Characterizing Scholar Popularity: A Case Study in the Computer Science Research Community

Glauber D. Gonçalves, Flavio Figueiredo, Jussara M. Almeida, Marcos A. Gonçalves
Department of Computer Science, Universidade Federal de Minas Gerais, Brazil
{ggoncalves,flaviov,jussara,mgoncalv}@dcc.ufmg.br

ABSTRACT

A common live debate among scholars regards the popularity, productivity and impact of research. This paper aims to contribute to such discussion by quantifying the impact of various academic features on a scholar popularity throughout her career. Using a list of over 2 million publications in the Computer Science research area obtained from two large digital libraries, we analyze how features that capture the number and rate of publications, number and quality of publication venues, and the importance of the scholar in the co-authorship network relate to the scholar popularity. We also investigate the temporal dynamics of scholar popularity, identifying a few common profiles, and characterizing scholars in each profile according to their academic features.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Citation analysis, scholar popularity dynamics, academic scholar features

1. INTRODUCTION

What factors contribute to a successful career in research? This is a fundamental and broad question that draws the attention of all scholars. Success in research can be assessed in terms of various measures. The acknowledgement by peers of the value of a researcher's publications (e.g., by citing them) is one of the most sought-after measures of scholarly success, as it can be seen as an estimate of her influence and visibility in the community, and ultimately of her scholarly popularity [1].

Investigating the factors that impact the popularity of a scholar can shed light into proactive actions that might guide decisions to shape a career in research. Moreover, from a system standpoint, it can draw useful insights into the design of cost-effective popularity prediction methods, which, in turn, can be exploited for improving various services (e.g. expert or academic collaboration recommendation services [2]). Some questions of interest are: *to which extent do the quantity and the quality of the publication venues impact a scholar popularity? To which extent is the number of publications related to popularity? What is the role that the co-authorship network plays in the scholar popularity?*

Influence, productivity and popularity in research has already been tackled in various prior studies since the early 20^{th} century [3, 4]. For example, many authors have characterized influential publications [5, 6], some with the goal of designing models to predict the number of times a particular piece of work will be cited [7, 8, 9]. Others have analyzed various impact factors to evaluate the influence of publication venues (e.g., journals) [10, 11] and researchers [12, 13, 14] from citations.

In this paper, our goal is to investigate and *quantify* the factors that impact the popularity of a scholar during her career. Complementing prior analyses [15, 1, 9], we intend to assess the impact of various academic characteristics, or *features*, on the popularity a scholar achieves during the career. We also want to study the temporal popularity dynamics of various scholars, identifying a few common profiles, and characterizing scholars in each profile according to their popularity and academic features.

As in [16, 17, 15], we use the *total number of citations* to estimate the popularity of a scholar. Although other indices, such as weighted metrics [6, 1] and PageRank [10, 18], could be adopted, our choice is based on two factors. Firstly, some authors [16] have argued that citation counts are better indicators of the scientific contribution of researchers, disciplines or nations than impact factors (such as the h-index [12]). Secondly, considering qualitative aspects of the research is very subjective and would require the use of metrics with debatable biases and criticisms [19, 11, 20]. Thus, we assess scholar popularity by the total number of citations, leaving the analysis of other indices to future work.

Specifically, we focus on scholars of a specific research field - Computer Science, and crawl statistics about their publication records from two large digital libraries, namely ArnetMiner[1] and Microsoft Academic[2]. We start by studying how various academic features[3] are correlated with the popularity of a scholar. The following features are analyzed: total number of publications, yearly rate

[1]http://arnetminer.org
[2]http://academic.research.microsoft.com
[3]Features that reflect the academic activity of a scholar.

978-1-4799-5569-5/14/$31.00

of publications, number of distinct publication venues as well as the venues' quality, and the importance of the author in the co-authorship network (estimated by various centrality metrics). We also explore regression based models to quantify the relative importance of each feature to the final popularity of the scholar. Moreover, since the impact of a feature may vary over time, as the scholar's career progresses, these analyses are performed separately for different groups of scholars, categorized based on the number of years of academic experience.

Next, we employ a time series clustering algorithm named K-Spectral Clustering [21], recently proposed to study the popularity of online content, to identify profiles of scholar popularity dynamics. We also characterize scholars in each profile in terms of their popularity and academic features.

Our results indicate that although most analyzed features are strongly correlated with popularity, only two features are required to explain practically all the variation in popularity across different scholars, even considering only the less experienced scholars. These two features are: the total number of publications and the average quality of the publication venues of the scholar (estimated by the average number of citations per publication of the venue). Out of them, the number of publications is the most important one, with a major impact on popularity which increases with the scholar's experience. The relative importance of the average quality of the publication venues in turn decreases for the most experienced scholars. We also uncovered five different profiles of popularity temporal dynamics. Three profiles correspond to scholars who succeed in becoming more and more popular with time, whereas the other two correspond to scholars whose popularity curve exhibits a clear decay after a popularity peak. Our characterization of these profiles also suggest that the most popular scholars, who fall in the first three profiles, are typically those who keep publishing over time.

The rest of this paper is organized as follows. Section 2 discusses related work, and Section 3 presents the datasets and academic features analyzed. The assessment of the importance of each feature to scholar popularity is presented in Section 4, whereas profiles of scholar popularity dynamics are identified and characterized in Section 5. Conclusions and future work are offered in Section 6.

2. RELATED WORK

Quantitative measures of research have long been studied. Cason and Lubotsky [3] conducted one of the earliest citation analysis studies with focus on measuring dependences among journals. A few decades later, Pinski and Narin [4] evaluated the influence of journals by taking both the number of citations and the importance of the citing journal into account. Since then, various studies have proposed techniques to measure and/or analyze the influence, popularity or productivity in scientific research.

Some authors focused on proposing different metrics of research performance. For example, Ding and Cronin [1] distinguished between weighted and unweighted citation counts, using the former as a measure of scholar prestige and the latter as a measure of scholar popularity. As argued in [20], there are advantages and disadvantages associated with each type of metric, as each one has its own bias. For instance, the impact factor is a widely used measure to compare the influence of publication venues [11]. However, impact factor does not reflect the influence of individual papers and authors [19, 11]. Similarly, the h-index is commonly used to measure the performance of researchers or publication venues [12]. However, it might lead to counter-intuitive results due to its attempt to bring together measures of productivity (e.g., total number of papers, references, and citations) and impact factor under a single denominator [20]. Accordingly, novel approaches to measure prestige or influence in research have also been proposed, including improved versions of h-index, such as g-index [13] and h_m-index [14], automatically learned metrics based on machine learning techniques [22], as well as customized indices sensitive to the productivity of researchers in different research fields [23].

