



# Do open citations give insights on the qualitative peer-review evaluation in research assessments? An analysis of the Italian National Scientific Qualification

Federica Bologna<sup>1</sup> · Angelo Di Iorio<sup>2</sup> · Silvio Peroni<sup>3</sup> · Francesco Poggi<sup>4</sup>

Received: 8 March 2021 / Accepted: 25 October 2022  
© The Author(s) 2022

## Abstract

In the past, several works have investigated ways for combining quantitative and qualitative methods in research assessment exercises. Indeed, the Italian National Scientific Qualification (NSQ), i.e. the national assessment exercise which aims at deciding whether a scholar can apply to professorial academic positions as Associate Professor and Full Professor, adopts a quantitative and qualitative evaluation process: it makes use of bibliometrics followed by a peer-review process of candidates' CVs. The NSQ divides academic disciplines into two categories, i.e. *citation-based* disciplines (CDs) and *non-citation-based* disciplines (NDs), a division that affects the metrics used for assessing the candidates of that discipline in the first part of the process, which is based on bibliometrics. In this work, we aim at exploring whether citation-based metrics, calculated only considering open bibliographic and citation data, can support the human peer-review of NDs and yield insights on how it is conducted. To understand if and what citation-based (and, possibly, other) metrics provide relevant information, we created a series of machine learning models to replicate the decisions of the NSQ committees. As one of the main outcomes of our study, we noticed that the strength of the citational relationship between the candidate and the commission in charge of assessing his/her CV seems to play a role in the peer-review phase of the NSQ of NDs.

**Keywords** Scientometrics · Open access · Bibliometrics · Citation-based metrics · Citation networks · Academic evaluation · Open science · Dimensions of proximity

## Introduction

A key concept of scientometrics is the study and quantification of the scientific impact of a work. In the sociology of science, two competing theories of citing behavior have been developed: the normative theory and the social constructivist view. Following Merton's (1973) sociological theory of science, the *normative theory* affirms that scientists, through the citation of a scientific work, recognize a credit to colleagues whose results

---

✉ Francesco Poggi  
francesco.poggi@cnr.it

Extended author information available on the last page of the article

they have used. As summarized by Bornmann (2016), “this means that citations represent an intellectual or cognitive influence on their scientific work”. A second view on citing behavior is *social constructivist* grounded on theory advanced in the sociology of science (Knorr-Cetina, 1981; Latour & Woolgar, 1979). Constructivists claim that the cognitive content of papers has poor influence on how they are received. This approach questions the base assumptions of normative theory, and so challenges the validity of using citation analysis for evaluation purposes. As stated by Bornmann (2016), “scientific knowledge is socially constructed through the manipulation of political and financial resources and the use of rhetorical devices” (Knorr-Cetina, 1991). The reasons that lead an author to a citation are complex and variously socially constructed, and these motivations have long been acknowledged in the field (Garfield, 1962).

Empirical tests on the validity of the two theoretical approaches were undertaken in Baldi (1998) and White (2004), and investigations on the possibility for a rapprochement of the normative and constructivist theories have been performed by Small (2004).

Several works in the field of scientometrics have investigated and given increasing importance to combined quantitative and qualitative methods of assessment. In 1987, a workshop was organized by John Irvine, Loet Leydesdorff, and Anthony van Raan on “The relations between qualitative theory and scientometric methods in science and technology studies”, which resulted in a special issue of *Scientometrics* (1989). In the introduction, Leydesdorff (1989) claims that there is increasing recognition of the need to integrate qualitative theories with quantitative techniques provided. The various contributions in the issue elucidate the mutual relevance of theory and empirical scientometric work. For example, Luukkonen (1989) advocates for the use of multi-faceted approaches in assessment studies. Moreover, Kranakis and Leydesdorff (1989) put in relation the results of a historical approach and a scientometric analysis (word occurrences and patterns) for the reconstruction of the cognitive organization of teletraffic science.

After this first issue, an increasing number of works have focused on the divide and connection between qualitative and quantitative science studies. For example, Milojevic et al. (2014) find that chapters from both quantitative and qualitative handbooks of Science and Technology Studies (STS) shared similar academic interests; and Wyatt et al. (2017) argue for the integration of quantitative methods within qualitative analyses. One of the most recent contributions to the topic is the recent issue of *Quantitative Science Studies* on “Bridging the divide between qualitative and quantitative science studies” curated by Loet Leydesdorff, Ismael Ràfols, and Staša Milojevic (2020). Here, Frenken (2020) assembles several theoretical traditions into a single analytical framework and research design by conceptualizing the diffusion of scientific knowledge as different forms of proximity. In addition, Fox (2020) shows that scholarship on gender inequalities in science could benefit from multi-level studies that utilize different theoretical and methodological approaches. Finally, Marres and de Rijcke (2020) present an example of methodology development that combines quantitative approaches with interpretative ones, and propose that the former should be devised according to their context of use, .

