Sarigöl et al. *EPJ Data Science* 2014, **2014**:9
http://www.epjdatascience.com/content/2014/1/9

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

# Predicting scientific success based on coauthorship networks

Emre Sarigöl, René Pfitzner*, Ingo Scholtes, Antonios Garas and Frank Schweitzer

*Correspondence: rpfitzner@ethz.ch
Chair of Systems Design, ETH Zurich, Weinbergstrasse 56/58, Zurich, 8004, Switzerland

**Abstract**

We address the question to what extent the success of scientific articles is due to social influence. Analyzing a data set of over 100,000 publications from the field of Computer Science, we study how centrality in the coauthorship network differs between authors who have highly cited papers and those who do not. We further show that a Machine Learning classifier, based only on coauthorship network centrality metrics measured at the time of publication, is able to predict with high precision whether an article will be highly cited five years after publication. By this we provide quantitative insight into the social dimension of scientific publishing – challenging the perception of citations as an objective, socially unbiased measure of scientific success.

**Keywords:** scientometrics; complex networks

## 1 Introduction

Quantitative measures are increasingly used to evaluate the performance of research institutions, departments, and individual scientists. Measures like the absolute or relative number of published research articles are frequently applied to quantify the *productivity* of scientists. To measure the *impact* of research, citation-based measures like the total number of citations, the number of citations per published article or the *h*-index [1], have been proposed. Proponents of such citation-based measures or rankings argue that they allow to quantitatively and objectively assess the *quality* of research, thus encouraging their use as simple proxies for the *success* of scientists, institutions or even whole research fields. The intriguing idea that by means of citation metrics the task of assessing research quality can be "outsourced" to the *collective intelligence* of the scientific community, has resulted in citation-based measures becoming increasingly popular among research administrations and governmental decision makers. As a result, such measures are used as one criterion in the evaluation of grant proposals and research institutes or in hiring committees for faculty positions. Considering the potential impact for the careers of – especially young – scientists, it is reasonable to take a step back and ask a simple question: To what extent do *social factors* influence the number of citations of their articles? Arguably, this question challenges the perception of science as a systematic pursuit for objective truth, which ideally should be free of personal beliefs, biases or social influence. On the other hand, quoting Werner Heisenberg [2], "*science is done by humans*", it would be surprising if specifically scientific activities were free from the influences of social aspects.

Whereas often the term "social influence" has a negative connotation, we don't think that social influence in science necessarily stems from malicious or unethical behavior, like e.g. nepotism, prejudicial judgments, discrimination or in-group favoritism. We rather suspect that, as a response to the increasing amount of published research articles and our limited ability to keep track of potentially relevant works, a growing importance of social factors in citation behavior is due to natural mechanisms of *social cognition* and *social information filtering*.

In this paper we address this issue by studying the influence of social structures on scholarly citation behavior. Using a data set comprising more than 100,000 scholarly publications by more than 160,000 authors, we extract time-evolving coauthorship networks and utilize them as a (simple) proxy for the evolving social (collaboration) network of the scientific discipline *computer science*. Based on the assumption that the centrality of scientists in the coauthorship network is indicative for the *visibility* of their work, we then study to what extent the "*success*" of research articles in terms of citations can be predicted using only knowledge about the embedding of authors in the coauthorship network *at the time of publication*. Our prediction method is based on a random forest classifier and utilizes a set of complementary network centrality measures. We find strong evidence for our hypothesis that authors whose papers are highly cited in the future have – on average – a significantly higher centrality in the coauthorship network at the time of publication. Remarkably, we are able to predict whether an article will belong to the 10% most cited articles with a precision of 60%. We argue that this result quantifies the existence of a *social bias*, manifesting itself in terms of visibility and attention, and influencing measurable citation "success" of researchers. The presence of such a social bias threatens the interpretation of citations as *objectively awarded esteem*, which is the justification for using citation-based measures as universal proxies of *quality* and *success*.

The remainder of this article is structured as follows: In Section 2 we review a number of works that have studied scientific collaboration structures as well as their relation to citation behavior. In Section 3 we describe our data set and provide details of how we construct time-evolving coauthorship networks. We further introduce a set of network-theoretical measures which we utilize to quantitatively assess the centrality and embedding of authors in the evolving coauthorship network. In Section 4 we introduce a number of hypotheses about the relations between the position of authors in the coauthorship network and the future success of their publications. We test these hypotheses and obtain a set of candidate measures which are the basis for our prediction method described in Section 5. We summarize and interpret our findings in Section 6 and discuss their implications for the application of citation-based measures in the quantitative assessment of research.

## 2 The complex character of citations

It is remarkable that, even though citation-based measures have been used to quantify research impact since almost sixty years [3], a complete *theory of citations* is still missing. In particular, researchers studying the social processes of science have long been arguing that citations have different, complex functions that go well beyond a mere attribution of credit [4]. For example, in [5] evidence was presented that papers, which have been publicly criticized via formal, published comments, are often highly cited. Furthermore, at the level of scientific articles, a citation can be interpreted as a "discursive relation", while at the level of authors citations have an additional meaning as expression of "professional relations" [4]. Additional interpretations have been identified at aggregate levels, like e.g. social groups,

institutions, scientific communities or even countries citing each other. These findings suggest that citations are indeed a complex phenomenon which have both cognitive and a social dimension [4, 6]. The complex character of scholarly citations was further emphasized recently [7, 8]. Here, the authors argue that, apart from an attribution of scientific merit, references in scientific literature often serve as a tool to guide and orient the reader, to simplify scientific writing and to associate the work with a particular scientific community. Furthermore, they highlight that citation numbers of articles are crucially influenced not only by the popularity of a research topic and the size of the scientific community, but also by the number of authors as well as their prominence and visibility. These findings question an oversimplified interpretation of citation counts as objective quality indicator.

Facilitated by the wide-spread availability of scholarly citation databases, some advances in the understanding of the dynamics of citations have been made in the last years. For an interesting study of bibliometric indicators on the author level, see e.g. [9]. Generally, citation practices seem to differ significantly across different scientific disciplines, which complicates the definition of universal citation-based impact measures. However, the remarkable finding that – independent of discipline – citations follow a log-normal distribution which can be rescaled in such a way that citation numbers become comparable [10, 11], suggests that the mechanisms behind citation practices are universal across disciplines, and differences are mainly due to differing community sizes.