Unlike these prior studies, our goal is not to propose metrics of research performance, but rather assess the importance of various factors to this performance as well as characterizing how the academic performance of scholars evolve over time. To that end, we focus on scholar popularity estimated by the total number of citations (as defined in [1]). Though simple and easy to compute, this metric has been shown to be very important for various types of analyses. For instance, in [16], the authors analyzed the productivity and impact of more than 700 biomedical researchers in Finland from 1966 to 2000, showing that actual publication and citation counts are better indicators of the scientific contribution of researchers, disciplines, or nations than impact factors. In [17], the authors used unweighted citation counts (i.e., popularity) to analyze the impact of experience and prestige on the number of references scholars use in their publications.

Others have tackled the prediction of popularity of publications or scholars. For example, in [7, 8], the authors used measures computed after a paper was published (e.g., number of downloads) to predict its future citation count. Yan *et al.*, in turn, exploited only features available at publication time, extracted from ArnetMiner, as inputs to linear and support vector regression models to perform such predictions [9]. In contrast, Acuna *et al.* [24] presented a model to predict the future h-index of a scholar using a linear combination of features related to the scholar's publications, citations, and funding. Our work complements these prior studies, as it aims at assessing the relative importance of various features to scholar popularity, drawing insights that can help improving existing prediction solutions [24] as well as designing new methods.

Another set of related studies applied complex network metrics to the co-authorship graph to determine the influence of a scholar within a research community [25]. Liu *et al.* [26] proposed AuthorRank, a weighted version of PageRank, to match the committee members of the Digital Libraries research community within its co-authorship graph. PageRank has also been exploited to assess the relative importance of publications, journals or authors in the co-authorship and citation networks [10, 5, 18]. Similarly, other centrality metrics, such as degree, closeness and betweenness, have also been shown to be significantly correlated with citation counts [15], whereas some recent efforts focused on structural properties that reflect the behavior of authors in the co-authorship network, such as interactions with authors in the largest connected component and reciprocity [27, 28]. Our work complements those studies as we are interested in comparing and

quantifying the importance of various scholar features to scholar popularity dynamics, including but not restricted to features related to the co-authorship network.

Finally, we are aware of only a couple of efforts to analyze the evolution of scholar popularity over time. Cronin and Meho [29] explored the relationship between scholar creativity, estimated by the number and total citation count of high impact works, and (both chronological and professional) age of 12 important scholars in the Information Science field. They found that creativity is expressed at different stages and with different intensities in the careers of those scholars. Ding and Cronin [1] analyzed the popularity and the prestige of the top 40 ranked authors in the Information Retrieval field for four "time bands" in their careers. Their main conclusion is that, unlike the prestige ranks of scholars whose behavior is stable, popularity ranks change over time. Although those studies provide interesting insights into the popularity evolution of successful scholars, they are focused on a small set of leading researchers in their fields. Thus, they do not identify common profiles of popularity dynamics across a large set of scholars, with various levels of popularity. We try to fulfill this gap by making use of the K-Spectral Clustering (KSC) algorithm [21], which was recently proposed to study popularity dynamics of on-line content. To our knowledge, we are the first to use such advanced tools to better understand scholar popularity dynamics.

3. METHODOLOGY

In this section we present the datasets (Section 3.1) and the academic features (Section 3.2) analyzed in this work.

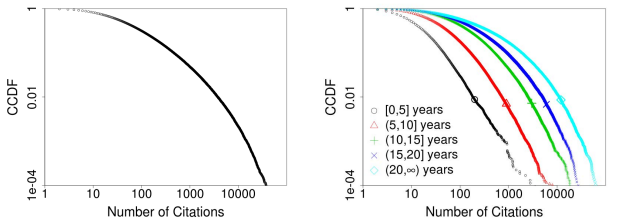
3.1 Datasets

We use datasets obtained from two large and very popular academic digital library services. The first dataset is publicly available at the ArnetMiner (AM) service[4], which is a free on-line service used to index and search academic social networks. The dataset consists of a list of publications from the Computer Science community, covering the period from 1936 to 2013. Each instance in the dataset is specified by the following attributes: author names, publication venue, year, total number of citations, and list of references. In total, our AM dataset includes 2,244,018 publications by 831,763 authors in 8,274 venues, which, collectively, received 38,770,182 citations. We use this dataset to characterize the importance of various academic features to scholar popularity, estimated by the total number of citations the scholar received in all her publications in this dataset.

Recall that we are also interested in characterizing the popularity dynamics of different scholars. We could estimate the popularity of each author in each year from the references attribute in the AM dataset, and thus build a set of popularity time series. However, we noticed that this attribute is not reliable as many publications have empty lists of references. Indeed, the popularity of each author inferred from this attribute was often much smaller than the aggregated popularity computed by summing up the citations of all the author's publications in the dataset.

Given that it was not possible to create consistent and accurate popularity time series for each author using the AM dataset, we relied on another data source, namely the Microsoft Academic Research (MS-AR) platform. Each author name that appeared in the AM dataset was submitted as a query to the MS-AR service to retrieve the citation time series and publication time series of the author. We considered the most common among all aliases provided by AM for the author name, and the best match returned by MS-AR, which also has its own disambiguation mechanisms for queries[5]. We successfully retrieve the time series for around 75% of the author names (i.e, 624,784 authors).

[4]http://arnetminer.org/citation



(a) All scholars (b) Scholars in each experience group

Figure 1: Scholar popularity distribution.

We note that in both datasets, but particularly in the AM dataset, a large number of scholars had a single publication. We removed these scholars from our datasets, focusing only on scholars with at least 2 publications. After this filtering, we were left with 402,720 scholars in the AM dataset and 437,446 popularity time series in the MS-AR dataset.

We also note that there are discrepancies in the total citation counts of the same author in both datasets, which might be due to different citation coverages of each digital library. However, we could find no clear trend towards a larger coverage of most authors in a single dataset[6]. We deal with such discrepancies in our investigation by focusing each analysis on a *single* dataset.

