Associations between author-level metrics in subsequent time periods

Ana C. M. Brito[1], Filipi N. Silva[2] and Diego R. Amancio[1]

[1]*Institute of Mathematics and Computer Science,*

University of São Paulo, São Carlos, Brazil

[2]*Indiana University Network Science Institute,*

Bloomington, Indiana 47408, USA

Abstract

Understanding the dynamics of authors is relevant to predict and quantify performance in science. While the relationship between recent and future citation counts is well-known, many relationships between scholarly metrics at the author-level remain unknown. In this context, we performed an analysis of author-level metrics extracted from subsequent periods, focusing on visibility, productivity and interdisciplinarity. First, we investigated how metrics controlled by the authors (such as references diversity and productivity) affect their visibility and citation diversity. We also explore the relation between authors' interdisciplinarity and citation counts. The analysis in a subset of Physics papers revealed that there is no strong correlation between authors' productivity and future visibility for most of the authors. A higher fraction of strong positive correlations though was found for those with a lower number of publications. We also found that reference diversity computed at the author-level may impact positively authors' future visibility. The analysis of metrics impacting future interdisciplinarity suggests that productivity may play a role only for low productivity authors. We also found a surprisingly strong positive correlation between references diversity and interdisciplinarity, suggesting that an increase in diverse citing behavior may be related to a future increase in authors interdisciplinarity. Finally, interdisciplinarity and visibility were found to be moderated positively associated: significant positive correlations were observed for 30% of authors with lower productivity.

1

I. INTRODUCTION

The age of information promoted several new discoveries in science, with many of them emerging from interdisciplinary endeavors [4, 10, 19]. At the same time, these communities are growing in size and productivity, resulting in an ever-increasing deluge of digital information available in the form of published articles [15], datasets, and algorithms, as well as across many platforms, such as cloud services and social media. However, the increase in digital resources has not leveled the playing field for researchers; inequality is rising in science [26].

Understanding the mechanisms leading to inequality in science can help policymakers and funding agencies better distribute research resources while also promoting a more just and democratic environment. Part of this problem relies on the fact that researchers compete among themselves for limited funding and attention. In such a system, an increase of researchers' visibility leads to better funding opportunities which, in turn, leads to more availability of resources for their institutions, thus allowing those researchers to attain even greater visibility.

The cycle in which researchers with the most resources are rewarded with even more resources over time is a source of inequality known as the Matthew Effect [12]. This is one of the reasons why understanding how the dynamics of authors visibility unfold over time is one of the most important problems in the field of Science of Science [8, 22]. However, not much attention was given to understand the relationships between other authors metrics besides citations [2, 7, 18, 21]. In particular, the literature lacks studies on metrics controlled by the authors, such as those based on their choices of references and their productivity; and the possible effects they may have on their received citations.

Here, we propose to explore the associations among different bibliometric measures for authors in subsequent time periods through correlation. Among the metrics we consider are interdisciplinarity, which is measured in terms of the subject diversity [19] of citations received by the authors, productivity and visibility of authors. Here, we are interested in addressing three main questions:

1. How metrics controlled by the authors – namely, their productivity and diversity in the choice of references – correlate with visibility metrics, such as the received number of citations per paper, in a subsequent time period?

2. How these characteristics correlate with the future interdisciplinarity of their publications based on citations?

3. How interdisciplinarity is related to future citations and vice versa?

In addition to productivity and visibility, we also studied interdisciplinarity as it plays an important role in modern science given the increasing number of authors bridging new different fields. Here, it is used as a descriptor for citation diversity, as adopted in related works [19]. In a similar fashion, we adopt a reference diversity for authors based on the fields of the employed references in their publications.

We employ the *American Physical Society* (APS) dataset, which incorporates all the citations and metadata for papers, mainly in Physics, published in any of the APS journals up to 2010. More specifically, we employ the dataset used in [20], which was supplemented with disambiguated authors from the *Microsoft Academic Graph* (MAG). First, we construct a co-occurrence network for the categories existing in the APS journals (PACS codes) which is used to define a metric of interdisciplinarity for authors based, in terms of the diversity of their received citations or the references they used. Next, we calculate the correlation between the considered author-level metrics for a window considering previous publications and another which considers subsequent publications and citations. Finally, we use a statistical framework based on null models to obtain the significance of the correlations between the considered metrics.

