseudocode for temporal impact-based peer grouping of scientists over time (phase 1)

average values of their members across all three clustering features (lines 18-23). Consequently, the high impact cluster is defined as the one with the biggest score in the weighted sum, whereas the low impact cluster obtains the lowest score. After ranking the clusters (lines 23-25), the k-th cluster represents the highest impact, the $(k-1)$-th one represents the moderate to high impact and so on.

4.1.2 Impact Grouping Evaluation and Meaning

Next, we are evaluating the proposed grouping by comparing the distribution of features amongst clusters and correlating the cluster segmentation to the degeneracy of the citation network. The clustering methodology is applied on age groups of scientists accounting for 0-50 years of activity with range 5 (0-5, 5-10, etc.) over time intervals of size δT=5 years. For years 1988, 1993, 1998, 2003 and 2008 a set of 102,735 data points (scientists in their various career stages) are included, from which 50,000 items were sampled preserving the relative percentages of each cluster's size to the total number of items. Table 2 displays the size of each age group in every timestamp (i.e. year) comprising our sample. This sample is used to estimate the probability density (PDF) and cumulative distribution function (CDF) depicted in Figure 4. It is observed that cluster 1 displays the steepest exponential decay across all metrics. The slope decreases as we move towards clusters of higher impact, with cluster 4 obtaining a higher probability (integral) of acquiring high values in all three features, which is compliant with our cluster ranking.

TABLE 2
Size of each group in each contemplated time interval (year) as represented in the 50,000 sample to calculate PDF and CDF functions

Age \ Year	1988	1993	1998	2003	2008
0-5	1,612	3,461	3,446	2,787	775
5-10	909	1,708	3,525	3,523	2,831
10-15	566	915	1,747	3,532	3,474
15-20	363	598	896	1,701	3,577
20-25	155	332	595	918	1,684
25-30	80	159	339	626	904
30-35	17	74	158	370	608
35-40	10	23	85	163	370
40-45	6	6	18	70	168
45-50	1	2	10	22	81

Even though our analysis so far highlights the correlation of the proposed grouping with the impact and different publishing patterns, it neither unfolds the significance of the scientists in each cluster nor their influence level to the scientific network structure. Therefore, we perform a k-core decomposition [40] of the citation graph at author level by degenerating the network of different timestamps into k-shells. For each timestamp, we collected the set of scientists having published before the end of that timestamp and

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

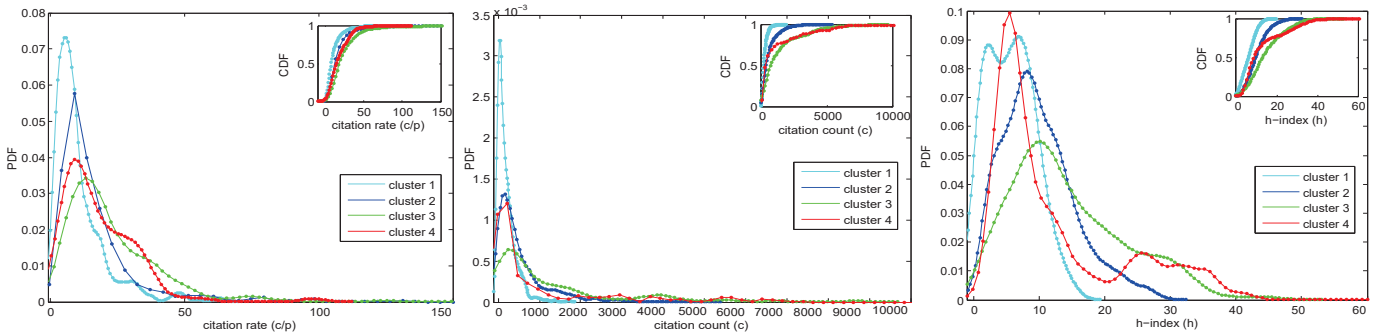

Fig. 4. Probability density for citation rate, citation counts and h-index values per cluster aggregated across timestamps: 1988,1993,1998,2003,2008

retrieve all the citations collected in the same period. One link is created between each author of the citing publication and each author of the cited one. Given a graph G, the k-core (denoted as G_k) is the maximal non-empty subgraph of G, where all vertices have in-degree larger than or equal to k. It is mentioned that k-cores are nested in the sense that $G_{k+1} \subseteq G_k$. The k-shell of a graph is the subgraph of G induced by the edges contained in the k-th core but not in (k+1)-th core. In the case of the citation graph, when a scientist is assigned a k-core number, this represents the largest integer k for this person to exist in a citation graph, where all scientists have received at least k citations. The maximum core, k_{max}-core, is the highest core number of all vertices.

To allow for long term influence shifts to be displayed, we contemplate four different timestamps (years) with interval $\delta T = 10$ (Y_1=1983, Y_2=1993, Y_3=2003 and Y_4=2013) in our decomposition with their aggregated citation networks including citations received by core scientists publishing before Y_i. We start the decomposition by recursively removing all scientists that have received only one citation and assign them to k=1-shell. Subsequently, we remove scientists with $2, 3, \ldots, k_{max}$ connections until all scientists have been assigned to a shell. The nodes belonging to the k_{max} comprise a well-connected globally distributed subset of the network. As we move towards smaller k values, we reach the *periphery* of the network. Through this segmentation we observe how each cluster's attachment to a particular shell evolves over time and how each of them contributes to the central part of the citation network.

Due to the varying number of k-shells over the different years, in Figure 5 we display only the five most highly populated k-shells and the k_{max}-core for selected years. The innermost circle represents the k_{max}-core, whereas the rest concentric circles represent the other five selected k-shells moving from central to peripheral ones (outermost circle). The four quadrants represent the four impact clusters in gray scale: quadrant I is cluster 1, II is cluster 2 and so on. Figure 5 also depicts the proportion of each cluster belonging to the six selected k-shells normalized to each cluster size. In other words, it illustrates how each cluster is dispersed along these k-shells. The high impact clusters dominate the k_{max}-core, while their proportion decreases as we move towards the network periphery. On the other hand, the lower impact clusters increase their proportional size in the peripheral shells with cluster 1 being almost absent from the k_{max}-core in years 1983 and 1993.