Figure 1-a) shows the distribution of popularity (number of citations) for all scholars in our AM dataset. Note the logarithm scale in both axes. The figure shows that, consistently with prior results [17], the distribution of scholar popularity is heavy tailed. This heavily skewed distribution may reflect a natural heterogeneity across scholars, but it may also reflect a bias due to different levels of experience. Thus, we split the scholars in each dataset into 5 *experience groups* based on their time of research activity (up to 2013). We estimate the time of activity t of a scholar by the year of her first publication in the AM dataset, and group scholars according to the following ranges: $t \leq 5$ years; $5 < t \leq 10$ years; $10 < t \leq 15$ years; $15 < t \leq 20$ years; and $t > 20$ years. Table 1 provides the numbers of authors (AM dataset) and popularity time series (MS-AR dataset) in each experience group, after the aforementioned filtering. Moreover, Figure 1-b) shows that the distribution of scholar popularity remains very skewed even considering a single experience group, for all groups. Thus, there is a lot of heterogeneity, in terms of popularity, even among scholars with approximately the same time of experience.

[5]In other words, we relied on the name disambiguation solutions provided by both MS-AR and AM services [30, 31] to help solving possible author name inconsistencies.

[6]Indeed, we analyzed the coverage of each dataset for a few authors and did observe a significant number of missing publications in both datasets.

Table 1: Distribution of scholars and popularity time series across experience groups in the *filtered* datasets (ranges specify time of research activity).

	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
# authors	53,642	148,567	91,932	50,435	58,144
# popularity time series	76,980	133,501	99,234	60,521	67,210

Table 2: Scholar academic features.

Notation	Description
$nPubs$	total number of publications
$yPubRate$	yearly publication rate
$nVenues$	number of distinct venues
$CitVen_{max}$	maximum number of citations of any venue
$CitVen_{avg}$	average number of citations per venue
$CitPubVen_{max}$	maximum number of citations per publication of any venue
$CitPubVent_{avg}$	average number of citations per publication per venue
$nCoauthors$	number of co-authors
$closeness$	closeness in the co-authorship network
$PageRank$	PageRank in the co-authorship network

3.2 Scholar Features

The popularity of a scholar may depend on a multitude of academic, social and even economic factors. We here focus on academic factors, and relate the popularity of a scholar to features that capture, quantitatively and qualitatively, her productivity as a publication author, as well as her importance in the co-authorship network. Unless otherwise noted, these features are computed from the AM dataset.

The productivity of a scholar is here estimated by the total number of publications over the period covered by the dataset, as well as by the yearly publication rate, given by the ratio of the total number of publications to the time period between the first and last publications.

We also use the total number of unique publication venues as well as different estimates of venue quality as scholar features. We characterize the quality of the publication venues of a scholar in terms of both *average* and *maximum* quality, using two estimates of venue quality: total number of citations of all publications and average number of citations per publication, both computed for all publication venues of the scholar. Thus, we use 4 features to capture the quality of the publication venues of a scholart.

We also relate the popularity of a scholar to her importance as a co-author. To that end, we build a co-authorship network where edges between two authors are added if they co-authored at least one publication. We assess the importance of an author in this network using three centrality metrics: degree, closeness and PageRank. The vertex degree is the number of co-authors of the scholar. The closeness is defined by the inverse of the shortest path distances from the vertex to all other vertices in the network[7]. The PageRank [18], in turn, can also be seen as a measure of the influence of the scholar given her position in the co-authorship network. Indeed, its use as an index of scholar productivity has already been proposed [10, 5, 18].

In sum, we characterize scholars in terms of the 10 academic features shown in Table 2.

[7]In case of unreachable vertices, the shortest path distance is assigned to the total number of vertices in the graph.

Table 3: Pearson correlations between features and scholar popularity (after logarithm transformation).

Academic Feature	Experience Group				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.544	0.656	0.704	0.753	0.813
$yPubRate$	0.194	0.383	0.528	0.572	0.592
$nVenues$	0.362	0.558	0.633	0.700	0.772
$CitVen_{max}$	0.330	0.503	0.551	0.562	0.556
$CitVen_{avg}$	0.297	0.416	0.409	0.353	0.266
$CitPubVen_{max}$	0.435	0.604	0.649	0.642	0.634
$CitPubVen_{avg}$	0.400	0.509	0.479	0.377	0.289
$nCoauthors$	0.340	0.474	0.572	0.639	0.705
$closeness$	0.230	0.385	0.546	0.621	0.705
$PageRank$	0.279	0.410	0.524	0.601	0.670

4. IMPACT OF ACADEMIC FEATURES ON SCHOLAR POPULARITY

In this section, we analyze how each academic feature is related to the scholar popularity. We start by quantifying the correlations between each feature and popularity (Section 4.1). Next, we make use of a regression model to quantify the importance of each feature to popularity.

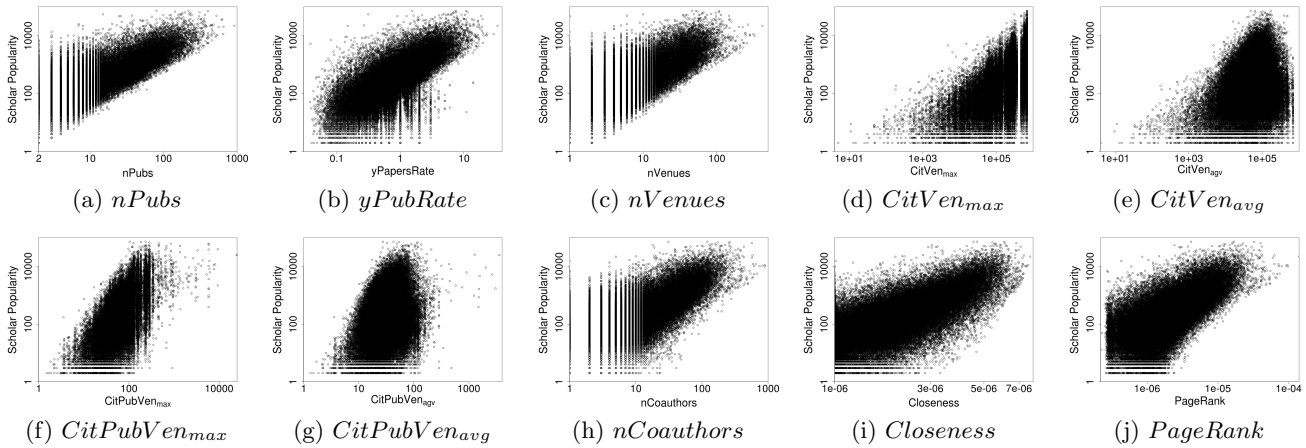
By assessing the relative importance of each feature to popularity and quantifying the strength of their relationship, we intend to provide insights that can drive the future design of methods to predict the popularity of a scholar, which in turn can be exploited for improving various recommendation services (e.g., expert recommendation, academic collaboration recommendation [2], etc).