Additionally to investigations of the differences across scientific communities, the relations between citations and coauthorships were studied in recent works. Using data from a number of scientific journals, it was shown that the citation count of an article is correlated both with the number of authors and the number of institutions involved in its production [12, 13]. Studying data from eight highly ranked scientific journals, it was shown [14] that (a) single author publications consistently received the lowest number of citations and (b) publications with less than five coauthors received less citations than the average article. Studying citations between individuals rather than articles, in [15] it was observed that coauthors tend to cite each other sooner after the publication of a paper (compared to non-coauthors). Further, the authors showed that a strong tendency towards reciprocal citation patterns exists. These findings already indicate that social aspects influence citing behavior. In this work we are going to quantitatively reveal the extent of this influence.

Going beyond a mere study of direct coauthorship relations, first attempts to study *both* citation and coauthorship structures from a *network perspective* have been made recently. Aiming at a measure that captures both the *amount* as well as the *reach* of citations in a scientific community, a citation index that incorporates the distance of citing authors in the collaboration network was proposed [16]. Another recent study [17] used the topological distance between citing authors in the coauthorship network to extend the notion of self-citations. Interestingly, apart from direct self-citations, this study could not find a strong tendency to cite authors that are close in the coauthorship network.

Different from previous works, in this article we study correlations between the *centrality* of authors in collaboration networks and the *citation success* of their research articles. By this we particularly extend previous works that use a network perspective on coauthorship structures and citation patterns. Stressing the fact that *social relations* of authors play an important role for how much attention and recognition their research receives, we further contribute a quantitative view on previously hypothesized relations between the *visibility* of authors and citation patterns.

## 3 Time-evolving collaboration and citation networks

In this work we analyze a data set of scholarly citations and collaborations obtained from the Microsoft Academic Search (MSAS, http://academic.research.microsoft.com) service. The MSAS is a scholarly database containing more than 35 Million publication records from 15 scientific disciplines. Using the Application Programming Interface (API) of this service, we extracted a subset of more than 100,000 computer science articles, published between 1996 and 2008, in the following way: First, we retrieved unique numerical identifiers (IDs) of the 20,000 highest ranked authors in the field of *computer science*. This ranking is the result of an MSAS internal "field rating", taking into account several scholarly metrics of an author (number of publications, citations, $h$-index) and comparing them to the typical values of these metrics within a certain research field. In order to build coauthorship and citation networks of reasonable size, in a second step we chose 1,000 authors i.i.d. uniformly from the set of these 20,000 authors. In the third step, we obtained information on coauthors, publication date, as well as the list and publication date of citing works for all the publications authored by these 1,000 authors between 1996 and 2008. This results in a data set consisting of a total of 108,758 publications from the field of computer science, coauthored by a total of 160,891 researchers. Each publication record contains a list of author IDs, which, by means of disambiguation heuristics internally applied by the MSAS service, uniquely identify authors independent of name spelling variations. The absence of name ambiguities is one feature that sets this data set apart from other data sets on scholarly publications that are used frequently. Based on this data set we extracted a *coauthorship network*, where nodes represent authors and links represent coauthorship relations between authors. In addition, using the information about citing papers, we extracted *citation dynamics*, i.e. the time evolution of the number of citations of all publications in our data set. Similar to earlier works, we argue that the coauthorship network can be considered a first-order approximation of the complete scientific collaboration network [15]. Based on the publication date of an article, we additionally assign time stamps to the extracted coauthor links – thus obtaining time-evolving coauthorship networks.

We analyze the evolution of the coauthorship network using a sliding window of two years in which we aggregate all coauthorships occurring within that time. Starting with 1996, we slide this window in one year increments and obtain a total of 11 time slices representing the evolution of collaboration structures between 1996 and 2008. We use an extended time-window of two years to account for the continuing effect of a coauthorship in terms of awareness about the coauthors works. Although larger time windows are certainly possible (and their effects interesting to investigate), in this work we are less concerned with the optimal time-window size and consistently use the above described approach. However, consistency checks performed with varying time-window sizes suggest that our results are robust.

Table 1 summarizes the number of nodes and links in the coauthorship network, the number of publications in each time slice as well as the fractional size of the largest connected component (LCC). Note that the time-aggregated network (nearly) forms one giant connected component with only a minor fraction of isolated nodes. In contrast, some of the time slices fall apart in several larger disconnected components. Note also that the size of the largest connected component is increasing with time, which may indicate either a possible bias in the coverage of the MSAS database to favor newer articles, or an

**Table 1 Number of papers and size of the collaboration network 2-year subgraphs between 1995-2008 used in our study.**

| Year | LCC fraction | Links | Nodes | Publications |
|---|---|---|---|---|
| 1996-1997 | 0.18 | 61,046 | 2,845 | 1,160 |
| 1997-1998 | 0.37 | 130,938 | 6,381 | 3,070 |
| 1998-1999 | 0.45 | 153,412 | 8,470 | 4,054 |
| 1999-2000 | 0.50 | 186,318 | 10,413 | 5,320 |
| 2000-2001 | 0.60 | 358,188 | 13,451 | 6,561 |
| 2001-2002 | 0.63 | 413,846 | 15,309 | 7,026 |
| 2002-2003 | 0.74 | 542,912 | 20,238 | 9,193 |
| 2003-2004 | 0.77 | 653,224 | 23,624 | 10,608 |
| 2004-2005 | 0.79 | 745,352 | 26,258 | 11,430 |
| 2005-2006 | 0.83 | 889,996 | 29,886 | 12,919 |
| 2006-2007 | 0.84 | 914,614 | 32,412 | 13,568 |
| 2007-2008 | 0.86 | 858,554 | 35,255 | 14,214 |
| Overall | 0.99 | 5,324,330 | 160,891 | 108,758 |

increase of "collaborativeness" in science. As we are going to perform a social network analysis of the coauthorship time slices – and some measures (like eigenvector centrality) are not well-defined for unconnected graphs – we limit our following analysis on the largest connected component. For each network corresponding to one two-year time slice, we compute a number of node-level metrics that allow us to quantitatively monitor the evolution of network positions for all authors. In particular, we compute *degree centrality*, *eigenvector centrality*, *betweenness centrality* and *k-core centrality* of authors. For further details on the centrality measures used in this study, we refer the reader to the Supplementary Material (Additional file 1) or a standard network analysis textbook, e.g. [18]. Here we utilize implementations of these measures provided by the igraph package [19].