It is interesting that in more recent years (2003 and 2013), when the citation network has increased both in size and density, the k_{max}-core becomes more heavily populated, as compared to previous years. This complies with the evolution of a social network [41]; in earlier years the citation network is rather sparse and only high impact scientists are well-connected, ergo influential. With the network expanding over the years, scientists of all levels become interconnected increasing their outreach. As happens with the Web graph or Online Social Network (OSN) graphs [42], a large real world network with social interactions displays a "gravity" force over time that pulls nodes to the central part leaving the periphery sparser and populated with disconnected (low-impact or inactive) nodes. Even in the recent years though, high impact clusters dominate the core and their distribution over the periphery reduces at a higher rate compared to clusters 1 and 2.

4.2 Phase 2: Predicting scientific impact

4.2.1 Feature Selection

Once a scientist at timestamp t has been assigned to a cluster, a set of features ($\mathcal{F}2$) are calculated as possible influential factors of the her/his future impact. This set consists of six categories of features that address diverse aspects of research activity, as some have been utilized in literature [43]–[47], but employed here over an aggregating combination of networks formed around scholars. They are expressed by the subsequent metrics:

Productivity: A high productivity may lead to increased outreach of a scientist's work and higher probability of producing a seminal high impact publication [1].

- *Number of publications (p)*: the total number of articles a scientist has authored.

Impact: Bearing in mind the rich-getting-richer phenomenon in science and the exponential growth in citation counts, a scientist with elevated current impact is more likely to accumulate further citations in the future.

- *Citation count (c)*: the total number of so far accumulated citations for a scientist.

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

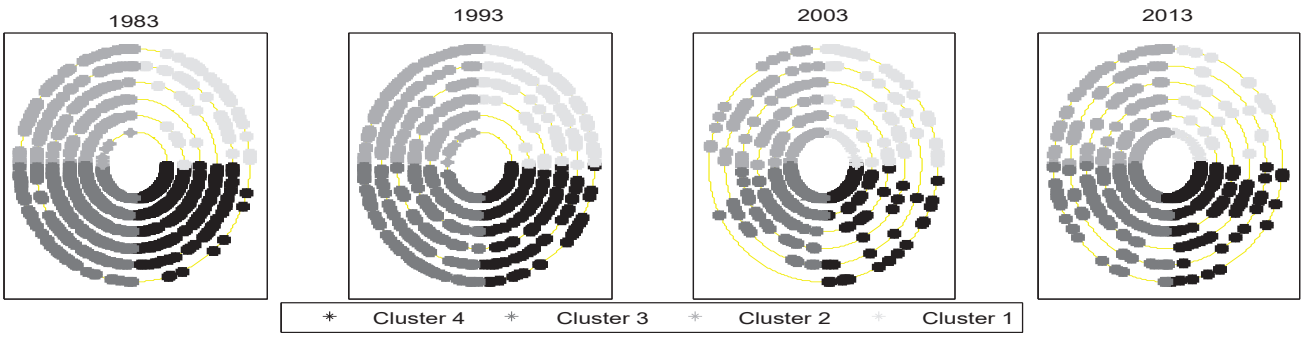

Fig. 5. Gray scale mapping of the composition of different k-shells with respect to four impact clusters for timestamps 1983,1993,2003 and 2013. The innermost circle is the core, while the outermost is the periphery with the four clusters depicted in the respective four quadrants

- *h-index (h)*: which attempts to measure both productivity and impact, but focuses on the number of publications that managed to achieve a threshold of citations.
- *e-index (e)*: which takes into account the *excess* citations [48], i.e. the higher than h citations acquired by a scientist's publications. It constitutes a measure of excellence, since the maximum number of citations achieved in a scientist's portfolio, i.e. one's seminal publication, is often underrated in the calculation of the h-index.

Co-authorship: Science is expanded through collaboration and the effects of co-authorship relationships on scientific impact have been extensively studied [49]. Identifying scientific collaborators helps map the evolution of research communities. We choose two different metrics to represent the level of collaboration in a scientist's career as well as the status of her/his co-authors that reflects not only on the research one shares with them but also on one's entire portfolio.

- *Number of single author papers ($singAuth$)*: indicating how heavily one depends on her/his collaborations and whether or not collaboration is beneficial for a scientist's outreach.
- *Average h-index of co-authors ($meanHCoA$)*: which expresses the shared status from one's co-authors.

Venue features: Similar to a scientist's reputation, publishing venues also have a social status that influences the attention their publications attract. There is a common belief in the scientific community that publishing in certain journals raises the probability of getting cited.

- *Journal Impact Factor (JIF)*: which is the most common proxy for journal reputation, measuring the yearly average citations to recent articles published in that journal. The JIF reflects the number of citations to the number of publications in the last two years; in other words, it is a measure of relative importance. Therefore, in the context of this work, we calculate it based on the citation and publication records in our data instead of relying on an external ranking (e.g. Journal Citation Report, Scientific Journal Rankings, etc.). This way a representative JIF can be calculated even for past years (e.g. 1980) based solely on data available at that time.

Social features: In science, "standing on the shoulders of giants" is a common way to produce and disseminate research. As a result, the position a scientist holds amongst others with whom s/he is connected through citation links, represents her/his level of influence. The citation network in our framework is created at author level, indicating that for every publication p_i that cites another p_j, all authors of p_j are connected with an incoming link from the authors of p_i. A scientist with a central and focal position in the citation network not only gets cited frequently but also helps in connecting groups of researchers and topics.

- *PageRank (PR)*: the well-known score representing the probability that a scientist gets cited from a random newly inserted publication.
- *Core number (cn)*: which represents the largest integer k for a node such that this node exists in a graph of degree $> k$ (see Section 4.1).
- *Betweenness centrality (bc)*: which represents the fraction of the shortest paths from any node to any other that pass through a specific node (i.e. a scientist). Scientists with high betweenness centrality values tend to act as "bridges" being located at the intersection of communities. This means that they bring together different cohesive research groups or they may link related publications.
- *Closeness centrality (cc)*: which is the mean geodesic path from one node to all others. In the case of the citation network, as it contains disconnected (less cited) scientists, we used the harmonic mean to calculate representative values for the closeness centrality [50]. In the citation network, scientists with a high cc value might have a more direct influence on others and disseminate their research more effectively due to their proximity to a large number of scientists.

Temporal features: Apart from the current state of a scientist, a pivotal indicator of her/his future citation trajectory is the past shift in her/his impact that may help unravel citation patterns and use them for future estimations.

- *Difference in citations (Dc)*: the increase in citations in the past δT years, assuming that one's career is longer than δT. Otherwise, this value equals zero.
- *Difference in h-index (Dh)*: similarly to Dc this metric represents the latest increase in h-index in the past δT years.