4.1 Correlation Analysis

We quantify the correlations between each feature and popularity using both the Pearson linear correlation coefficient (ρ_p) and the Speaman's rank correlation coefficient (ρ_s) [32]. The latter is a nonparametric measure of statistical dependence between two variables that does not assume linear relationships. We note that, as observed for popularity (Figure 1), all features exhibit great variability across authors, even considering authors of a single group. Thus, before computing these correlations, we first apply a logarithm transformation on the scholar popularity and feature values to reduce their large variability (as in [17]).

The Pearson correlation coefficients between each feature and popularity for authors in each experience group are shown in Table 3. The Spearman correlation coefficients are very similar, and thus are omitted. For illustration purposes, Figure 2 presents scatter plots of each feature and popularity for scholars with more than 20 years of experience. Note the log scale in both axes.

Table 3 shows that, for most features, the correlations with popularity tend to strengthen with the scholar experience. Less experienced scholars, who are still building their careers, might be subject to other factors (e.g., lack of funding, low visibility in the academic community) that impact their popularity. The exceptions are the venue quality features: their correlations with scholar popularity tend to decrease for the most experienced scholars (scholars with more than 20 years of experience), particularly for features related to the average quality of the venues. This might reflect that such scholars are already very known in the community. Thus, their popularity is less influenced by the venues where they publish.



(a) $nPubs$ (b) $yPubRate$ (c) $nVenues$ (d) $CitVen_{max}$ (e) $CitVen_{avg}$

(f) $CitPubVen_{max}$ (g) $CitPubVen_{avg}$ (h) $nCoauthors$ (i) $Closeness$ (j) $PageRank$

Figure 2: Correlations between academic features and popularity (scholars with more than 20 years of experience; log scale on both axes).

Focusing on any experience group, the feature that is most strongly correlated with popularity is the number of publications[8]: the correlation reaches 0.81 for the most experienced scholars (both Pearson and Spearman coefficients). Even for scholars with at most 5 years of experience, the correlation is still quite strong (above 0.5). Similarly, the number of distinct venues is also strongly correlated with popularity, particularly for scholars with more than 5 years of experience.

Regarding the other features, we observe an inversion of roles across the experience groups. For less experienced scholars, features that capture the quality of the publication venues, particularly the quality of the best venue (maximum quality), are more strongly correlated with popularity than the centrality metrics (notably PageRank and closeness), possibly because these scholars are still building their co-authorship network. Indeed, the number of distinct co-authors is only moderately correlated with popularity ($\rho_p < 0.5$) for scholars with up to 10 years of experience. As the time of experience increases, all centrality metrics become increasingly more correlated with popularity[9]. That is, the scholar's role in the co-authorship network becomes more important. For the most experienced scholars, the correlations between each of these features and popularity are between 0.67 and 0.7. In contrast, as mentioned, metrics related to venue quality become less correlated with popularity as the time of experience increases. The yearly publication rate is not very strongly correlated with popularity for any experience group.

These correlations provide evidence of features that are strongly related to scholar popularity, and thus can help explain it. They may also be interpreted as providing some "advice" for scholars at different stages of their careers. For instance, based on the observed patterns, less experienced scholars might want to focus on publishing in the highest quality venues as given by the average number of citations per publication and work on building their citation networks. More experienced researchers, in turn, should never loose the focus on the number of publications, but, given that their names are already known within their communities, they might also be more flexible regarding their choices of publication venues, occasionally publishing in smaller events which can bring other benefits (e.g., close interactions) compared to larger and possibly more highly cited venues. We note, however, that these are mainly speculative suggestions. We cannot claim any causality relation from the correlations, as such claims would require specific causality tests, which are left to future work.

We note that some of these features, despite being highly correlated with popularity, might be redundant to explain the popularity of a scholar, as they are strongly correlated between themselves. For example, the Pearson correlation between number of publications and number of venues is above 0.75 for all experience groups, exceeding 0.9 for the three most experienced groups. Both features may not be needed *jointly* to capture scholar popularity. Which of the considered features are redundant? Is there a subset of the features that explain most of the popularity variations observed across scholars? The answers to these questions can provide key insights into which factors must be considered to design effective and efficient scholar popularity prediction models. We address these questions next.

4.2 Regression Analysis

We now make use of regression models to further assess the relative importance of each academic feature to scholar popularity. This investigation complements the correlation analysis in the previous section, as our goals here are: (1) identify which of the considered features are required to build a model that *describes* reasonably well the popularity of scholars in each experience group, (2) quantify the importance of each feature, as well as (3) identify and disregard redundant and unnecessary features.

We employ ordinary least squares (OLS) multivariate linear regression model to estimate a response variable $\mathcal{R}$ as a linear function of k predictor variables (i.e., features) x_1, x_2, ... x_k, that is:

$$log(\mathcal{R}) = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + \cdots \beta_k log(x_k)$$

[8]This is consistent with [24], but in a different context.
[9]This is consistent with [15], but again in a different context.

Table 4: Quality of regression with all features and after removing one feature at a time.

Regression Model	Model Quality (R^2)				
	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
All Features	0.450	0.621	0.696	0.737	0.785
$nPubs$ (-)	0.337	0.566	0.656	0.699	0.741
$yPubRate$ (-)	0.449	0.619	0.694	0.736	0.785
$nVenues$ (-)	0.449	0.621	0.696	0.736	0.785
$CitVen_{max}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitVen_{avg}$ (-)	0.450	0.621	0.696	0.736	0.785
$CitPubVen_{max}$ (-)	0.445	0.617	0.694	0.735	0.784
$CitPubVen_{avg}$ (-)	0.430	0.591	0.666	0.709	0.763
$nCoauthors$ (-)	0.444	0.617	0.693	0.733	0.781
$closeness$ (-)	0.450	0.621	0.696	0.735	0.784
$PageRank$ (-)	0.450	0.620	0.695	0.734	0.782

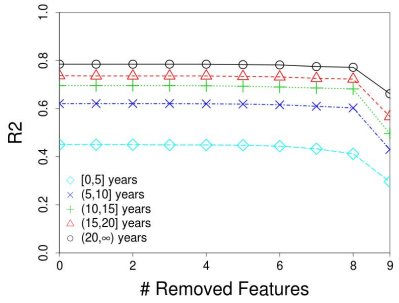


Figure 3: Quality of regression models (R^2) as features are removed in order of importance.