A major focus of our work is to assess the predictive power of an author's position in the coauthorship network for the citation success of her future articles. To do so we adopt a so-called *hindcasting approach*: For each publication $p$ published in year $t$, we extract the list of coauthors as well as the LCC of the coauthorship network in the time slice $[t-2, t]$, and calculate the centrality measures. Based on the citation data, we furthermore calculate the number of citations $c_p$ paper $p$ gained within a time frame of *five years* after publication, i.e. in the time slice $[t, t+5]$.

In particular, we are interested in those publications that are among the most successful ones. Defining *success* is generally an ambiguous endeavor. As justified in the Introduction, here we take the (controversial) viewpoint that success is directly measurable in number of citations. We specifically focus on a simple notion of success in terms of having *highly cited papers* and, similar to [20], assume that a paper is *successful* if five years after publication it has more citations than 90% of all papers published in the same year. We refer to the set of successful papers published in year $t$ as $P_{\uparrow}(t)$. The set of remaining papers, i.e. those published at time $t$ that are cited less frequently than the top 10%, is denoted as $P_{\downarrow}(t)$.

## 4 Statistical dependence of coauthorship structures and citations

Having a large social network and "knowing the right people" often is a prerequisite for career success. However, science is often thought to be one of the few fields of human endeavor where success depends on the quality of an authors' work, rather than on her social connectedness. Given the time evolving coauthorship network, as well as the observed success (or lack thereof) of a publication, we investigate two research questions, aiming to quantify the aspect of social influence on citation success. First, we examine

whether there is a general tendency of central authors in the coauthorship network to publish papers that are more successful than those of non-central authors. Second, we investigate the inverse effect and ask whether the success of a paper influences the future coauthorship centrality of its authors.

### 4.1 Effects of author centrality on citation success

To answer the first research question we test the following hypothesis.

**Hypothesis 1** *At the time of publication, authors of papers in $P_{\uparrow}(t)$ are more central in the coauthorship network than authors of articles in $P_{\downarrow}(t)$.*

As papers often have more than one author, for each paper we only consider the coauthorship network centralities of the author with the highest coauthorship degree, and refer to this as the *coauthorship centrality of the paper*. This choice is motivated by the intuition that the centrality of the best connected coauthor should provide the major amount of (socially triggered) visibility for the publication. One might argue that this procedure introduces a centrality bias towards papers with a large number of authors. However, as the number of coauthors in our dataset is rather narrow with a mean of 3.95, a median of 3 and a standard deviation of 5.41 authors, a seizable bias cannot be expected. We test Hypothesis 1 by comparing coauthorship centrality distributions of papers in $P_{\uparrow}(t)$ and $P_{\downarrow}(t)$ for each year $t$. In order to compare the centrality distributions, we apply a *Wilcoxon-Mann-Whitney two-sample rank-sum-test* [21]. For each of the four centrality metrics we test the null hypothesis that coauthorship centrality distributions of papers in $P_{\uparrow}(t)$ and $P_{\downarrow}(t)$ are the same against the alternative hypothesis that the centrality distribution of papers in $P_{\uparrow}(t)$ is stochastically larger than that of papers in $P_{\downarrow}(t)$. The $p$-values of the tests as well as the corresponding averages and variances of the four considered centrality metrics in the two sets are shown in Table 2. Additionally, Figures 1, 2, 3 and 4 show kernel density estimates of these distributions. For all considered centrality metrics $p$-values are well below a significance level of 0.01. We can thus safely reject the null hypothesis, concluding that coauthorship centrality metrics of papers in $P_{\uparrow}(t)$ are stochastically larger than those of papers in $P_{\downarrow}(t)$. This result indicates that centrality metrics in the coauthorship network, at the time of publication of a paper, are indicative for future paper success. Note however, that this statistical dependency is more complicated than the linear Pearson or the more general Spearman correlation. Indeed, all the considered social network metrics are only weakly, if at all, correlated with citation numbers (see Supplementary Material (Additional file 1)). Table 3 summarizes to what extent citation success and coauthorship network centrality are statistically dependent. The left entry of each cell indicates

**Table 2 $p$-values of one sided Wilcoxon-Mann-Whitney test. This quantifies whether the centrality distributions of authors of articles in $P_{\uparrow}$ are (in a statistical sense) larger than those of authors of articles in $P_{\downarrow}$. Also shown are the medians $M$ and variances var of the centrality metrics in the two sets.**

| | $p$-value | $M(P_{\downarrow})$ | $M(P_{\uparrow})$ | $\mathrm{var}P_{\downarrow}$ | $\mathrm{var}P_{\uparrow}$ |
|---|---|---|---|---|---|
| $k$-core | $1.28 \times 10^{-115}$ | 16 | 26 | $1.20 \times 10^{4}$ | $7.18 \times 10^{3}$ |
| Eigenvector | $2.52 \times 10^{-34}$ | $9.67 \times 10^{-18}$ | $2.08 \times 10^{-17}$ | $2.58 \times 10^{-3}$ | $5.40 \times 10^{-4}$ |
| Betweenness | $1.19 \times 10^{-68}$ | 19.38 | $11.4 \times 10^{4}$ | $4.19 \times 10^{12}$ | $1.58 \times 10^{13}$ |
| Degree | $5.63 \times 10^{-125}$ | 28 | 57 | $1.02 \times 10^{5}$ | $1.13 \times 10^{5}$ |

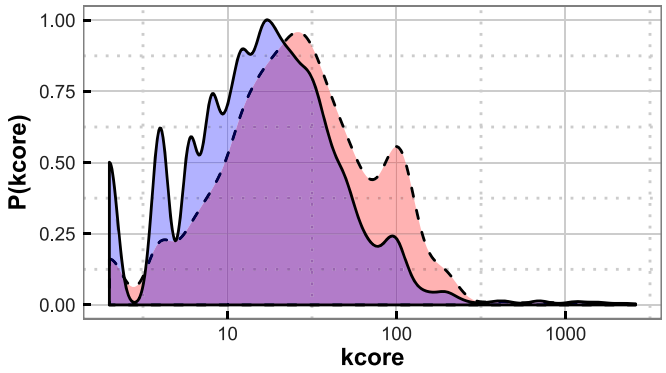


**Figure 1 *k*-core centrality shift.** Kernel density estimate of *k*-core coauthorship centrality of papers in $P_{\downarrow}$ (solid blue) and $P_{\uparrow}$ (dashed red).