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

4.2.2 Data Collection Bias and Model Selection

Our data-driven approach is based on a "data-point" modeling of a scientist's current status compared to her/his future status after δT years. In other words, a single data point is comprised of a vector of features ($\mathcal{F}2$) that are calculated at timestamp t based on all the publication and citation records of our dataset till that time, and a future status which is represented as citation count at timestamp $t + \delta T$. Therefore, no information exposure can occur as a scientist's current state is estimated using only the records until the contemplated timestamp. For instance, the social features or the JIF are assessing citation networks or a venue's reputation respectively at year 2000 based on the status acquired no later than 2000.

Another common issue with temporally sensitive predictive models is the selected time window for training and testing. In our approach we follow a "data-point" model instead of a time series model. Since we are evaluating our models on real data, selecting for instance years 1980-2008 as training and years 2009-2013 for testing ($\delta T = 1, 2, 3, 4, 5$ respectively) would produce significant data bias related to the density and detail of citation records across different time periods or to the thoroughness of a given database over time. It is often the case that different databases provide more detailed records for certain years, while they are sparser for others depending on their sources and update rate. To alleviate these issues and given that our definition of a "data-point" ensures no future information regarding a scientist's status is exposed in earlier years, we employ the following sampling technique:

- Given a time interval of size $\delta T = \{1, 2, 3, 4, 5\}$, ($\mathcal{F}2 = \{f_1, f_2, \ldots, f_{13}\}$) is calculated characterizing every scientist publishing prior to each of the contemplated years $\{t_{initial}, t_{initial} + \delta T, t_{initial} + 2 * \delta T, \ldots, t_{final}\}$ with $t_{initial} \geq 1980$ and $t_{final} \leq 2013$ in our experiments.
- After constructing all our data points, we shuffle them and perform 100 Monte Carlo simulations with 70%-30% training and test set respectively out of the total of 700,302 instances for cross validation purposes [51]. This way we minimize the data bias as well as the variance in the performance evaluation of our framework.

The prediction phase of our proposed framework uses Support Vector Regression (SVR) [52]. An important advantage of Support Vectors (SVs) is that the weights can be completely described as a linear combination of the training patterns (SVs), meaning that the complexity of a predictive function's representation by SVs is independent of the dimensionality of the input space and depends only on the number of SVs. In our case, we use ϵ-SVR where the goal is to find a function $\phi(x)$ that has at most ϵ-deviation from the actually obtained targets for all training data and at the same time is as flat as possible. SVRs have also been identified as the most appropriate prediction model for citation measurements compared to linear regression or k-NN [45]. The implementation of SVR in LibSVM [53] is employed with default ϵ and Radial Basis as kernel function, due to the non-necessarily linear relationship between current and future state (impact can rise exponentially or steadily). Next, we will provide a comparison of our proposed method with other widely used approaches.

4.2.3 Performance Evaluation and Baseline Comparisons

For the evaluation of our proposed framework we aggregate all results from our Monte Carlo simulations, with different training and test sets randomly sampled in the proportion 70%-30%. Performance metrics are averaged over all simulations for each δT and impact cluster. Figure 6 displays two selected metrics, R^2 and relative Root Mean Square Error ($rRMSE$), for all clusters across all time intervals. The coefficient of determination R^2 is primarily used as a metric of relative correlation between the actual target variable and the predicted one based on a set of related features. It expresses the *goodness of fit*, i.e. the proportion of variability of one factor (data set) that can be caused by its relationship to another factor (predictive model). An R^2 equal to zero means that the target variable cannot be predicted using the independent variables (features). Conversely, if it equals one, this means that the target future state is always predicted by the proposed model.

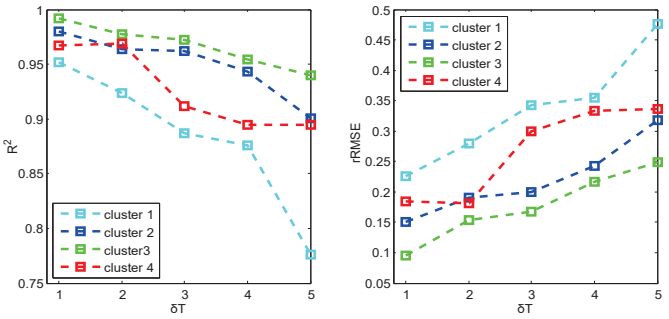

Fig. 6. Values of R^2 and relative $RMSE$ per impact cluster for δT=1,2,3,4,5

An additional metric used for performance evaluation is the relative Root Mean Square Error ($rRMSE$). Traditional $RMSE$ is the square root of the average of squared error between predicted and real values. In our case, with citation counts displaying a wide range of values with huge differences amongst impact clusters, comparisons using $RMSE$ would not be feasible. Therefore, we opt for a normalized scale-free version by dividing the $RMSE$ with the error occurring if the predicted values were to be naively calculated as the average of the actual values within each cluster, thus obtaining a relative $RMSE$ for each cluster. Equation 3 gives the definition of $rRMSE$:

$$rRMSE = \sqrt{\frac{\sum_{t=1}^{n}(f - y)^2}{\sum_{t=1}^{n}(y - \widehat{y})^2}} \tag{3}$$

where f denotes the predicted values, y denotes the actual values and $\widehat{y}$ is the average of the actual values for each cluster. Other considered metrics were $MAPE$ (Mean Absolute Percentage Error) and MAE (Mean Absolute Error) with MAE being scale-dependent and $MAPE$ being sensitive to possible outliers (highly cited scientists).

As seen in Figure 6, it is more difficult to predict the future status of the two "extreme" impact clusters, i.e. low impact cluster (cluster 1) and high impact (cluster 4),

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

compared to the ones of moderate impact. Additionally, a significant performance drop occurs for larger time intervals (δT=4 or 5) in the case of clusters 1 and 4, with lower R^2 values and higher $rRMSE$. These challenges can be attributed to the versatility of the low impact cohort of scientists. A number of them may slowly become inactive in the subsequent years or they may significantly increase their output achieving unexpected high impact long term (5 years ahead). The top impact cluster, in turn, is comprised of exceptional high performing scientists whose performance is also subject to bursts that are hard to predict. On the other hand, moderate impact clusters (clusters 2 and 3) include steadily performing scientists, thus allowing for better predictions to occur even in the long run.