As in the previous section, we apply a logarithm transformation in the raw data before building the model. Thus, the response $\mathcal{R}$ and the predictors x_i are, respectively, the logarithms of the popularity and feature values. We build one model for each experience group, determining parameters β_0, β_1, $\cdots$, β_k by the minimization of the least squared errors over the data for all authors in the group. We use all authors in the group to build each model as our interest is in *describing* scholar popularity[10]. The quality of the model is estimated by the coefficient of determination R^2, which captures the fraction of the total variation in the response $\mathcal{R}$ that can be explained by the predictors [32][11].

The R^2 values for the models produced using all academic features ($k = 10$) for each experience group are shown in the first line of Table 4. We note that the models can reasonably well explain the popularity of scholars with more than 5 years of experience, with R^2 exceeding 0.6 and reaching 0.78 for the most experienced group. For the least experienced scholars, the R^2 is smaller (0.45), though similar to other regression based analysis of scholar citing behavior [17], reflecting that, as previously mentioned, other factors are also important to explain popularity for young researchers.

We tested the statistical significance of each model parameter to identify those that can be disregarded. To that end, we set up a series of hypothesis tests, one for each model parameter β_i, specified by a null hypothesis $H_o : \beta_i = 0$ However, we found that most parameters are statistically significant (i.e., non-zero), with 95% confidence level, for all 5 models. This implies that the effects of most features to scholar popularity cannot be neglected. The only exceptions (p-value > 0.05) are the coefficients associated with closeness (in the model for scholars with 5-10 years of experience) and $CitVen_{max}$ (in the model for scholars with up to 5 years).

However, despite statistically significant (i.e., non-zero effect), some features might still be redundant to the model because they are strongly (linearly) correlated with other features. In other words, the complete model, with 10 predictors, might be unnecessarily complex, with parameters (β_i) that are hard to interpret. Thus, our goal is to identify the smallest set of non-redundant features from which we can build a regression model that explains scholar popularity as accurately as the full model.

To that end, we first assess the importance of each feature *individually* to model quality by removing the feature, building a new regression model with the other k=9 features, and evaluating the impact of the removal on model quality. Table 4 shows the quality of the models built when each feature, identified by (-), is removed. Note that, consistently for all groups, the removal of either the number of papers – $nPubs$ – or the average venue quality based on number of citations per publication – $CitPubVenue_{avg}$ – produces the greatest reductions on model accuracy. This suggests that each of the other 8 features, despite being (strongly) correlated with popularity, is redundant *when taken in combination with the others*. That is, each of them can be individually removed, as its impact on popularity is mostly captured by some of the remaining features.

To identify the smallest set of features that *jointly* capture the impact of all considered features on popularity, we first sort all features by their importance, estimated by the impact on R^2 caused by its removal. Then, we build new regression models removing one feature at a time *cumulatively*, starting with the least important one. That is, we first build a model with the 9 most important features; then with the 8 most important features, and so forth up to a model with a single (most important) feature.

Figure 3 shows the R^2 of the models produced as a function of the number of removed features, for all experience groups. The removal of up to 8 features has practically no impact on model accuracy. This means that only two of the considered features are necessary to explain all the variations in popularity that can be explained by the full model (i.e., with 10 features). These two features are: (1) the number of publications ($nPubs$), and (2) the average quality of the publication venues estimated by number of citations per publication in the venue ($CitPubVen_{avg}$).

Table 5 shows the values of parameters β_i in the models produced with the two best features (including the intercept β_0). The parameters associated to both features are statistically significant (i.e., non-zero) with 95% confidence for all models. Moreover, the correlations between both features ($nPubs$ and $CitPubVen_{avg}$) are very weak (near zero): the Pearson correlations ρ_p are 0.11, 0.15, 0.07,-0.02 and -0.05 for the [0;5], (5;10]; (10;15]; (15;20] and (20; ∞) groups, respectively. Thus, the impact of these two features on scholar popularity mostly complement each other.

[10]Alternatively, regression models could also be exploited to *predict* scholar popularity, in which case the data should first be split into training (model parameterization) and test sets (model evaluation). The design of prediction models is outside the present scope, and is left for future work.

[11]We also computed the adjusted R^2, which takes into account the number of predictors in the model, finding quantitatively similar results in all cases.

Table 5: Coefficients β_i of the regression models with the two best features.

Predictor	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
Intercept (β_0)	0.004	-0.050	-0.240	-0.409	-0.486
$nPubs$	1.251	1.100	1.083	1.155	1.250
$CitPubVen_{avg}$	0.455	0.756	1.013	1.082	1.029

We note that $nPubs$ is clearly the most important feature to explain scholar popularity, and this importance increases with the scholar experience. This can be attested by the correlations observed in Table 3, which are higher for $nPubs$ and increase with the scholar experience. Alternatively, we can quantify the relative importance of each feature by running two linear regressions with a single predictor. The R^2 of the produced models represent the fractions of the total variation in scholar popularity that can be explained by each considered predictor *individually*[12]. These R^2 values are shown in Table 6. Note that if we compare the R^2 of the single-predictor models with the R^2 of the two-predictor models (also shown in the table), we conclude, once again, that: (1) $nPubs$ clearly can explain a much larger fraction of the total variation that can be explained for all experience groups; and (2) the relative importance of $CitPubVen_{avg}$ decreases for the two most experienced groups. These results are consistent with our discussion in Section 4.1.

Table 6: Relative importance of each of the two best features to popularity: R^2 of regression models with a single predictor and with two predictors.

Predictors	[0;5]	(5;10]	(10;15]	(15;20]	(20,∞)
$nPubs$	0.30	0.43	0.50	0.57	0.66
$CitPubVen_{avg}$	0.16	0.26	0.23	0.14	0.08
Both	0.41	0.60	0.68	0.72	0.77

5. TEMPORAL DYNAMICS OF SCHOLAR POPULARITY

The previous section focused on the total popularity acquired by a scholar in her career. We now analyze the *evolution* of scholar popularity over time. Our goals are twofold: (1) identify common *profiles* of popularity temporal dynamics (Section 5.1); and (2) characterize scholars in each profile in terms of their academic features (Section 5.2).