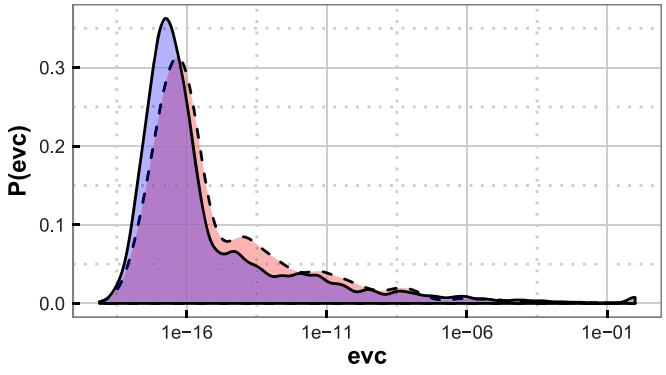


**Figure 2 Eigenvector centrality shift.** Kernel density estimate of eigenvector coauthorship centrality of papers in $P_{\downarrow}$ (solid blue) and $P_{\uparrow}$ (dashed red).

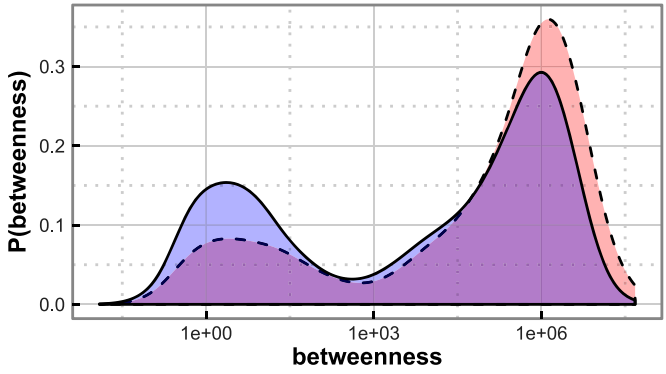


**Figure 3 Betweenness centrality shift.** Kernel density estimate of betweenness coauthorship centrality of papers in $P_{\downarrow}$ (solid blue) and $P_{\uparrow}$ (dashed red).

$P$(toppaper|topmetric), i.e. the fraction of papers belonging to the top $x$% successful papers, given that their authors have top $x$% centrality metrics. The right entry of each cell indicates $P$(topmetric|toppaper), i.e. the fraction of papers that have authors which are within the set of authors with top $x$% centrality metrics, given that the papers are within

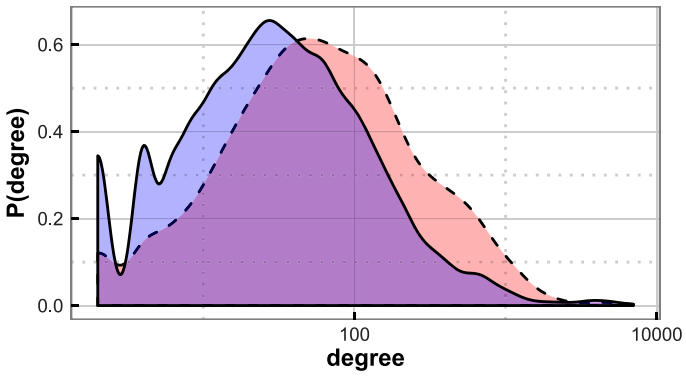


**Figure 4 Degree centrality shift.** Kernel density estimate of degree coauthorship centrality of papers in $P_{\downarrow}$ (solid blue) and $P_{\uparrow}$ (dashed red).

**Table 3 The left entry of each cell indicates *P*(toppaper|topmetric), i.e. the fraction of papers belonging to the top *x*% successful papers, given that their authors have top *x*% centrality metrics. The right entry of each cell indicates *P*(topmetric|toppaper), i.e. the fraction of papers that have authors which are within the set of authors with top *x*% centrality metrics, given that the papers are within the top *x*% successful papers. Row *Intersection* indicates the intersection of all the above considered centrality metrics.**

| | Top 10% | Top 5% | Top 2% | Top 1% |
|---|---|---|---|---|
| *k*-core | 0.22\|\|0.21 | 0.17\|\|0.16 | 0.07\|\|0.07 | 0.01\|\|0.01 |
| Eigenvector | 0.11\|\|0.11 | 0.06\|\|0.06 | 0.01\|\|0.01 | 0.01\|\|0.01 |
| Betweenness | 0.20\|\|0.20 | 0.13\|\|0.13 | 0.11\|\|0.11 | 0.11\|\|0.11 |
| Degree | 0.20\|\|0.20 | 0.15\|\|0.15 | 0.10\|\|0.09 | 0.07\|\|0.07 |
| Intersection | 0.36\|\|0.15 | 0.27\|\|0.11 | 0.17\|\|0.06 | 0.12\|\|0.04 |
| # papers | 3,700 | 1,844 | 730 | 362 |

the top *x*% successful papers. From these results, we conclude two observations: First, the probabilities in each cell are well below 1, indicating the absence of a simple linear (Pearson) correlation. Second, especially considering *k*-core centrality, knowing a paper is top 10% successful, the conditional probability that it was written by an author with top 10% *k*-core centrality, is $P(\text{topmetric}|\text{toppaper}) = 0.21$. Additionally, Table 3 shows that vice versa $P(\text{toppaper}|\text{topmetric}) = 0.22$ of all papers that are published by authors with top 10% *k*-core centrality, are among the most successful ones. In addition, we consider the intersection of all four centrality metrics. Here we even find that $P(\text{toppaper}|\text{topmetric}) = 0.36$ of all papers published authors with top 10% centrality w.r.t. *all four* centrality metrics, are among the top 10% most cited papers. We will use this observation as basis for a naive Bayes classifier in Section 5.