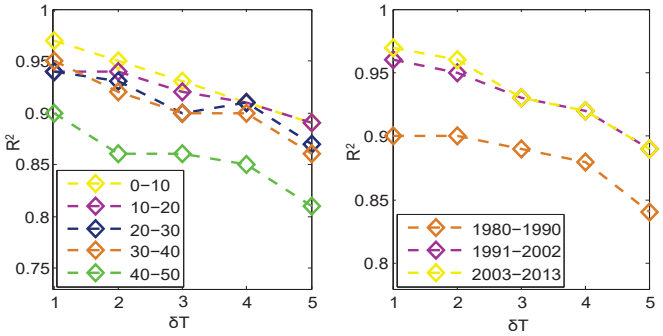

Fig. 7. Values of R^2 across different age groups (left) and different time periods (right) over all time intervals δT=1,2,3,4,5

The proposed framework's performance is also evaluated on various age groups of scientists and different time periods to test its robustness and effectiveness. As stated in Section 4.1, the first step of our framework introducing impact clustering across different values of academic ages, allows for comparisons amongst peers to occur. But how does this affect the prediction performance, since in the same cluster scientists of various academic ages can coexist sharing similar relative impact compared to their cohort? Furthermore, the rate with which research expands has grown significantly in recent years as opposed to earlier decades. Since our analysis expands over a period of 34 years, we need to explore the accuracy of predictions across various time periods when high performance can be interpreted differently and bibliometric data are of different scale. In the left part of Figure 7, values of R^2 are illustrated for all our contemplated age groups of Section 4.1 each one averaged out over all impact levels (clusters). Predictive performance is high for all age groups, with a slight drop in the group with 40-50 years of academic activity. This is a generally sparse age group populated mostly with retired or inactive scientists who keep accumulating citations for existing publications. Therefore, their future impact cannot be fitted to a particular pattern of active performance, e.g. they may have one seminal publication that still gathers citations but the rest of their portfolio is stagnant. On the right of Figure 7, R^2 values for three time periods are depicted. It is interesting that performance for all time periods matches the average values over all clusters (see Table 3), indicating that the exponentiality in citation accumulation is more prominent across different impact levels than across different time periods. A small deviation in performance occurs for early decades (1980-1990), which can be attributed to the fact that in these early years most scientists are new comers and they are often placed in cluster 1 that is the most challenging to predict.

To better investigate performance we implement as baseline a commonly used model to predict bibliometric quantities, namely the generalized linear regression with elastic net regularization as used by Acuna [14] and numerous others [6], [16], [22]. Generalized linear regression (GLM) and its simplest form, i.e. linear regression, have been extensively employed in the "science of science" [23], [33]. Moreover, we combine GLM with elastic net regularization to

TABLE 3
R^2 values per cluster for our proposed model and three baseline approaches

Time Interval	Clusters	SVR-Clust		ElGLM-Clust		SVR-NoClust	ElGLM-NoClust
δT=1	cl1	0.95	mean of all clusters 0.97	0.89	mean of all clusters 0.89	0.89	0.80
	cl2	0.98		0.86			
	cl3	0.99		0.91			
	cl4	0.97		0.90			
δT=2	cl1	0.92	mean of all clusters 0.96	0.88	mean of all clusters 0.89	0.84	0.78
	cl2	0.97		0.87			
	cl3	0.98		0.91			
	cl4	0.97		0.91			
δT=3	cl1	0.89	mean of all clusters 0.93	0.81	mean of all clusters 0.87	0.80	0.72
	cl2	0.96		0.90			
	cl3	0.97		0.91			
	cl4	0.92		0.86			
δT=4	cl1	0.88	mean of all clusters 0.92	0.83	mean of all clusters 0.88	0.80	0.70
	cl2	0.94		0.88			
	cl3	0.95		0.91			
	cl4	0.90		0.89			
δT=5	cl1	0.79	mean of all clusters 0.89	0.75	mean of all clusters 0.82	0.79	0.69
	cl2	0.91		0.84			
	cl3	0.95		0.86			
	cl4	0.91		0.84			

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

determine better fitting coefficients for the selected features and eliminate irrelevant or highly correlated ones. The three implemented baseline models are:

- *Elastic net GLM without impact clustering* (*ElGLM-NoClust*), where one linear model is fitted to the entire data set,
- *Elastic net GLM with impact clustering* (*ElGLM-Clust*), where one linear model per cluster is fitted, and
- *Support Vector Regression without clustering* (*SVR-NoClust*) to explore not only the selection of regression approach but the performance of our entire framework.

Table 3 displays the achieved performance for all models trained and tested over the same data and averaged out over 100 Monte Carlo simulations. We observe that short term predictions display a consistently higher performance as compared to long term ones and cluster 1 (low impact) is the most difficult to predict for all models. However, our model (SVR-Clust) for clusters 2, 3 and 4 achieves R^2 higher than 0.9 across all δTs.

Another interesting finding is how impact clustering improves performance for both SVR and ElGLM; there is a bigger difference, though, between SVR-NoClust and SVR-Clust and between ElGLM-Clust and ElGLM-NoClust, as opposed to ElGLM-Clust versus SVRClust. In other words, defining a scientist's cohort effectively is more important than model selection in estimating her/his future impact. In any case, SVR significantly outperforms ElGLM in both cluster and non-cluster versions, which can be attributed to the not necessarily linear relationships between our proposed features and the future citation count. Considering that ElGLM is the existing literature baseline, our combination of SVR with dynamic grouping achieves a 22% higher performance in short term predictions (δT=1 year) and a 30% improvement in long term ones (δT=5 years).

Next, we detail a series of plots for selected scientists from our final test set (100-th simulation), where their actual citation trajectory is compared with the predicted ones from our proposed model and the aforementioned baselines. These scientists were selected randomly with the condition that they are present in our test set for more than 60% of their entire career length (as covered in our dataset). The left column of Figure 8 contains random scientists, whereas the right contains selected ACM Fellows present in our dataset, so that performance can be illustrated for all impact levels. As expected, for δT=1 (first row of plots) all models are relatively close to the actual citation curve, but in larger δT the SVR-Clust evidently outperforms all other approaches. As for different impact levels, we observe that for ACM Fellows the baselines increase their performance slightly compared to random scientists, but are still outperformed by our framework. This may be due to the fact that, as discussed in Section 1, the "one to fit all" approaches usually favor more mature scientists over younger or moderately performing ones. Furthermore, we notice that ElGLM-NoClust, which is more commonly used in literature, tends to overestimate the citations accumulated by the majority of scientists.