With those goals, we aim at not only producing valuable knowledge to the Scientometrics field (we are not aware of any similar prior study), but also drawing insights that can help the design of effective scholar popularity prediction methods. For instance, in [33], the authors concluded that the prediction of popularity of on-line content can greatly benefit from the knowledge of popularity profiles. This is because, given the diversity of observed profiles, building a specialized prediction model for each profile produces more accurate predictions. Although this observation was made in a different context, the same general principle might hold also in the context of scholar popularity.

We use our MS-AR dataset to study popularity temporal dynamics, focusing on scholars in the two most experienced groups, i.e., scholars with more than 15 years of experience, as their long-term popularity dynamics have already stabilized. As shown in Table 1, each of these two groups accounts for at least 60,000 popularity time series extracted from the MS-AR dataset.

5.1 Identifying Popularity Profiles

To identify profiles of popularity dynamics, we make use of a recently proposed time series clustering algorithm, called K-Spectral Clustering (KSC) [21], which was used to study the patterns of popularity dynamics of on-line content. As far as we know, this is the first time it is used to understand popularity dynamics of scholarly research.

The KSC algorithm groups times series based on the *shape* of the curve, and thus respects invariants of scale in the popularity axis and shifts in the time axis. That is, two scholars that have their popularities evolving according to similar processes (e.g., linear growth) will be assigned to the same cluster by KSC, regardless of the popularity values. For example, two authors that have stable popularity over time except for a peak in a single year will be clustered together, regardless of the time when the peak occurred and the peak value. These invariants allow us to focus on the *patterns* of popularity evolution, rather than on specific time intervals and popularity values that lead to such patterns.

KSC requires that all time series have the same number of points. Thus, we represent each scholar in an experience group by a vector s of n elements, with each element representing the scholar popularity (i.e., number of citations) in one year, starting in the year of its first publication. We define n to be the minimum number of years of experience in the group, so as to meet the requirement of equal size time series. Although there are scholars with more years of experience in each group, looking into the first 15 or 20 years of experience should be enough to understand their long-term popularity dynamics.

KSC is mostly a direct translation of the K-Means algorithm [21], except for the distance metric used to capture the similarity between the popularity curves of two scholars with scale and time shifting invariants. Given the popularity curves of scholars a and b represented by vectors s_a and s_b, respectively, KSC uses the following distance metric:

$$dist(s_a, s_b) = \min_{\alpha,q} \frac{||s_a - \alpha s_{b(q)}||}{||s_a||}, \quad (1)$$

where $s_{b(q)}$ is the operation of shifting vector s_b by q units and $||\cdot||$ is the l_2 norm[13]. For a fixed q, the exact solution for α, obtained by computing the minimum of $dist$, is: $\alpha = \frac{s_a^T s_{b(q)}}{||s_b||}$. However, there is no simple way to compute q. Thus, in our implementation of KSC[14], we search for the optimal value of q considering all integers in the range $(-n, n)$[15]. We refer the reader to the original paper for more details [21].

Like K-means, KSC requires the choice of a number k of clusters. We chose this number primarily based on the β_{CV} clustering quality metric. The β_{CV} is the ratio of the coefficient of variation (CV)[16] of the distances

[12]We note that the R^2 of the single-predictor model is equal to the square of the linear correlation between predictor and response.

[13]The l_2 norm of a vector s_k is $||s_k|| = \sqrt{\sum_{i=1}^{n} s_k^2(i)}$.

[14]We used of an open source implementation of the algorithm available at: http://github.com/flaviovdf/pyksc

[15]Shifts are performed in a rolling manner, where the element at the end of the vector returns to the beginning, to maintain the symmetric nature of $dist(s_a, s_b)$.

[16]CV is the ratio of standard deviation to the mean.

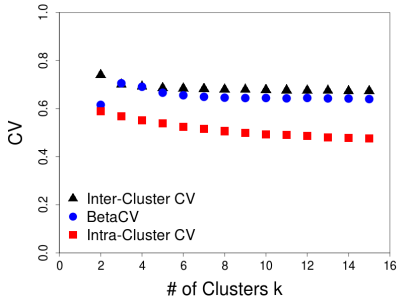


Figure 4: β_{CV} clustering quality metric (scholars with more than 20 years of experience).

between members of the same cluster (intra-cluster CV) to the CV of the distances between members of different clusters (inter-cluster CV). The β_{CV} should be computed for increasing values of k. When it 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 variability in the dataset.

5.2 Profile Characterization

For both experience groups analyzed, we found that the β_{CV} stabilizes for $k = 5$, as shown in Figure 4 for one group. This choice of number of clusters also agreed with other heuristics we employed, including visual inspection. Figure 5 shows the centroids of the 5 clusters identified for scholars in the (20, ∞) group. Each centroid corresponds to an "average" popularity curve for scholars in the cluster, and represents a different profile of popularity dynamics. Scales on both axes are omitted to emphasize the scale and time shifting invariants. The fraction of scholars in each cluster is provided in the caption of the figure. Interestingly, we found the same number of clusters k *and* very similar centroids also for scholars in the (15;20] group (omitted).

Before discussing the identified profiles, we note that they may not perfectly match the popularity curves of all scholars analyzed, as there might be variations within each cluster. Indeed, our goal is not to perfectly model the popularity evolution of all scholars, but rather capture the most prevalent trends, respecting time shift and volume invariants. To illustrate this point, Figure 6 shows the popularity time series of one example author in each cluster.

Centroids C0, C1 and C2, shown in Figures 5(a-c), correspond to profiles of scholars who succeed in becoming increasingly popular, acquiring more and more citations with time. These profiles account for 66% of all scholars. Note that the sharp decay in the last years might be just an artifact of recent publications having fewer citations. Also, recall that each centroid represents an "average" popularity curve for all scholars in the cluster. By manually inspecting the popularity curves for various scholars in these three clusters, we found that the decay at the end was not clear in several individual curves, although others did exhibit it.

In particular, we note that many of the scholars grouped in cluster C0 exhibit roughly stable popularity over time (with occasional peaks and decays), after an initial increase during their first years of activity (see example in Figure 6-a). Clusters C1 and C2, in turn, are more dominated by scholars whose popularity curves exhibit a more clear growth trend over time, differing basically in the growth rates. In contrast, profiles C3 and C4, shown in Figures 5(d-e), describe scholars who grow in popularity, experiencing a clear peak, but fail to remain popular afterwards. Once again, the main difference between these two profiles is on the growth and decay rates before and after the peak. Note that, in general, the centroids in Figure 5 approximate well the general *trends* of the individual curves in Figure 6.