### 4.2 Coevolution of coauthorship and citation success

In the previous section we studied the question whether the centrality of authors in the coauthorship network is indicative for the success of publications in terms of citations. Our results suggested that centrality in coauthorship networks is indeed indicative for citation success. In the following we study the inverse relation and ask whether a shift in citation success of authors is indicative for their future position in the coauthorship network. To answer this question, we consider all authors who published an article both at time $t$ and five years later at $t+5$. We then categorize them based on the citation success

of their articles published at time $t$ and time $t+5$. We introduce two sets of authors: Set $A_{\searrow}(t)$ is the set of authors who at time $t$ had at least one publication in class $P_{\uparrow}(t)$, but who at time $t+5$ did not have an article in class $P_{\uparrow}(t+5)$ anymore. Set $A_{\nearrow}(t)$ contains all authors who at time $t$ had no article in class $P_{\uparrow}(t)$ but who at time $t+5$ published at least one article that falls in class $P_{\uparrow}(t+5)$. In addition, we record the coauthorship centralities of authors in these two sets for two time windows $[t-2,2]$ and $[t+3,t+5]$. For authors in set $A_{\nearrow}$ we test the following hypothesis:

**Hypothesis 2** *Authors that experience a positive shift in their citation success* (*i.e. authors in* $A_{\nearrow}$) *will become more central in the coauthorship network.*

Complementary to Hypothesis 2, for authors in set $A_{\searrow}$ we hypothesize:

**Hypothesis 3** *Authors that experience a negative shift in their citation success* (*i.e. authors in* $A_{\searrow}$) *will become less central in the coauthorship network.*

In order to test for Hypothesis 2 and Hypothesis 3, we apply a *pairwise Wilcoxon-Mann-Whitney* test. To verify Hypothesis 2 we test if the centralities of authors have decreased in the case of a decrease in publication success from time $t$ to $t+5$. To verify Hypothesis 3 we test if the centralities of authors have increased in the case of an increase in publication success from time $t$ to $t+5$. Results of these hypotheses tests are presented in Table 4. Testing Hypothesis 2, for authors in $A_{\nearrow}$ we observe that $p$-values are much lower than the 0.01 significance threshold. We hence find evidence that authors in $A_{\nearrow}$ experience a significant *increase* in $k$-core, betweenness and degree centrality. Reversely, results for authors in $A_{\searrow}$ suggest a significant *decrease* in $k$-core, eigenvector and degree centrality. Based on these results we cannot reject Hypothesis 2 and Hypothesis 3, indicating that citation success significantly influences the future centrality of authors in the coauthorship network.

As an illustration of citation and coauthorship dynamics, Figures 5 and 6 show part of the coauthorship network. Color intensity of the nodes is scaled to their degree centrality, while node size is scaled to their betweenness centrality. A very strong community structure is clearly visible. Furthermore, we highlighted in red one particular author from the set $A_{\nearrow}(t)$ (2002), i.e. an author who did not have a paper in $P_{\uparrow}$ in 2002, but did so in 2007. In the considered five year span the highlighted author moved from a position in the periphery of the coauthorship network to a position in the center. Not only did the authors'

**Table 4 *p*-values of Wilcoxon-Mann-Whitney test for different centrality metrics and alternative hypotheses. Column $A_{\searrow}$ presents *p*-values for authors in set $A_{\searrow}$, column $A_{\nearrow}$ presents *p*-values for authors in set $A_{\nearrow}$.**

| Centrality measure & alternative | $A_{\searrow}$ | $A_{\nearrow}$ |
|---|---|---|
| $k$-core($t$) > $k$-core($t+5$) | $3.15 \times 10^{-11}$ | 1 |
| $k$-core($t$) < $k$-core($t+5$) | 1 | $3.04 \times 10^{-55}$ |
| ev-centr($t$) > ev-centr($t+5$) | $5.18 \times 10^{-14}$ | 0.86 |
| ev-centr($t$) < ev-centr($t+5$) | 1 | 0.14 |
| bw-centr($t$) > bw-centr($t+5$) | 0.23 | 1 |
| bw-centr($t$) < bw-centr($t+5$) | 0.77 | $7.29 \times 10^{-30}$ |
| degree($t$) > degree($t+5$) | $6.69 \times 10^{-11}$ | 1 |
| degree($t$) < degree($t+5$) | 1 | $7.72 \times 10^{-62}$ |
| # authors | 521 | 648 |

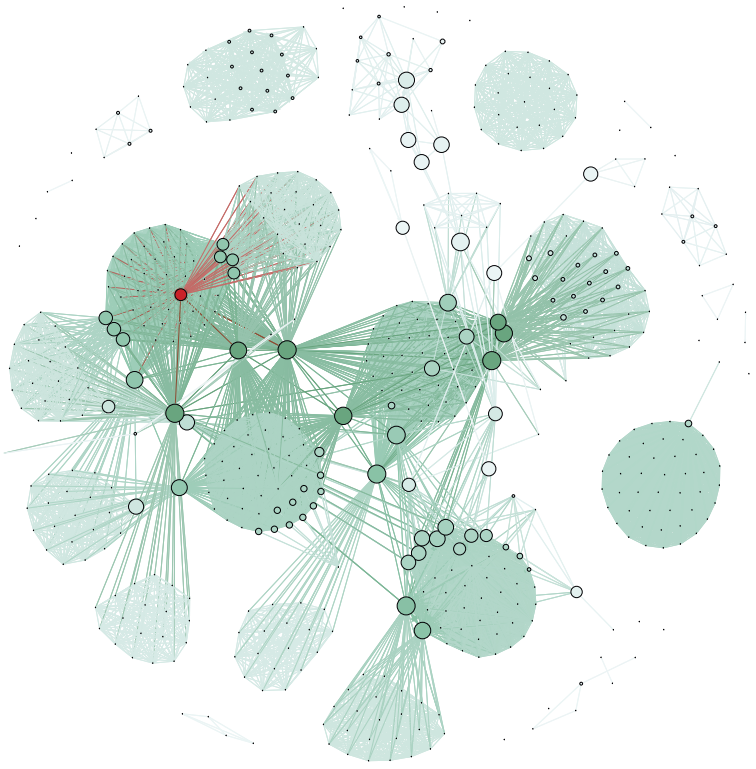

**Figure 5 Citation success 2002.** Illustration of correlation between citation success and centrality in the coauthorship network for year 2002. Color intensity of the nodes is scaled according to their degree centrality and size of nodes is scaled according to their betweenness centrality. Red node highlights one and the same author as in Figure 6.

degree centrality increase (see size of the node as well as joined red-colored links), but also the author's betweenness centrality largely increased.