Several interesting results have been obtained in our analysis. We found that the diversity of references may impact positively the observed future visibility for 1/3 of low-productivity authors. This effect is minimized when analyzing more productive authors, yet the fraction of authors that were positively affected varied between 22% and 25%. A weaker association between productivity and citation counts was found: the highest fraction of authors with a significant positive correlation was 21%. When comparing the fraction of authors displaying significant positive and negative correlations, both productivity and reference diversity turned out to be more positively than negatively correlated with authors' visibility. Surprisingly, we also found that reference diversity and future interdisciplinarity are strongly positively correlated for roughly 50% of authors. Finally, the association between interdisciplinarity and visibility revealed that an increase in interdisciplinarity is more likely to be linked to an increase in visibility for low productivity authors. Such positive significant

correlations were observed in roughly 30% of authors in that class. We believe our results can provide further insights into better understanding researchers' career dynamics.

II.   RELATED WORKS

In this paper, among other relationships, we analyze which factors affect the visibility of authors (measured in terms of citations). At the paper level, some correlations between paper features and the number of citations have been studied in the last few years. An important factor that has been found to affect the visibility of paper is related to the *interdisciplinarity* of venues in which they are disseminated. Different aspects of scientific pieces have been used to define interdisciplinarity indexes. In [19], journal citation networks are used to quantify how interdisciplinary a journal is. For a given journal, the diversity of citations from different areas is used to gauge interdisciplinary. Such a diversity is computed using the concept of *true diversity*, a measure widely used to express how diverse a set of elements from different classes is [5, 23, 25]. Subject areas and citation data were extracted from the *Journal Citation Reports* dataset. Some interesting conclusions were the positive correlation between the proposed interdisciplinary index and journals impact factor. In other words, interdisciplinary journals tend to have a higher impact factor than specialized journals.

Using a different approach, the study conducted in [4] also quantified journals interdisciplinarity. The authors used Scopus data comprising *Information and Communication Technology* publications. The relationship between scholars and journals was represented via bipartite graphs. After a SVD dimension reduction, a spectral co-clustering method was used to identify communities of scholars and journals. The diversity (i.e. the interdisciplinarity) of a journal was then defined by analyzing the unevenness of authors distribution over the obtained network communities. Such a dispersion was computed via Shannon entropy, Simpson diversity, and Rao-Stirling index [11]. High values of disparity metrics were found to occur in journals appearing between communities. Conversely, low diversity was observed mostly in network community cores.

A correlation between interdisciplinarity and citation impact was investigated in [27]. Three aspects of interdisciplinary were investigated at the paper level: variety, balance, and disparity. Variety is the total number of different disciplines (or *Web of Science* categories) cited by the paper, while balance corresponds to the evenness of the disciplines distribution,

computed via Shannon diversity. Disparity measures how different are the disciplines in the reference set. The authors analyzed the impact of papers using the Normalized Citation Score (NCS). The data set used was papers from Science Citation Index-Expanded (2005). A regression estimation analysis revealed that variety was positively associated with NCS. In contrast, both balance and disparity were negatively associated with NCS.

The impact of citing interdisciplinary papers on papers visibility was investigated in [10]. The authors characterized interdisciplinarity at the paper level by using papers references. Subdisciplines were defined by the UCSD map of science [3]. According to this map, the similarity between journals is based on the number of shared references (via bibliographic coupling) and keywords. An average-linkage clustering strategy generates a cluster of 13 different categories and the pairwise cluster distance is represented in a 3D Fruchterman-Reingold layout. An analysis of 25,000 documents showed that papers citing interdisciplinary sub-disciplines tend to receive more citations than papers with fewer references to interdisciplinary sub-disciplines. This study also grouped sub-disciplines by distance in the UCSD map and demonstrated that papers citing distant sub-disciplines tend to have higher relative citation rates than papers citing similar sub-disciplines.