4.2.4 Sensitivity Analysis

To investigate the effect of each feature to our prediction model and their correlation with future output and with each other, we perform a sensitivity analysis on the SVR-Clust model by iteratively removing or adding subsets of the features. Our feature set for the prediction model consists of six groups which expresses different aspects of scientific status. Firstly, we remove one group of features at a time to explore the shift in model performance (see Table 4). For the majority of feature categories there is a slight drop in performance when removing them from the model as compared to the complete model (see SVR-Clust in Table 3). The *impact* related category of features are found to be the most influential ones, whose removal causes a slightly bigger drop in R^2 values. However, it is observed that performance still remains high (on average higher than 0.8) indicating that the factors influencing future output are interrelated and that the setting of the framework is robust, without substantial dependencies on individual features.

Next, we move in the opposite direction to train the model (see Table 5), adding single features individually and not as a group to further explore the contribution of each single factor. As expected, by using only one feature performance decreases significantly. However, certain features are proven to be highly influential producing a well performing model even when used individually. These features include impact features, i.e. e-index, c and h-index as well as one temporal feature, Dc, and a social feature, the core number cn. Interestingly, out of the impact features, e-index appears to be highly effective in citation count prediction and in particular more effective than h-index, since it provides insight on the maximum citation counts recorded in a scientist's portfolio. The core number, expressing the level of influence a scientist displays amongst her/his citing and cited authors, also contributes effectively to the prediction of future status. Less effective features but still contributing to an efficient prediction include number of publications (p), mean h-index of coauthors ($meanHCoA$), number of single author publications ($singAuth$) and two social features, betweenness and closeness centrality (bc, cc).

The number of single author publications has a significantly higher effect on predictions for high impact clusters than for low impact ones. When an established scientist produces a single author paper, it often contains seminal personal work thus increasing her/his impact. On the contrary, the mean h-index of coauthors affects all clusters seemingly the same, meaning that scientists of all levels reflect the quality of their collaborators. The least efficient factors, which lead to poor performance in predictive modeling when used on their own, are the JIF, the difference in h-index values (Dh) and the PageRank score (PR). Despite the fact that PageRank score has been used for evaluation of citation networks, it seems to be outperformed by other social measures (e.g. core numbers) that more clearly identify cliques and communities. Moreover, the Impact Factor of journals in which a scientist publishes offers limited predictive power over her/his future impact, thus questioning the whole journal ranking process and its relation to actual impact.

The final stage of our sensitivity analysis entails pro-

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

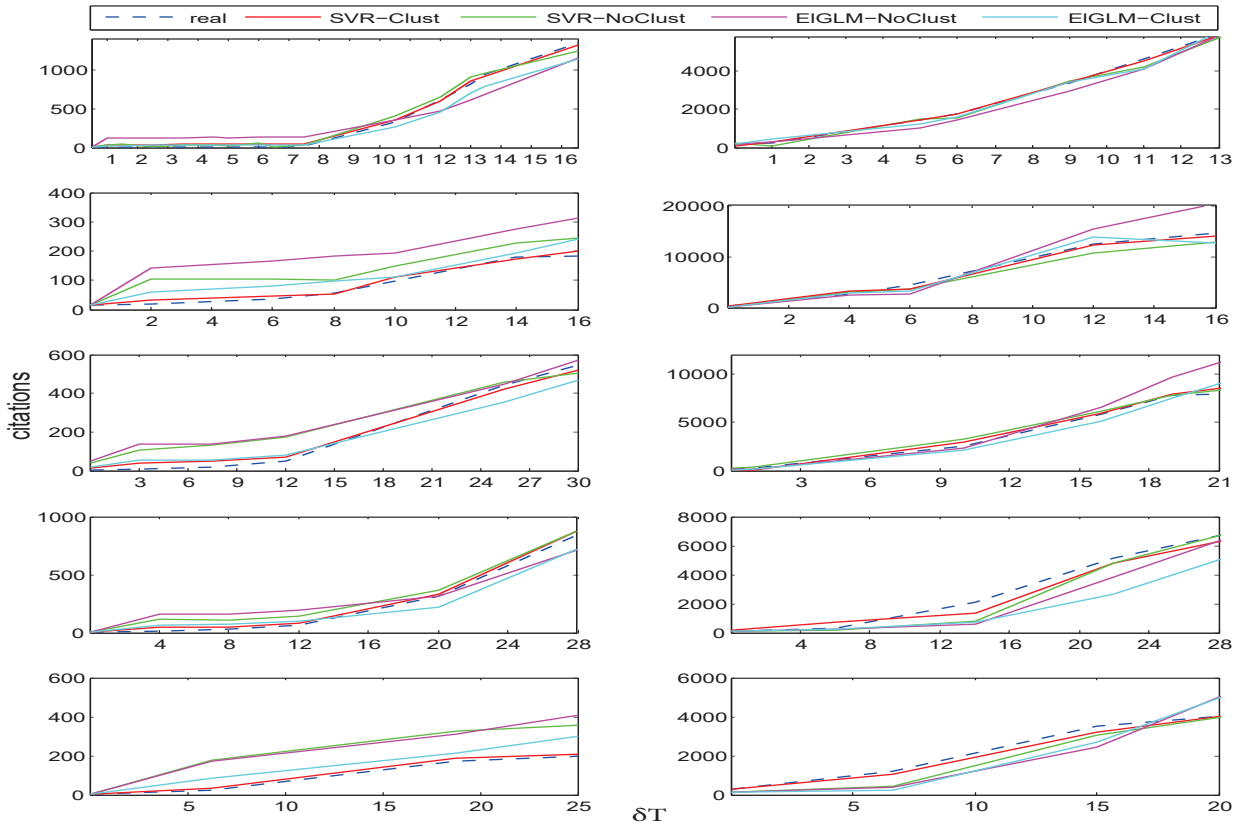

Fig. 8. Actual citation trajectories of scientists and the predicted ones by our model and three baselines. The rows represent δT values, with δT=1, 2, 3, 4 and 5 starting from top rows to bottom ones. The left column is populated with random scientists, while the right contains ACM Fellows

TABLE 4
R^2 values for our proposed framework trained with the entire set of features minus one subset of features at a time