We next characterize the scholars in each profile by analyzing the distributions of academic features computed for scholars in each profile. Since the popularity curves were extracted from the MS-AR dataset, and given the discrepancies observed in both MS-AR and AM datasets, we used only data from the MS-AR service to build these distributions. Specifically, we used the citation and publication time series collected for each scholar to compute the scholar's total citation count and total number of publications. The former accounts for the scholar popularity, as defined in this paper, and the latter is the most important academic feature (among those we analyzed) to explain scholar popularity (see Section 4.2).

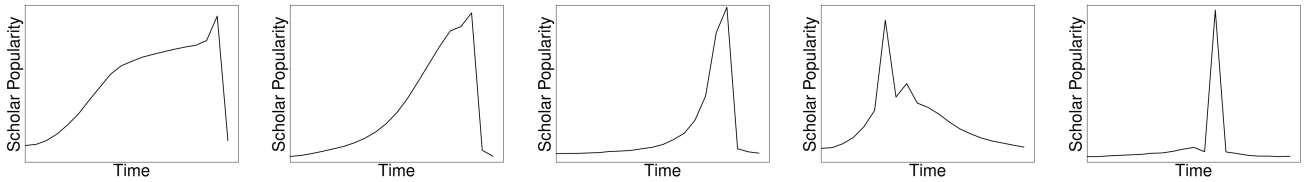
Figure 7 summarizes the distributions of popularity and number of publications of scholars in each profile with box plots. In each box plot, the central rectangle spans the first to the third quartiles, the segment inside is the median (second quartile), whereas whiskers above and below the rectangle represent the 9^{th} and 91^{th} percentiles. Each box plot also shows the mean value of the distribution. We focus the discussion on the (20;∞) group, though the conclusions hold for the other experience group as well.

In general, we find that the most popular scholars as well as those with the largest numbers of publications are in cluster $C1$. For both features, the median, mean, third quartile and 91^{th} percentile are larger for this cluster than for the others. C1 is also the largest cluster, with 27% of all analyzed popularity curves in the group. As mentioned, these are scholars who succeed in acquiring more and more popularity with time. For example, the mean popularity of these scholars is 535 citations, and the mean number of publications is 61. Moreover, for 9% of the authors in this cluster, the number of citations exceeds 1,410, and the number of publications surpasses 144 (91^{th} percentiles).

In contrast, scholars in cluster C2, who also exhibit a clear trend towards increasing popularity over time (as shown in Figure 5-c), have distributions much more skewed towards fewer citations and fewer publications. Note the concentration of both distributions around smaller values and the smaller span from the 9^{th} to 91^{th} percentiles, compared to the distributions for C1. For comparison purposes, the mean popularity in this cluster is only 143 citations, whereas the mean number of publications is 31.

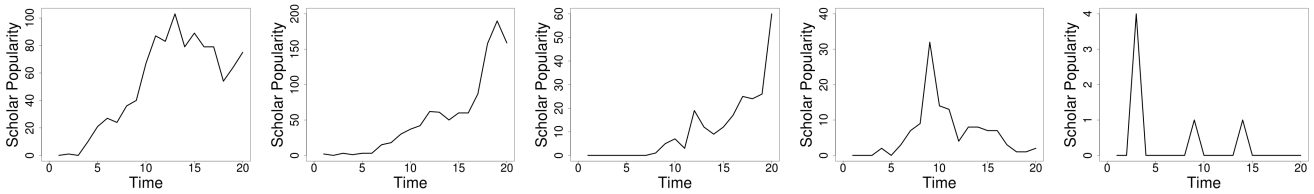
Cluster C0, which accounts for 23% of all scholars in the group, lie between both C1 and C2. The mean popularity and the mean number of publications are 393 and 38, respectively. Note, however, that some of these scholars do succeed in becoming very popular: the 9% most popular scholars in this cluster have at least 1,001 citations.

In contrast, clusters C3 and C4, which exhibit a clear decay in popularity after the peak, consists of scholars who tend to have far fewer papers and thus become much less popular than scholars in the other clusters. C3 and C4 also exhibit much less variability across scholars in terms of both features (note the smaller span between the 9^{th} and



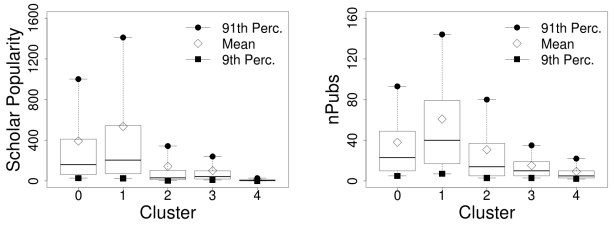
(a) C0 (23%) (b) C1 (27%) (c) C2 (16%) (d) C3 (19%) (e) C4 (15%)

Figure 5: Profiles of popularity dynamics (scholars with more than 20 years of experience): percentage of scholars in each profile shown in parentheses.



(a) Example in C0 (b) Example in C1 (c) Example in C2 (d) Example in C3 (e) Example in C4

Figure 6: Examples of popularity time series in each profile (scholars with more than 20 years of experience).



(a) Popularity (b) Number of publications

Figure 7: Distributions of popularity and number of publications for each profile (scholars with more than 20 years of experience).

91^{th} percentiles). In particular, scholars in C4 are the least popular ones in the set, and the popularity distribution for this cluster is very concentrated around the mean. Similarly, they tend to have very few publications. For example, the mean popularity and the mean number of publications of scholars in C3 are 101 and 15, respectively, whereas corresponding values for scholars in C4 are only 13 and 9.

One important observation can be drawn from these results. For most cases, scholars who do not publish frequently (captured by the number of publications), will likely attract only some attention over small time windows (i.e., their popularity will follow the trends of C3 and C4). Thus, such scholars's research will *likely* have little impact over time as measured by the number of citations. This result reflects the culture of "publish or perish", which serves as an incentive for scholars to continue publishing new research throughout their careers, and remain getting cited over time (as those in C0 and C1). While this is not a rule *per se*, since there are some scholars with very few publications who remain popular over time, it shows that, in a field like Computer Science, remaining active (publishing) over time is very important for popularity.