At the author level, the study carried out in [17] investigated the effects of interdisciplinarity on scientists careers. The APS dataset was used, considering papers published between 1980 and 2009. The hierarchical system of subdisciplines classification – referred to as *Physics and Astronomy Classification Scheme* (PACS) – was used to measure the interdisciplinarity of an author. They proposed an index combining the total of PACS codes used during the entire author career and the average number of different classes appearing simultaneously in the author papers. Using this value, authors were grouped by different levels of interdisciplinarity: low, medium, and high. Based on these groups, it was observed that higher interdisciplinarity affects positively productivity. A statistical model was proposed to reproduce the original data. The factors considered in the model were the proposed interdisciplinarity index, the number of publications in each class, the number of citations, talent, reputation, and luck. The model reproducing the properties of the studied system revealed that authors with medium-high talent are the most successful ones. In addition, luck turned out to play an important role in career success. Surprisingly, it was found to be even more relevant than interdisciplinarity factors in some cases.

Another different source of factor concerns the well-known rich-get-richer paradigm. In

other words, if an author has received several citations, he/she has a higher tendency of receiving more citations if they have received a higher citation rate in the past. In [20], the authors describe a model for reproducing the distribution of authors citations in the APS dataset. Unlike other models, they included a recency factor so that more recent citation data receives a higher weight in the preferential attachment model. This model showed that the rich-get-richer paradigm describes the citation distribution for authors publishing in APS journals. Most importantly, they also found that recency plays an important role to define how broad the burstiness of citations are. The number of citations received by authors is strongly dependent on the total of citations received in the last 1-2 years [20].

III. METHODOLOGY

The methodology adopted in this paper can be divided into the following steps:

1. *Creation of PACS networks*: this phase is responsible for establishing and identifying the subfields inside the considered dataset. Groups of strongly connected subareas are grouped into network communities. The latter is used to identify an area, which in turn is used to define some of the variables of interest. The dataset used to create the networks is described in Section III A. The process of creating and identifying communities of co-occurring PACS is described in Section III B.

2. *Definition of diversity indexes*: here we use diversity indexes to quantify how diverse authors cite or are cited by other papers. The diversity takes as reference the subareas (communities) identified in the PACS network. The adopted diversity index is defined in Section III C. Diversity indexes are among the author-level metrics of interest in this paper.

3. *Quantifying the relationship between variables of interest*: here we quantify there are correlations between variables of interest quantified in subsequent time intervals. The methodology adopted to quantify the fraction of authors displaying significant positive/negative correlations between the variables of interest in described in Section III D.

A. Dataset

The dataset consists of papers published by the American Physical Society (APS) journals between 1991 and 2010. The dataset comprises 299,930 publications from APS journals. While the dataset provides several article metadata, we used for each paper the list of authors and the reference list. We also used the list of subfields codes provided by the authors and selected from the *Physics and Astronomy Classification Scheme* (PACS). This classification scheme is a hierarchical code system used to organize the main fields and subfields in Physics journals.

When addressing any issue at the author level, one should be aware that ambiguities and name split may arise [1, 13]. To address this problem, we used the Microsoft Academic Graph (MAG) dataset, which is a more extensive set of publications with authors' names disambiguated [20]. We mapped the APS dataset into the MAG database by matching DOIs values.

B. PACS Networks

In this work, we use the notion of subfields to compute the degree of interdisciplinarity inside the Physics area (for APS journals). Subfields were derived from PACS co-occurrence networks [16]. Each publication in the APS dataset has its PACS codes, and this information of area is provided by the authors, among a list of possible codes. We used this information to generate networks where nodes are PACS codes. Figure 1 shows an example of PACS co-occurrence network extracted from a set of papers. As suggested by other works, PACS were analyzed at the first two levels [16]. Two codes are linked whenever they appear together in one or more papers. Here we take the view that a subfield in the considered subset of Physics papers can be seen as a subset of highly connected codes. In this way, each subfield is defined as a community in the respective co-occurrence PACS network. While our results are based on the Louvain community detection algorithm [24], a preliminary analysis revealed that there is no large difference when other methods are used to detect communities. Considering the most recent years of the dataset, using the Louvain method, we found 10 network communities. An analysis of the obtained communities considering data from the last 5 years showed that the four largest communities are mainly composed of papers in

the following subjects: (i) *magnetic properties and materials*; (ii) *quantum mechanics, fiel theories, and special relativity*; (iii) *structure of solids and liquids; crystallography;* and (iv) *statistical physics, thermodynamics, and nonlinear dynamical systems.*

FIG. 1. Schematic representation of the components needed to calculate *citations* and *references* diversity.