Time Interval	Cluster	productivity	impact	co-authorship	venue	social	temporal
δT=1	cl1	0.94	0.82	0.94	0.95	0.92	0.95
	cl2	0.98	0.87	0.98	0.97	0.98	0.98
	cl3	0.98	0.91	0.99	0.99	0.97	0.99
	cl4	0.97	0.91	0.96	0.97	0.93	0.97
δT=2	cl1	0.92	0.84	0.92	0.91	0.92	0.92
	cl2	0.94	0.89	0.95	0.95	0.97	0.96
	cl3	0.98	0.93	0.98	0.98	0.97	0.98
	cl4	0.95	0.91	0.97	0.94	0.93	0.96
δT=3	cl1	0.82	0.83	0.84	0.83	0.88	0.89
	cl2	0.94	0.90	0.96	0.96	0.95	0.94
	cl3	0.95	0.92	0.96	0.96	0.96	0.96
	cl4	0.91	0.92	0.92	0.90	0.92	0.92
δT=4	cl1	0.88	0.80	0.85	0.86	0.83	0.85
	cl2	0.92	0.90	0.92	0.93	0.92	0.91
	cl3	0.94	0.93	0.95	0.93	0.94	0.91
	cl4	0.91	0.88	0.89	0.90	0.90	0.90
δT=5	cl1	0.79	0.75	0.78	0.78	0.76	0.77
	cl2	0.90	0.86	0.91	0.90	0.89	0.89
	cl3	0.93	0.92	0.95	0.93	0.93	0.93
	cl4	0.89	0.90	0.91	0.89	0.91	0.91

jecting the whole set of features on the Principal Component space for all clusters and two selected δT values of 1 and 5 years (short and long term impact). Figures 9 and 10 display the features as vectors and the colored points represent scientists belonging to each cluster. PageRank (PR) appears to be almost orthogonal with the target (future citation count) indicating a very small correlation with future status, which comes in accordance with our previous observations. Even though JIF seems closer to target, it displays rather small values (small sized vector) in both components, which

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

TABLE 5
R^2 values for our proposed framework trained with only one of the features at a time

Time Interval	Cluster	productivity	impact			co-authorship		venue	social				temporal	
		p	c	h	e	sing Auth	meanH CoA	JIF	PR	cn	bc	cc	Dc	Dh
δT=1	cl1	0.36	0.71	0.60	0.80	0.14	0.32	0.11	0.01	0.57	0.26	0.23	0.35	0.01
	cl2	0.47	0.79	0.72	0.80	0.20	0.35	0.07	0.01	0.51	0.30	0.26	0.44	0.04
	cl3	0.56	0.80	0.82	0.82	0.26	0.33	0.06	0.02	0.40	0.37	0.28	0.50	0.04
	cl4	0.65	0.80	0.82	0.81	0.34	0.40	0.06	0.05	0.38	0.51	0.33	0.66	0.08
δT=2	cl1	0.40	0.77	0.60	0.70	0.15	0.31	0.13	0.01	0.54	0.26	0.24	0.56	0.01
	cl2	0.48	0.77	0.69	0.77	0.18	0.35	0.09	0.02	0.48	0.32	0.27	0.60	0.02
	cl3	0.56	0.79	0.73	0.82	0.25	0.35	0.07	0.01	0.41	0.41	0.27	0.66	0.02
	cl4	0.62	0.79	0.77	0.82	0.32	0.34	0.07	0.04	0.33	0.53	0.33	0.69	0.09
δT=3	cl1	0.42	0.71	0.64	0.68	0.15	0.28	0.16	0.01	0.50	0.27	0.23	0.61	0.19
	cl2	0.50	0.72	0.68	0.71	0.19	0.31	0.12	0.02	0.45	0.34	0.21	0.65	0.1
	cl3	0.53	0.72	0.72	0.75	0.24	0.31	0.12	0.01	0.37	0.42	0.25	0.68	0.1
	cl4	0.63	0.73	0.72	0.76	0.41	0.33	0.10	0.05	0.34	0.57	0.27	0.68	0.25
δT=4	cl1	0.45	0.62	0.55	0.58	0.18	0.29	0.14	0.01	0.47	0.29	0.22	0.55	0.07
	cl2	0.50	0.65	0.60	0.63	0.18	0.29	0.17	0.01	0.47	0.40	0.22	0.60	0.08
	cl3	0.55	0.66	0.63	0.65	0.25	0.27	0.10	0.01	0.40	0.44	0.25	0.62	0.08
	cl4	0.65	0.68	0.63	0.67	0.38	0.37	0.10	0.07	0.43	0.59	0.31	0.62	0.19
δT=5	cl1	0.39	0.57	0.47	0.56	0.14	0.22	0.15	0.01	0.39	0.30	0.17	0.46	0.08
	cl2	0.47	0.59	0.49	0.58	0.20	0.25	0.10	0.01	0.41	0.38	0.19	0.55	0.10
	cl3	0.56	0.60	0.53	0.60	0.27	0.26	0.11	0.01	0.38	0.49	0.23	0.59	0.17
	cl4	0.58	0.61	0.54	0.61	0.32	0.38	0.10	0.01	0.44	0.60	0.28	0.59	0.34

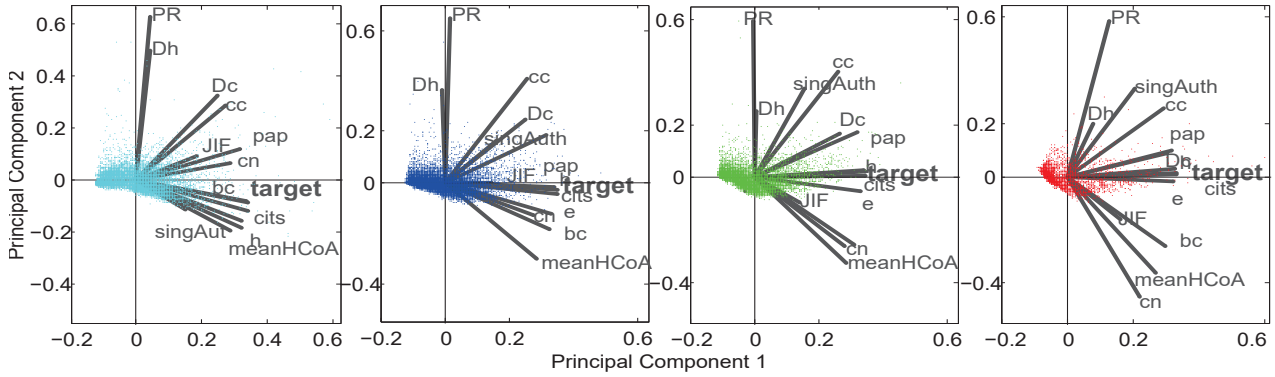

Fig. 9. Feature correlation on Principal Component space for all clusters and δT=1 year