6. CONCLUSIONS AND FUTURE WORK

We have investigated the importance of various academic features to scholar popularity. Two large scholarly datasets were used to quantify the impact of features on total popularity (citation count), and uncover trends of popularity temporal dynamics. Our analyses showed that, even though most of the considered features are strongly correlated with popularity, only two features - number of publications and average quality of the scholar's publication venues – are needed to explain practically all the variation in popularity across different scholars. We also uncovered five profiles of scholar popularity dynamics. Three of them correspond to scholars who succeed in becoming increasingly popular with time, varying only in terms of the popularity growth rate, while the others correspond to scholars who fail to keep being cited after a popularity peak. Our results also suggest that scholars who succeed in getting cited over time most likely are those who remain publishing through their careers.

Future work includes comparing the popularity dynamics of scholars in different research fields and different countries, extending our study to other popularity metrics, and developing scholar popularity prediction methods.

Acknowledgements

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant Number 573871/2008-6), by CNPq, CAPES, and FAPEMIG.

7. REFERENCES

[1] Y. Ding and B. Cronin, "Popular and/or Prestigious? Measures of Scholarly Esteem," *Information Processing & Management*, vol. 47, no. 1, pp. 80–96, 2011.

[2] M. A. Brandão, M. M. Moro, G. R. Lopes, and J. P. Oliveira, "Using Link Semantics to Recommend

Collaborations in Academic Social Networks," in *International Conference on World Wide Web Companion*, 2013, pp. 833–840.

[3] H. Cason and M. Lubotsky, "The Influence and Dependence of Psychological Journals on each Other," *Psychological Bulletin*, vol. 33, no. 2, pp. 19–103, 1936.

[4] G. Pinski and F. Narin, "Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics," *Information Processing & Management*, vol. 12, no. 5, pp. 297–312, 1976.

[5] N. Ma, J. Guan, and Y. Zhao, "Bringing PageRank to the Citation Analysis," *Information Processing & Management*, vol. 44, no. 2, pp. 800–810, 2008.

[6] E. Yan and Y. Ding, "Weighted Citation: An Indicator of an Article's Prestige," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 8, pp. 1635–1643, 2010.

[7] T. Brody, S. Harnad, and L. Carr, "Earlier Web Usage Statistics As Predictors of Later Citation Impact: Research Articles," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 8, pp. 1060–1072, 2006.

[8] C. Castillo, D. Donato, and A. Gionis, "Estimating Number of Citations Using Author Reputation," in *International Conference on String Processing and Information Retrieval*, 2007, pp. 107–117.

[9] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, "To Better Stand on the Shoulder of Giants," in *Joint Conference on Digital Libraries*, 2012, pp. 51–60.

[10] J. Bollen, M. A. Rodriquez, and H. V. de Sompel, "Journal Status," *Scientometrics*, vol. 69, no. 2, pp. 669–687, 2006.

[11] A. Fersht, "The Most Influential Journals: Impact Factor and Eigenfactor," *National Academy of Sciences*, vol. 106, no. 17, pp. 6883–6884, 2009.

[12] J. E. Hirsch, "An Index to Quantify an Individual's Scientific Research Output," *National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[13] L. Egghe, "Theory and Practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.

[14] M. Schreiber, "To Share the Fame in a Fair Way, h m Modifies h for Multi-authored Manuscripts," *New Journal of Physics*, vol. 10, no. 4, p. 040201, 2008.

[15] E. Yan and Y. Ding, "Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 10, pp. 2107–2118, 2009.

[16] P. Riikonen and M. Vihinen, "National Research Contributions: A Case Study on Finnish Biomedical Research," *Scientometrics*, vol. 77, no. 2, pp. 253–280, 2008.

[17] T. F. Frandsen and J. Nicolaisen, "Effects of Academic Experience and Prestige on Researchers' Citing Behavior," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 64–71, 2012.

[18] E. Yan and Y. Ding, "Discovering Author Impact: A PageRank Perspective," *Information Processing & Management*, vol. 47, no. 1, pp. 125–134, 2011.

[19] E. Garfield, "Journal Impact Factor: a Brief Review," *Canadian Medical Association Journal*, vol. 161, no. 8, pp. 979–980, 1999.

[20] L. Leydesdorff, "How are New Citation-based Journal Indicators Adding to the Bibliometric Toolbox?" *Journal of the American Society for Information Science and Technology*, vol. 60, no. 7, pp. 1327–1336, 2009.

[21] J. Yang and J. Leskovec, "Patterns of Temporal Variation in Online Media," in *International Conference on Web Search and Data Mining*, 2011, pp. 177–186.

[22] S. Bergsma, R. L. Mandryk, and G. McCalla, "Learning to Measure Influence in a Scientific Social Network," *Lecture Notes in Computer Science*, vol. 8436, pp. 35–46, 2014.

[23] H. Lima, T. H. Silva, M. M. Moro, R. L. Santos, W. Meira, Jr., and A. H. Laender, "Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas," in *Joint Conference on Digital Libraries*, 2013, pp. 97–106.

[24] D. E. Acuna, S. Allesina, and K. P. Kording, "Future Impact: Predicting Scientific Success," *Nature*, vol. 489, no. 7415, pp. 201–202, 2012.

[25] M. Newman, "The Structure of Scientific Collaboration Networks," *National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[26] X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel, "Co-authorship Networks in the Digital Library Research Community," *Information Processing & Management*, vol. 41, no. 6, pp. 1462–1480, 2005.

[27] B. Keegan, D. B. Horn, T. A. Finholt, and J. Kaye, "Structure and Dynamics of Coauthorship, Citation, and Impact within CSCW," *CoRR*, vol. 1307.7172, 2013.

[28] T. Martin, B. Ball, B. Karrer, and M. Newman, "Coauthorship and Citation in Scientific Publishing," *CoRR*, vol. 1304.0473, 2013.

[29] B. Cronin and L. I. Meho, "Timelines of Creativity: a Study of Intellectual Innovators in Information Science," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 1948–1959, 2007.

[30] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and Mining of Academic Social Networks," in *International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 990–998.

[31] S. Roy, M. Cock, V. Mandava, S. Savanna, B. Dalessandro, C. Perlich, W. Cukierski, and B. Hamner, "The Microsoft Academic Search Dataset and KDD Cup 2013," in *Knowledge Discovery and Data Mining Cup Workshop*, 2013.

[32] R. K. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley & Sons, 1991.

[33] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using Early View Patterns to Predict the Popularity of Youtube Videos," in *International Conference on Web Search and Data Mining*, 2013, pp. 365–374.