C.   Diversity indexes

Here we employ a diversity index for authors based on the diversity of fields being cited (*citations diversity*) or referenced (*references diversity*) by their papers. Because usually *citation diversity* is related to *interdisciplinary* [19], we use both terms to describe the same concept. To assign a distribution of fields of a given author A, first, we look at all the papers $P_i^{(A)}$ citing publications co-authored by A during the considered time window. For each citing paper we obtain the communities associated to the PACS listed in the paper. Figure 1 illustrates the necessary components employed to calculate the *in*-diversity index for authors. Next, we derive the weights $w_{\text{in}}(P_i, C_j)$ relating a paper P_i to a PACS community

C_j, defined as the ratio of the number of PACS in C_j listed in P_i, i.e.

$$w(P_i, C_j) = \frac{|\text{PACS}(P_i) \cap C_j|}{|\text{PACS}(P_i)|}, \tag{1}$$

where $\text{PACS}(P_i)$ is the set of PACS listed in paper P_i. Next, we assign a weight $\bar{w}_{\text{cit}}(A, C_j)$ relating an author A to each PACS communities C_j based on the citing papers. Each citation to a paper from author A counts as a unit that is distributed among the communities, so that $\bar{w}(A, C_j)$ is defined as

$$\bar{w}_{\text{cit}}(A, C_j) = \sum_{P_i} n_{\text{cit}}(P_i, A) w(P_i, C_j), \tag{2}$$

where $n_{\text{cit}}(P_i, A)$ is the number of citations from P_i to author A Finally, we normalize $\bar{w}(A, C_j)$ across all the received citations, thus obtaining a probability-like measure $p_{\text{cit}}(A, C_j)$ of relatedness between an author A and a community C_j, given by

$$p_{\text{cit}}(A, C_j) = \frac{\bar{w}_{\text{in}}(A, C_j)}{\sum_{C_k} \bar{w}_{\text{in}}(A, C_k)}. \tag{3}$$

The *citation* diversity index $\text{citDiv}(A)$ is then defined as the exponential of entropy of $p_{\text{cit}}(A, C_j)$ [19], i.e.,

$$\text{citDiv}(A) = \exp\left[-\sum_{C_j} p_{\text{cit}}(A, C_j) \log p_{\text{cit}}(A, C_j) \right]. \tag{4}$$

Similarly, to obtain *references* diversity index, we use the papers P_i referenced by works authored by author A instead of the received citations. Thus, the weight linking an author and a PACS community is defined as

$$\bar{w}_{\text{ref}}(A, C_j) = \sum_{P_i} n_{\text{ref}}(A, P_i) w(P_i, C_j), \tag{5}$$

where $n_{\text{ref}}(A, P_i)$ is the number of times author A cited the paper P_i. The probability analogous to p_{cit} (i.e. p_{ref}) is then normalized as:

$$p_{\text{ref}}(A, C_j) = \frac{\bar{w}_{\text{ref}}(A, C_j)}{\sum_{C_k} \bar{w}_{\text{ref}}(A, C_k)}, \tag{6}$$

and the *references* diversity refDiv(A) is calculated as

$$\text{refDiv}(A) = \exp\left[-\sum_{C_j} p_{\text{ref}}(A, C_j) \log p_{\text{ref}}(A, C_j)\right].\tag{7}$$

Both equations 3 and 6 have been used to measure diversity in many contexts [5, 6, 19]. Because the computation of p_{cit} and p_{cit} are not reliable when only a few data is available, these quantities were computed for authors with more than ten references and citations in the dataset.