Note that already in 2002 the author had comparatively high betweenness and degree centrality, which – according to our previous discussion – provided an ideal starting point for citation success in 2007.

## 5 Predicting successful publications

In the previous sections we presented evidence for the existence of statistical dependencies between authors' coauthorship centrality and the success of their publications. Results suggest that several coauthorship centrality metrics are indicative for citation success. However, we did not identify one single centrality metric whose magnitude is sufficient to predict whether the paper will become highly cited. In particular, we did not find that this would be true for the mere number of coauthors. Instead, we can guess that importance in the collaboration network is *multi-faceted* and thus influences by more than one network measure. In this section we thus develop a Machine Learning classifier which – taking into account several features of the authors position in the coauthorship network – is able to predict whether a publication will be highly cited.

Previous works have already attempted to predict citation success. For example in [22], the predictive power of the past $h$-index for the future $h$-index of a scientist was presented. Furthermore, in [23] additional indicators like, e.g. the length of the career or the

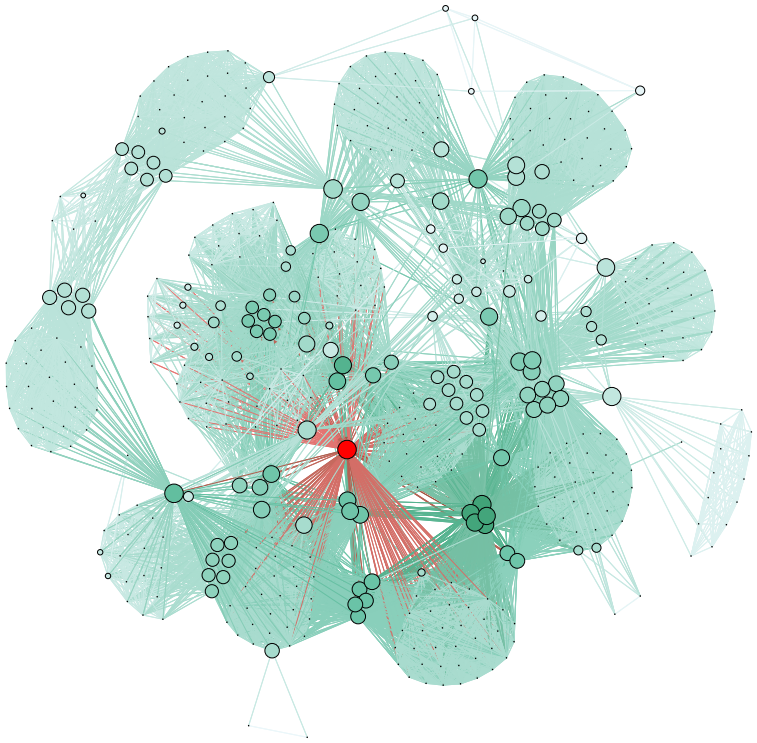

**Figure 6 Citation success 2007.** Illustration of correlation between citation success and centrality in the coauthorship network for year 2007. Color intensity of the nodes is scaled according to their degree centrality and size of nodes is scaled according to their betweenness centrality. Red node highlights one and the same author as in Figure 5.

number of articles in certain journals, have been integrated into a model to predict the future $h$-index of scientists. The authors of [20] compare the number of citations an article has received at a given point in time with the expected value in a preferential attachment model for the citation network. Deriving a $z$-score, the authors present a prediction of which papers will be highly cited in the future. Recently the authors reevaluate their earlier predictions and confirm the predictive power of their approach [24]. Whereas these three approaches attempt to predict success based on past citation dynamics, they do not investigate the underlying mechanisms that lead to citation success. Here we address this fundamental question and try to predict citation success merely based on centrality measures of authors in the coauthorship network. Clearly, many different factors will contribute to scientific success. In this work, however, we focus on the social component (based on the coauthorship network) in order to highlight the influence of social, and not necessarily merit-based, mechanisms on publication success.

In Section 4.1 we presented insights about the statistical dependency of citation success and several social network centrality measures (see Table 3). These results suggest that a naive Bayes predictor for citation success can already yield quite useful results, predicting whether or not a paper will be toppaper, given ex ante knowledge about topmetric of the authors. Using $k$-core centrality as a basis, we apply the following classification rule:

> *If a paper is authored by a top* 10% *k-core centrality author, then the paper will be among the top* 10% *most cited papers five years after publication.*

To evaluate the goodness of this prediction, we consider the error measures precision and recall (see Additional file 1 for a general definition of precision and recall). Observing that for $k$-core centrality in a 10% success scenario it is $P(\mathsf{topmetric}|\mathsf{toppaper}) = 0.21\%$ as well as $P(\mathsf{toppaper}|\mathsf{topmetric}) = 0.22\%$ and the fact that for a naive Bayes classifier $\mathsf{recall} = P(\mathsf{topmetric}|\mathsf{toppaper})$ and $\mathsf{precision} = P(\mathsf{toppaper}|\mathsf{topmetric})$ holds, one sees that a classifier with the above rule yields $\mathsf{recall} = 21\%$ and $\mathsf{precision} = 22\%$. Similarly, instead of $k$-core centrality other network measures presented in Table 3 can be used as basis for the above classification rule. As earlier works have tried to predict the success of papers based on the number of coauthors [14], using degree centrality as basis for the above classification rule directly extends these attempts, yielding a recall of 20% and a precision of 20%. Note, however, that degree centrality accumulates all coauthorships that have been established within the two-year sliding window of our analysis, not just the coauthorships of the paper under consideration.