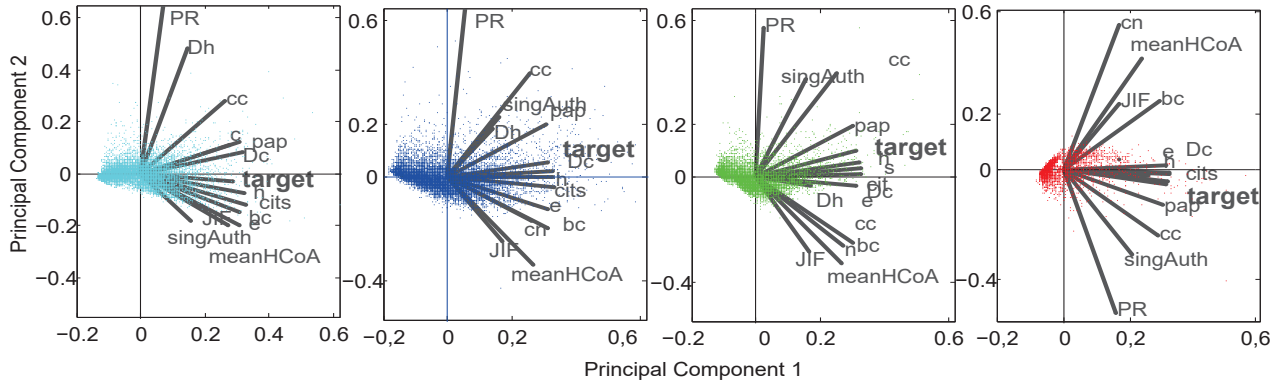

Fig. 10. Feature correlation on Principal Component space for all clusters and δT=5 years

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

suggests weak interrelations with the target and the rest of the factors. For higher impact clusters (3 and 4) e, h, Dc and c are mostly correlated with each other and the target compared to the remaining features which appear disconnected from target. For low impact clusters, on the other hand, most factors display similar level of correlation with each other and future status meaning that for up and comers who have not yet obtained established status all factors constitute an interplay to decipher future impact. In short term versus long term predictions limited changes occur in the interrelations amongst factors, with the temporal features being the only ones with substantial shifts. It is observed that temporal factors (Dh and Dc) are more related to the target in long term impact rather than in short term one. Intuitively, citation patterns arise over longer periods of time and temporal information can help identify them. JIF also seems more relevant with clusters 1 and 2 in short term impact prediction.

5 The Dynamics of Science

Based on the above sensitivity analysis and performance evaluation of our model, a number of interesting findings came out regarding the dynamics of science evolution over time and the factors influencing future impact. The general pattern that arises is that scientific output is the result of a complex interplay of personal, social and temporal factors. More specifically, the well-known JIF appears to have limited effect on scholar status, whereas in previous studies it has been found correlated with individual publication impact. A scientist's position in the social network consisting of citers and cited scholars appears to be of crucial importance for future impact. The core number appears to reveal considerable information about a scholar's impact. Since science is a social process, for all levels of performance it is of high importance to enter the well-connected core of the scientific community and form close citing relationships with seminal scientists to increase visibility of your work.

An interesting finding is the effect of past h-index shifts in future citation counts; since h-index rises slowly and remains within a limited range, the citation accumulation process may exceed the evolution of h-index. Past patterns in h-index values offer limited insight to future citation accumulation, therefore may prove misleading in assessing future potential. Regarding a scholar's collaborations, it seems that whom you publish with reflects upon your personal status even in the long run. For mature high impact scientists in particular, their choice to publish on their own (single author papers) is proven to be of high influence to their status. This can be attributed to their fame and recognition that shines through even when they do not share a strong co-authorship network. For younger scientists though, this approach may not prove as effective as they appear more dependent on their social and collaboration relationships.

6 Conclusion

In the context of this work, it is not without caution that we attempt to provide estimations of future scientific development. The goal is to account for the diversity in scientific work, through the adaptation of different predictive models per impact cluster and provide a relative cohort assignment to each scientist at any given time. It is undeniable that by relying heavily on the predictability of scientific output and automatic decision support, novel fields of research or under-represented groups can end up marginalized and discouraged. Therefore, we contemplated different academic ages and career levels, without predefined thresholds or a priori assumptions on distribution models to which any new scientist needs to adhere, thus allowing for flexibility, transparency and interpretability. In any case, the inevitable emergence of data intelligence based evaluation systems aims neither to replace human insight nor predefine the future of science, but to ensure that no scientist gets overlooked.

Acknowledgment

The authors would like to thank Dimitrios Katsaros, Antonis Sidiropoulos and Zenonas Theodosiou for their support.

References

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[2] E. Garfield, "The Impact Factor and using it correctly," *Der Unfallchirurg*, vol. 6, no. 101, pp. 413–414, 1998.

[3] D. Wang, C. Song, and A. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[4] A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Reputation and impact in academic careers," *Proceedings of the National Academy of Sciences*, vol. 111, no. 43, pp. 15 316–15 321, 2014.

[5] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[6] M. Schreiber, "How relevant is the predictive power of the h-index? a case study of the time-dependent Hirsch index," *Journal of Informetrics*, vol. 7, no. 2, pp. 325–329, 2013.

[7] A. Gogoglou, A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, "A scientist's impact over time: The predictive power of clustering with peers," in *Proceedings of the 20th International Database Engineering & Applications Symposium (IDEAS)*, 2016, pp. 334–339.

[8] S. Wang, S. Xie, X. Zhang, Z. Li, P. S. Yu, and Y. He, "Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement," *ACM Transactions on Intelligent Systems Technology*, vol. 7, no. 4, pp. 64:1–64:28, 2016.

[9] E. S. Vieira, J. A. Cabral, and J. A. Gomes, "How good is a model based on bibliometric indicators in predicting the final decisions made by peers?" *Journal of Informetrics*, vol. 8, no. 2, pp. 390–405, 2014.

[10] L. Wildgaard, J. W. Schneider, and B. Larsen, "A review of the characteristics of 108 author-level bibliometric indicators," *Scientometrics*, vol. 101, no. 1, pp. 125–158, 2014.

[11] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[12] H. W. Shen and A. Barabási, "Collective credit allocation in science," *Proceedings of the National Academy of Sciences*, vol. 111, no. 34, pp. 12 325–12 330, 2014.

[13] D. J. de Solla Price, "Networks of scientific papers," *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[14] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, no. 7415, pp. 201–202, 2012.

[15] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted?" *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[16] T. Yu, G. Yu, P. Li, and L. Wang, "Citation impact prediction for scientific papers using stepwise regression analysis," *Scientometrics*, vol. 101, no. 2, pp. 1233–1252, 2014.