D. Past and Future scholarly time series

We propose a framework to analyze how a scholarly metric or diversity at a certain point in time for an author A may impact his future metrics. First, we define two moving windows, one for the past and another for the future, respectively a 5 years window before the time under consideration t, and a 3 years window after t, as illustrated in Figure 2a. For each window, we calculate the scholarly metrics of A. In particular, for the Past window, we calculate the number of papers, citations received per paper, and references diversity, only considering publications in the period.

FIG. 2. Schematic representation of the proposed methodology. (a) Given two subsequent windows (past and future) that moves over time, we calculate the time series of the considered metrics. (b) For each time series we derive a null model based on shuffling them along time. (c) We draw the correlation distribution (gray) obtained from the data time series and highlight negative (blue) and positive (yellow) values that are significant in comparison to the null models. The average null model distribution is also shown for comparison in red.

For the Future window, we calculate the number of citations received in that window from

papers published by A during the Past window. In the same fashion, we calculate citation diversity by considering only publications in the Past windows and citations in the Future window. By moving the windows along t for a period from 1995 to 2010, we obtain Past (number of papers, citations per author, and the references diversity), and Future (citations received per paper and citation diversity) time series for each author based on the calculated scholarly metrics.

In order to draw relationships between the scholarly metrics from past and future windows, we adopted the Pearson correlation. However, as these metrics may have characteristics that can lead to spurious correlations, such as the presence of outliers or long-tail distributions, we employed a statistical approach based to measure the significance of the obtained correlations. First, for each time series of each author, we obtain a set of $10,000$ surrogates generated by shuffling the original data along time, which is regarded as a null model, as illustrated in Figure 2b.

When calculating the correlations between two scholarly metrics, we also calculate the respective correlation distribution from their null models. This distribution is used to calculate a p-value associated with each author and a pair of past and future metrics. The p-value is defined as the probability of the null model resulting in a *absolute* correlation that is higher than what was found for the data. Finally, the results are presented in the form of a correlation distribution alongside the percentage of negative and positive significant relationships by considering a threshold of 5×10^{-2} for the p-values. This is illustrated in Figure 2c.

IV. RESULTS AND DISCUSSION

Here we analyze the relationship between relevant author-level metrics. More specifically, we analyze, if the diversity of references, the numbers of papers, and the number of references are correlated with citation counts and citation diversity. We first focus on the relationship between variables that authors can control in the first 5-year window (e.g. the number and diversity of references) and variables that are not directly self-dependent (such as the number of citations and citation diversity) and are measured in the following 3-year window. The correlations between paper/reference features and citation counts are discussed in Section IV A. The correlations between paper/reference features and citation diversity are discussed

in Section IV B. Because interesting relationships between interdisciplinarity (i.e. citation diversity) and citation counts have been reported at different levels [4, 14, 19], we also analyzed the correlations between interdisciplinarity and citations at the author level. This is reported in Section IV C.

A. Correlations between reference features and citations

The simplest reference feature that can be used in our analysis is the total number of references. For the sake of clarity, we will use instead of the number of papers (i.e. the authors productivity) in our analysis because the total number of references is strongly related to the number of papers. In addition, the results using either number of references or the number of papers are very similar.

We start our analysis by analyzing whether productivity – i.e. the number of published papers – is correlated with the total number of citations per paper. This result is shown in Figure 3. As mentioned in the methodology (see Section III D), the histograms shows the distribution of authors in different degrees of correlation between the variables of interest. Each panel corresponds to a different class of authors, according to its productivity. The authors analyzed in subpanels (a)-(d) are those who published the following amount of papers over all the considered period: (a) 5–25; (b) 26–36; (c) 37–58; and (d) 59–359 papers. The considered thresholds in the number of publications were chosen so that each class comprises 25% of all authors in the dataset. In other words, each panel corresponds to a quartile of authors. In this figure, the distribution of correlations observed using the null model is represented by the red curve (see Section III D). The fraction of authors displaying significant positive and negative correlations between the considered variables are represented in yellow and blue, respectively.