We now ask whether a multi-dimensional naive Bayes classifier can improve this single metric classification result. Taking into account the intersection of all considered centrality metrics, we consider the following classification rule:

> *If a paper is authored by an author with a top* 10% *betweenness centrality, degree centrality, k-core centrality and eigenvector centrality, then the paper will be among the top* 10% *most cited papers five years after publication.*

Using this classifier, we achieve even better classification with a precision of 36%, however diminishing recall to 15%. Whereas these results already show that a naive Bayes classifier can yield interesting insights, in the following we will present a more sophisticated Machine Learning approach, taking multiple network centrality features into account and improving classification errors.

We first construct a feature vector for every publication as follows. For each publication appearing in year $t$, we extract all coauthors and compute the maximum and minimum of their centralities in the coauthorship network constructed based on the time window $[t-2, t]$. Then, for each publication we build a feature vector with ten features containing the maximum and minimum of the centrality metrics considered earlier (*degree*, *eigenvector*, *betweenness* and *k-core*), as well as the number of coauthors and the cumulative number of authors a paper has referenced. We then classify all publications regarding whether they fall in $P_\uparrow$ or $P_\downarrow$ according to the aforementioned publication classes, with $P_\uparrow$ defined as the set of the top 10% cited publications and $P_\downarrow$ as the remaining 90%.

The classification is done using a Random Forest classifier [25], extending the concept of classification trees (we use the R package *randomForest*, available at http://cran.r-project.org/web/packages/randomForest/). In general, the Random Forest classifier is known to yield accurate classifications for data with a large number of features [25]. Furthermore, it is a highly scalable classification algorithm, eliminating the need for separate cross validation and error estimation, as these procedures are part of the internal classification routine (for details on the procedure and the error estimates we refer to Additional file 1).

Table 5 summarizes precision, recall, and $F$-score of the resulting classification (see Supplementary Material (Additional file 1) for details on these measures). Comparing this result with the expectation from a random guess, which will correctly pick one of the top 10% publications only in 10% of the cases, the achieved precision of 60% is striking. In particular, by only considering positional features of authors in the coauthorship network,

**Table 5 Error estimates of the Random Forest classifier to predict success of papers.**

| Nr. publications | Precision | Recall | *F*-score |
|---|---|---|---|
| 36,000 | 0.6 | 0.18 | 0.28 |

we are able to achieve *an increase of factor six in predictive power* compared to a random guess. Also, we obtain a *recall* value of 18%, meaning that our classifier correctly identified about one fifth of all of the top 10% papers in a given research field. As a random guess would yield a recall of 10%, the Random Forest classifier *improves recall by* 80%.

This result allows for two conclusions: First, the fact that a high-dimensional random forest classifier performs better than a naive Bayes classifier, makes clear that social influence on scientific success cannot be measured by a single network metric and is instead a *multi-faceted* concept. Second, and most importantly, our result show that by *solely considering metrics of social influence*, such a classifier is able to predict scientific success with high precision.

Let us note that here we focused on the social influence on *success*. However, one might equally ask whether the complementary effect is true as well: can social factors predict whether a paper will be in the *bottom* 10% of all papers? We tested this hypothesis as well and found that, using the same procedure, with an achieved recall of 1.8% and a precision of 22.8%, whether a paper will be in the bottom 10% of all papers is *nearly unpredictable* using metrics of social influence only. Our interpretation of this finding is that even authors that are socially well connected will have papers that are not highly cited, simply because their content did not raise interest in the scientific community. This leads us to conclude that social factors are *necessary* factors for success, but are *not sufficient* – which is, in our opinion, a very easing result for the scientific community.

## 6 Discussion and conclusions

Using a data set on more than 100,000 scholarly publications authored by more than 160,000 authors in the field of computer science, in this article we studied the relation between the centrality of authors in the coauthorship network and the future success of their publications. Clearly, there are certain limitations to our approach, which we discuss in the following.

First of all, any data-driven study of social behavior in general and citation behavior in particular is limited by the completeness and correctness of the used data set. In our data set name ambiguities are automatically resolved by the Microsoft Academic Search (MSAS) database by sophisticated and validated disambiguation heuristics. This provides a clear advantage over simpler heuristics that have been used in similar studies. Although we did manual consistency checks of ambiguities for the top authors in our dataset, it is nevertheless not possible to exclude that there are some name ambiguities. However, since additionally author profiles in MSAS are to large parts manually edited by authors themselves, we are confident that name ambiguities are nearly negligible.

In order to rule out effects that are due to different citation patterns in different disciplines, we limited our study to computer science, for which we expect the coverage of MSAS to be particular good. While this limits the generalization of our results to other fields, our work nevertheless represents – to the best of our knowledge – the first large-scale case study of social factors in citation practices. As publication practices seem to

vary widely across disciplines, it will be interesting to investigate whether our results hold for other research communities as well.

Clearly, any study that tries to evaluate the *importance* or *centrality* of actors in a social network needs to be concerned about the choice of suitable centrality measures. In order to not overemphasize one particular – out of the many – dimensions of centrality in networks, we chose to use *complementary centrality measures* that capture different aspects of importance at the same time. The results of our prediction highlight that the combination of different metrics is crucial – making clear that visibility and social influence are more complicated to capture than by a single centrality metric.

Finally, one may argue that our observation that authors with high centrality are cited more often is not a statement of a *direct causal relation* between centrality and citation numbers. After all, both centrality and citations could be secondary effects of, for instance, the scientific excellence of a particular researcher, which then translates into becoming central and highly cited at the same time. Clearly, we neither can – nor do we want – to rule out such possible explanations for our statistical findings. However, considering our finding of strong statistical dependence between social centrality and citation success, one could provocatively state the following: if citation-based measures were to be good proxies for scientific success, so should be measures of centrality in the social network. We assume that not many researchers would approve having the quality of their work be evaluated by means of such measures.