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2884505, IEEE Transactions on Big Data

[17] F. Davletov, A. S. Aydin, and A. Cakmak, "High impact academic paper prediction using temporal and topological features," in *Proceedings of the 23rd ACM International Conference on Conference on Information & Knowledge Management (CIKM)*, 2014, pp. 491–498.
[18] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *Proceedings of the 14th International Symposium on String Processing & Information Retrieval (SPIRE)*, 2007, pp. 107–117.
[19] D. G. Brizan, K. Gallagher, A. Jahangir, and T. Brown, "Predicting citation patterns: Defining and determining influence," *Scientometrics*, vol. 108, no. 1, pp. 183–200, 2016.
[20] X. Cao, Y. Chen, and K. R. Liu, "A data analytic approach to quantifying scientific impact," *Journal of Informetrics*, vol. 10, no. 2, pp. 471–484, 2016.
[21] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying sleeping beauties in science," *Proceedings of the National Academy of Sciences*, vol. 112, no. 24, pp. 7426–7431, 2015.
[22] A. Mazloumian, "Predicting scholars' scientific impact," *PLOS ONE*, vol. 7, no. 11, pp. 1–5, 2012.
[23] D. McNamara, P. Wong, P. Christen, and K. S. Ng, *Predicting high Impact academic papers using citation network features*. Springer, 2013, pp. 14–25.
[24] N. Pobiedina and R. Ichise, "Citation count prediction as a link prediction problem," *Applied Intelligence*, vol. 44, no. 2, pp. 252–268, 2016.
[25] O. Penner, R. K. Pan, A. M. Petersen, and S. Fortunato, "The case for caution in predicting scientists' future impact," *Physics Today*, vol. 66, no. 4, p. 8, 2013.
[26] H. Shen, D. Wang, C. Song, and A. Barabási, "Modeling and predicting popularity dynamics via reinforced Poisson processes," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, 2014, pp. 291–297.
[27] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zhu, "On modeling and predicting individual paper citation count over time," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016, pp. 2676–2682.
[28] A. Clauset, D. B. Larremore, and R. Sinatra, "Data-driven predictions in the science of science," *Science*, vol. 355, no. 6324, pp. 477–480, 2017.
[29] P. Z. Revesz, "A method for predicting citations to the scientific publications of individual researchers," in *Proceedings of the 18th International Database Engineering & Applications Symposium (IDEAS)*, 2014, pp. 9–18.
[30] M. Nezhadbiglari, M. A. Gonçalves, and J. M. Almeida, "Early prediction of scholar popularity," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, 2016, pp. 181–190.
[31] J. Zhang, Z. Ning, X. Bai, W. Wang, S. Yu, and F. Xia, "Who are the rising stars in academia?" in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, 2016, pp. 211–212.
[32] S. Datta, P. Basuchowdhuri, S. Acharya, and S. Majumder, "The habits of highly effective researchers: An empirical study," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 3–17, 2017.
[33] J. Garner, A. L. Porter, and N. C. Newman, "Distance and velocity measures: Using citations to determine breadth and speed of research impact," *Scientometrics*, vol. 100, no. 3, pp. 687–703, 2014.
[34] M. Wang, G. Yu, and D. Yu, "Effect of the age of papers on the preferential attachment in citation networks," *Physica A: Statistical Mechanics & its Applications*, vol. 388, no. 19, pp. 4273–4276, 2009.
[35] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee, "On the categorization of scientific citation profiles in computer science," *Communications of the ACM*, vol. 58, no. 9, pp. 82–90, 2015.
[36] A. Sidiropoulos, A. Gogoglou, D. Katsaros, and Y. Manolopoulos, "Gazing at the skyline for star scientists," *Journal of Informetrics*, vol. 10, no. 3, pp. 789–813, 2016.
[37] M. J. Zaki, M. J. W., and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.
[38] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.
[39] M. Y. Kiang, "Extending the kohonen self-organizing map networks for clustering analysis," *Computational Statistics & Data Analysis*, vol. 38, no. 2, pp. 161–180, 2001.
[40] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11 150–11 154, 2007.
[41] K. Shin, T. Eliassi-Rad, and C. Faloutsos, "Corescope: Graph mining using k-core analysis - patterns, anomalies and algorithms," in *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 2016, pp. 469–478.
[42] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
[43] Y. Dong, R. A. Johnson, and N. V. Chawla, "Will this paper increase your h-index?: Scientific impact prediction," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: ACM, 2015, pp. 149–158.
[44] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 990–998.
[45] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, "To better stand on the shoulder of giants," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2012, pp. 51–60.
[46] C. Mccarty, J. W. Jawitz, A. Hopkins, and A. Goldman, "Predicting author h-index using characteristics of the co-author network," *Scientometrics*, vol. 96, no. 2, pp. 467–483, 2013.
[47] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.
[48] C.-T. Zhang, "The e-index, complementing the h-index for excess citations," *PLoS One*, vol. 4, no. 5, 2009.
[49] C. Biscaro and C. Giupponi, "Co-authorship and bibliographic coupling network effects on citations," *PLOS ONE*, vol. 9, no. 6, 2014.
[50] Y. Rochat, "Closeness centrality extended to unconnected graphs: The harmonic centrality index," in *Proceedings of the 6th Conference on Applications of Social Network Analysis (ASNA)*, 2009.
[51] Q.-S. Xu and Y.-Z. Liang, "Monte carlo cross validation," *Hemometrics & Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.
[52] A. J. Smola and B. Sch´olkopf, "A tutorial on Support Vector Regression," *Statistics & Computing*, vol. 14, no. 3, pp. 199–222, 2004.
[53] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

Antonia Gogoglou holds a 5-years Diploma degree in Electrical and Computer Engineering from Aristotle University of Thessaloniki. Currently, she is member of the Datalab team and has completed her PhD with the Department of Informatics of the same university. She has been awarded as Marie Curie RISE Fellow in 2016. Her research interests include complex network analysis, graph theory, data mining, knowledge discovery and predictive modeling with a focus towards Scientometrics and Social Networks.

Yannis Manolopoulos serves as Vice Rector of the Open University of Cyprus. He has been with the Aristotle University of Thessaloniki, the University of Toronto, the University of Maryland at College Park and the University of Cyprus. He has also served as Rector of the University of Western Macedonia in Greece and Vice-Chair of the Greek Computer Society. For his works in Data Management he received >12,000 citations with h-index=52. Currently, he is member of the Editorial Board of Information Systems, WWW Journal and The Computer Journal.

2332-7790 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.