The results in Figure 3 reveals that the observed distribution in all panels differs from the null model distribution. The discrepancy between real data and null model arises since very high or low values of correlations are unlikely to happen by chance, while the real data reveals an opposite effect: for a fraction of authors, the correlations are significant. Considering all four classes, 18-22% of authors displayed a *positive* correlation between productivity and visibility. On the other hand, a *negative* correlation was also observed in all classes of authors. The percentage of authors displaying a *negative* correlation between

FIG. 3. Correlation between the *total number of published papers* and *citations per paper*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The distribution of correlations obtained with the adopted null model is shown in the red curve.

productivity and visibility ranged between 10% and 15%. Because more than 64% of the observed correlations are not significant in all four classes of authors, the results suggest that for most of the authors the increase in productivity is not correlated with higher citation counts per paper.

In our analysis, we also compared the proportion of positive (f^+) and negative (f^-) correlations. The proportions are compared via q-index, defined as

$$q = \frac{f^+}{f^-}. \tag{8}$$

In this case, all values of q are higher than $q = 1$, suggesting thus that in all classes of

13

authors positive correlations are more likely to appear. The highest value of q was observed for authors with the lowest number of publications (see panel (a)). We found $q = 2.08$, meaning that positive correlations are twice more likely to appear than negative correlations considering this class of authors.

The results regarding productivity suggest that significant correlations between productivity and citation rates occur only to a small percentage of authors. This effect is more prominent in authors with lower productivity. This is an indication that, for most of the authors, increasing productivity does not improve authors' visibility in the near future.

In Figure 4, we show the histograms of correlations between the *diversity of references* and the *number of citations per paper*. A stronger positive correlation is observed specially for authors with lower productivity. In panel (a), one-third of authors displayed a positive correlation between references diversity and visibility, while in (b), the same behavior occurred for one-fourth of all authors. In both cases, positive correlations are more frequent than negative correlations. We found, $q = 5.22$ and $q = 2.65$, respectively for authors in classes (a) and (b). Authors in classes (c) and (d) displayed q values similar to those observed in class (b).

The analysis of reference diversity showed that the way in which authors cite other works may affect their visibility in the near future. This effect was found to be more relevant than the productivity since significant positive correlations were found in up to 25% of authors. This effect might be related to the fact that diverse references might attract attention from other subfields, favoring thus the dissemination of authors' visibility in other scientific communities. In fact, a similar effect has been reported at the journal analyses comparing the relationship between journals impact factor and interdisciplinary indexes [19]. Similar effects have also been observed in diffusion systems, where the presence across different communities benefits the spreading of agents [9]. While it is not possible to establish a causal effect, our results suggest that references diversity (inside a field) might play a role in predicting authors' visibility.

B. Correlations between reference features and citations diversity

While in the previous section we analyzed how references features correlate with visibility, here we investigate the relationship between references and the diversity of citations. Be-

FIG. 4. Correlation between *diversity of references* and *citations per paper*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The distribution of correlations obtained with the adopted null model is shown in the red curve.

cause citations diversity can be seen as an interdisciplinary index (see e.g. [19]), this section analyzes how the choice (and quantity) of references is related to authors interdisciplinarity.

Figure 5 depicts the correlations between the *number of published papers* and *citation diversity*. As observed in the results reported in the previous section, for most of the authors there is no significant correlation between the considered variables. However, a positive correlation is observed for 1/3 of all authors in class (a), while 1/4 of authors displayed a positive correlation in the other classes. A negative correlation is less frequent than positive correlations. In addition, the values of q decreases with productivity, since we obtained $q_A = 8.2$, $q_B = 3.3$, $q_C = 2.2$ and $q_D = 2.2$ respectively for classes (a), (b), (c) and (d). The

results suggest, therefore, that an increase in productivity is more likely to play a role in increasing interdisciplinary for the class of authors with a lower degree of productivity.

FIG. 5. Correlation between number of *published papers* and *citations diversity*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The red curve denotes the distribution of correlations obtained with the adopted null model.