In summary, the contributions of our work are threefold:

1. We provide the, to the best of our knowledge, first large-scale study that analyses relations between the position of researchers in scientific collaboration networks and citation dynamics, using a set of complementary network-based centrality measures. A specific feature of our method is that we study *time-evolving* collaboration networks and citation numbers, thus allowing us to investigate possible mechanisms of social influence at a microscopic scale.
2. We show that – at least for the measures of centrality investigated in this paper – there is no *single* notion of centrality in social networks that could accurately predict the future citation success of an author. We expect this finding to be of interest for any general attempt to predict the success of actors based on their centrality in social networks.
3. Using modern Machine Learning techniques, we present a supervised classification method based on a Random Forest classifier, using a multidimensional feature vector of collaboration network centrality metrics. We show that this method allows for a remarkably precise prediction of the future citation success of a paper, solely based on the social embedding of its authors. With this, our method provides a clear indication for a strong statistical dependence between author centrality and citation success. Additionally, we show evidence that author centrality is more of a necessary condition for success than a sufficient one.

In conclusion, we provided evidence for a strong relation between the position of authors in scientific collaboration networks and their future success in terms of citations. We would like to emphasize that by this we *do not* want to join in the line of – sometimes remarkably uncritical – proponents of citation-based evaluation techniques. Instead, we hope to contribute to the discussion about the manifold influencing factors of citation measures and their explanatory power concerning scientific success. Especially, we *do*

*not* see our contribution in the development of automated success prediction techniques, whose widespread adoption could possibly have devastating effects on the general scientific culture and attitude. Highlighting social influence mechanisms, we rather think that our findings are an important contribution to the ongoing debate about the meaningfulness and use of citation-based measures. We further hope that our work contributes to a better understanding of the multi-faceted, complex nature of citations and citation dynamics, which should be a prerequisite for any reasonable application of citation-based measures.

## Additional material

**Additional file 1: Supplementary material, including details on methods used in this research.**

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
All authors conceived and designed the research. ES and IS acquired the data. ES and RP analyzed the data. All authors discussed the research, wrote and approved the final version of the manuscript.

**Acknowledgements**
EM, IS and FS acknowledge funding by the Swiss National Science Foundation, grant no. CR31I1_140644/1. AG acknowledges funding by the EU FET project MULTIPLEX 317532. We especially thank Microsoft Research for granting unrestricted access to the Microsoft Academic Search service.

Received: 28 April 2014 Accepted: 29 July 2014 Published online: 25 September 2014

**References**

1. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569-16572. doi:10.1073/pnas.0507655102. http://www.pnas.org/content/102/46/16569.full.pdf+html
2. Heisenberg W (1969) Der Teil und Das Ganze: Gespräche Im Umkreis der Atomphysik. Piper und Co. Verlag, München
3. Garfield E (1955) Citation indexes for science: a new dimension in documentation through association of ideas. Science 122(3159):108-111. doi:10.1126/science.122.3159.108. http://www.sciencemag.org/content/122/3159/108.full.pdf
4. Leydesdorff L (1998) Theories of citation? Scientometrics 43(1):5-25. doi:10.1007/BF02458391
5. Radicchi F (2012) In science "there is no bad publicity": papers criticized in comments have high scientific impact. Sci Rep 2:815
6. Nicolaisen J (2003) The social act of citing: towards new horizons in citation theory. Proc Am Soc Inf Sci Technol 40(1):12-20. doi:10.1002/meet.1450400102
7. Laloë F, Mosseri R (2009) Bibliometric evaluation of individual researchers: not even right…not even wrong! Europhys News 40(5):26-29. doi:10.1051/epn/2009704
8. Bornmann L, Daniel H-D (2008) What do citation counts measure? A review of studies on citing behavior. J Doc 64(1):45-80
9. Radicchi F, Castellano C (2013) Analysis of bibliometric indicators for individual scholars in a large data set. Scientometrics 97(3):627-637
10. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. Proc Natl Acad Sci USA 105(45):17268-17272
11. Stringer MJ, Sales-Pardo M, Amaral LAN (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J Am Soc Inf Sci Technol 61(7):1377-1385. doi:10.1002/asi.21335
12. Katz JS, Hicks D (1997) How much is a collaboration worth? A calibrated bibliometric model. Scientometrics 40(3):541-554. doi:10.1007/BF02459299
13. Figg WD, Dunn L, Liewehr DJ, Steinberg SM, Thurman PW, Barrett JC, Birkinshaw J (2006) Scientific collaboration results in higher citation rates of published articles. Pharmacotherapy 26(6):759-767. doi:10.1592/phco.26.6.759
14. Hsu J, Huang D (2011) Correlation between impact and collaboration. Scientometrics 86(2):317-324
15. Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation in scientific publishing. arXiv:1304.0473
16. Bras-Amorós M, Domingo-Ferrer J, Torra V (2011) A bibliometric index based on the collaboration distance between cited and citing authors. J Informetr 5(2):248-264
17. Wallace ML, Lariviere V, Gingras Y (2012) A small world of citations? The influence of collaboration networks on citation practices. PLoS ONE 7(3):33339. doi:10.1371/journal.pone.0033339
18. Newman MEJ (2009) Networks: an introduction. Oxford University Press, New York
19. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal:1695
20. Newman MEJ (2009) The first-mover advantage in scientific publication. Europhys Lett 86(6):68001
21. Mann HB, Whitney DB (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50-60. doi:10.1214/aoms/1177730491

22. Hirsch JE (2007) Does the *h* index have predictive power? Proc Natl Acad Sci USA 104(49):19193-19198. doi:10.1073/pnas.0707962104. http://www.pnas.org/content/104/49/19193.full.pdf+html
23. Acuna DE, Allesina S, Kording KP (2012) Future impact: predicting scientific success. Nature 489:201-202. doi:10.1038/489201a
24. Newman MEJ (2013) Prediction of highly cited papers. arXiv:1310.8220
25. Breiman L (2001) Random forests. Mach Learn 45(1):5-32

doi:10.1140/epjds/s13688-014-0009-x
**Cite this article as:** Sarigöl et al.: **Predicting scientific success based on coauthorship networks.** *EPJ Data Science* 2014 **2014**:9.