The association between references and citation diversity was also analyzed. This result is shown in Figure 6. The observed correlations are much stronger than the ones analyzed so far. The null model distribution is clearly not compatible with the real data. Here, a significant relationship between reference and citation diversity arises for *more than 50% of all authors*. Surprisingly, virtually all significant correlations are positive. The percentage of positive correlations reaches roughly 50%, while significant negative correlations were observed for roughly 1.5% of all authors. Another distinctive feature of the relationship between reference and citation diversity lies in the fact that the relationship is similar for

FIG. 6. Correlation analysis between *reference diversity* and *citation diversity*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The red curve denotes the distribution of correlations obtained with the adopted null model.

all classes of authors. This result is, therefore, a strong evidence that researchers who cite papers from many other disciplines might be also cited by many other subareas. In other words, if authors display a diverse behavior when citing other papers, they also tend to be cited by other diverse subareas. Because citation diversity can be seen as a way to measure authors' interdisciplinary [19], most of the authors adopting larger reference diversity in a given period are expected to increase their interdisciplinarity indexes in the near future.

C. Interplay between interdisciplinarity and citations

Here we analyze the relationship between interdisciplinarity of authors (computed as citation diversity) and the number of citations. While studies have shown that a positive correlation exists between journals interdisciplinarity and impact factor [19], only a few studies have touched on this issue at the author level. Here we found that for most of the authors ($\geq 62\%$), there is no significant relationship between interdisciplinarity and the number of citation received by the authors. However, a positive correlation can be found for a considerable fraction of authors. This is more evident again for authors in class (a), as shown in Figure 7: roughly 30% displayed a significant positive correlation. Figure 7 also reveals a tendency of higher positive significant correlation over a negative one: the values of q for each class are $q_A = 3.9$, $q_B = 2.2$, $q_C = 1.8$ and $q_D = 1.4$. As observed in other associations studied here, higher values of q are found for authors in the group of lower productivity.

While Figure 7 only show the relationship between interdisciplinarity and future visibility, it would be still interesting to see if there is an inverse effect. To investigate if variation in citations is correlated to a future variation in interdisciplinarity we conducted an analysis similar to the one provided in Figure 7. The histograms of correlations are shown in Figure 8. Overall the histograms are similar to the ones depicted in Figure 7, but here the fraction of significant positive correlations are smaller. This is evident e.g. for authors in (a): the fraction of positive correlations drop from 29.1% to 24.7%. This suggests that a variability in citation counts is weakly correlated with the future author interdisciplinarity for most of authors. In other words, for most of the authors, a variability in visibility does not affect the future authors' interdisciplinary indexes.

V. CONCLUSION

In this paper, we analyzed whether relevant scholarly variables are correlated. We proposed a framework to probe if features extracted from authors' recent history are correlated with metrics observed a few years later. While some correlations are trivial and were not object of study (such as correlations between citations in subsequent time periods [20]), we studied the correlation between other variables of interest. We focused our analysis on sim-

FIG. 7. Correlation analysis between *citation diversity* and *citations per paper*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The red curve denotes the distribution of correlations obtained with the adopted null model.

ple, yet relevant metrics, including number of publications, number of citations, references diversity and authors' interdisciplinarity (measured via citation diversity).

Several interesting results have been obtained. Among the associations studied, we found that the strongest correlations were obtained between references diversity and authors' interdisciplinarity. Here we found a reciprocal tendency: if authors increase their diversity when citing other papers, received citations will also tend to increase. This pattern was observed for more than 50% of authors. The relationship between productivity and visibility was found to be more prominent for authors with a lower productivity. While no significant correlation exists for most of authors, about 20% showed a positive and significant correlation. A stronger association was obtained when analyzing the relationship between

FIG. 8. Correlation analysis between *citations per paper* and *citation diversity*. Panels (a)-(d) correspond to quartiles of authors sorted, in increasing order, by number of publications. The red curve denotes the distribution of correlations obtained with the adopted null model.

references diversity and future citation. For the class of authors with a lower productivity, we found that roughly 1/3 of authors displayed a significant positive correlation between references diversity and visibility. We also studied the association between references and citation diversity and found out that the fraction of positive significant correlations ranges between 18-30% across different classes of authors.

Our study shed lights into the relationship between current and future researchers' activity. The results obtained here could be extended in diverse studies to provide mechanisms to predict authors' behavior, given the recent researchers' history. Future research could dive into other research questions arising from our analysis. For example, while we found that significant positive correlations are more likely to happen than negative ones, it would

be interesting to probe which factors make authors display opposite behaviors for the same variables of interest. Another interesting feature that could be studied concerns the causality of the obtained correlations. Finally, a systematic study could be performed in different areas to check whether correlations are more significant in specific subfields.

ACKNOWLEDGMENTS

[1] D. R. Amancio, O. N. Oliveira Jr, and L. d. F. Costa. On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *EPL (Europhysics Letters)*, 99(4):48002, 2012.

[2] D. R. Amancio, O. N. Oliveira Jr, and L. da Fontoura Costa. Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index. *Journal of informetrics*, 6(3):427–434, 2012.

[3] K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Larivière, and K. W. Boyack. Design and update of a classification system: The ucsd map of science. *PloS one*, 7(7), 2012.

[4] C. Carusi and G. Bianchi. A look at interdisciplinarity using bipartite scholar/journal networks. *Scientometrics*, pages 1–28, 2019.

[5] E. A. Corrêa Jr, F. N. Silva, L. d. F. Costa, and D. R. Amancio. Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, 11(2):498–510, 2017.

[6] H. F. de Arruda, L. d. F. Costa, and D. R. Amancio. Using complex networks for text classification: Discriminating informative and imaginative documents. *EPL (Europhysics Letters)*, 113(2):28007, 2016.

[7] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PloS one*, 6(9):e24926, 2011.

[8] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379), 2018.

[9] M. Kaiser, M. Goerner, and C. C. Hilgetag. Criticality of spreading dynamics in hierarchical cluster networks without inhibition. *New Journal of Physics*, 9(5):110, 2007.

[10] V. Larivière, S. Haustein, and K. Börner. Long-distance interdisciplinarity leads to higher scientific impact. *Plos one*, 10(3), 2015.

[11] L. Leydesdorff, C. S. Wagner, and L. Bornmann. Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics*, 13(1):255–269, 2019.

[12] R. K. Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

[13] S. Milojević. Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, 7(4):767–773, 2013.

[14] K. Okamura. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1):1–9, 2019.

[15] R. K. Pan, A. M. Petersen, F. Pammolli, and S. Fortunato. The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3):656–678, 2018.

[16] R. K. Pan, S. Sinha, K. Kaski, and J. Saramäki. The evolution of interdisciplinarity in physics research. *Scientific reports*, 2(1):1–8, 2012.

[17] A. Pluchino, G. Burgio, A. Rapisarda, A. E. Biondo, A. Pulvirenti, A. Ferro, and T. Giorgino. Exploring the role of interdisciplinarity in physics: Success, talent and luck. *PloS one*, 14(6), 2019.

[18] F.-X. Ren, H.-W. Shen, and X.-Q. Cheng. Modeling the clustering in citation networks. *Physica A: Statistical Mechanics and its Applications*, 391(12):3533–3539, 2012.

[19] F. N. Silva, F. A. Rodrigues, O. N. Oliveira Jr, and L. d. F. Costa. Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, 7(2):469–477, 2013.

[20] F. N. Silva, A. Tandon, D. R. Amancio, A. Flammini, F. Menczer, S. Milojević, and S. Fortunato. Recency predicts bursts in the evolution of author citations. *Quantitative Science Studies*, 1(3):1298–1308, 2020.

[21] M. V. Simkin and V. P. Roychowdhury. Stochastic modeling of citation slips. *Scientometrics*, 62(3):367–384, 2005.

[22] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312), 2016.

[23] J. V. Tohalino and D. R. Amancio. Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539, 2018.

[24] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[25] H. Tuomisto. A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22, 2010.

[26] Y. Xie. "undemocracy": inequalities in science. *Science*, 344(6186):809–810, 2014.

[27] A. Yegros-Yegros, I. Rafols, and P. D'Este. Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PloS one*, 10(8), 2015.