This study both draws from and expands this discussion in the context of the Italian National Scientific Qualification (NSQ – in Italian, *Abilitazione Scientifica Nazionale*, or ASN). The NSQ is the national assessment exercise which aims at deciding whether a scholar can apply to professorial academic positions as Associate Professor and Full Professor. It consists of a quantitative and qualitative evaluation process, which first makes use of bibliometrics, to filter candidates who pass specific thresholds, then followed by a peer-review process, where candidates’ CVs are assessed by a committee (a.k.a. the commission) to get to the final decision.

The NSQ divides academic disciplines into two categories, i.e. *citation-based* disciplines (CDs) and *non-citation-based* disciplines (NDs)<sup>1</sup>. This division affects only the metrics used for assessing the candidates of that discipline in the first part of the process, which is based on bibliometrics. In this work, we aim at exploring whether citation-based metrics can support the human peer-review of NDs and yield insights on how it is conducted. Specifically, our work focuses on citation-centric metrics that are devised to capture the relationship between the candidate of the NSQ and the commission in charge of the decision upon the candidate, by measuring the overlap between their citation networks. Our hypothesis is that citations give insights on the qualitative peer-review evaluation that follows the preliminary assessment, even in the case of NDs. Therefore, citation metrics should act as a proxy for and mimic the outcome of the human evaluation phase. To measure this relation, we tried to analyze such citation data using a computational method that included two machine learning classifiers, namely SVM and Decision Trees. Our aim was also to experiment how AI-aided analysis can support Science of Science research, as suggested in (Fortunato et al. 2018).

Our study makes a further step: it exclusively utilizes open bibliographic data. Indeed, our citation-centric metrics have been calculated using data collected from open access datasets only. This makes it possible for other scholars to reproduce our analysis, and thus, support the quality of our study, and let us investigate the potentialities of such an approach and its repeatability.

In recent years, many scholars – led by internationally-coordinated efforts such as the Initiative for Open Citations (I4OC, <https://i4oc.org>) and the Initiative for Open Abstracts (I4OA, <https://i4oa.org>) – have come forward to advocate for the free accessibility of bibliographic and citation data. This push has resulted in numerous projects, such as COCI (Heibi et al., 2019), OUCI (Cheberkus & Nazarovets, 2019), Unpaywall (Else, 2018), and the extensions of VOSviewer (van Eck & Waltman, 2010) to handle open scholarly meta-data and open citations. Previous works have already shown open data’s research potential in the study of bibliometric disciplines – e.g. (Bedogni et al., 2022; Bologna et al., 2022; Chudlarský & Dvořák, 2020; Di Iorio et al., 2019; Huang et al., 2019; Martín-Martín et al., 2020; Peroni et al., 2020; Zhu et al., 2019).

Therefore, the research question of this paper can be summarized as follows: *can the citation-centric metrics we computed using open access datasets provide insights on the human evaluation of these disciplines? In particular, what role does the relationship between the candidate and the commission play in the peer-review phase of the NSQ?*

To answer these questions, we ground our analysis on the data of the candidates and commissions that took part in the 2016, 2017, and 2018 terms of the NSQ. A necessary, but not sufficient, condition is that the open access datasets are sufficiently rich to represent the scientific production of these candidates, on top of which we could calculate citation metrics. In fact, we measured how many publications listed in the candidates CVs are found in the open datasets and focused on the two disciplines that were the best covered,

---

<sup>1</sup> According to the NSQ rules and nomenclature, the names used to identify CDs and NDs are bibliometric disciplines and non-bibliometric disciplines respectively. Naively, this suggests that only disciplines of the first kind are evaluated using bibliographic metrics. However in the first step of the NSQ evaluation all candidates are evaluated using bibliometric information, making this terminology inconsistent with that used by the scientometric community. In order to better contextualize this study and its purposes, we decide to rename these disciplines into CDs and NDs. Indeed, there are two citation-based metrics among those used for assessing the first type of disciplines; whereas, no citation-based metric is used to evaluate the second type of disciplines.

among all the NDs and did not belong to the same Scientific Area: *Historical and General Linguistics* and *Mathematical Methods of Economics, Finance and Actuarial Sciences*, having Recruiting Fields<sup>2</sup> 10/G1 (*Antiquities, Philology, Literary Studies, Art History*) and 13/D4 (*Economics and Statistics*) respectively according to the NSQ classification. The coverage was high, as detailed in Section ‘[Methods and Materials](#)’, and we could compute reliable metrics.

We collected the bibliographic metadata and citation data from the following open datasets: Microsoft Academic Graph (<https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>) (Wang et al., 2020), OpenAIRE (<https://www.openaire.eu/>) (Rettberg and Schmidt, 2012), Crossref (<https://www.crossref.org/>) (Hendricks et al., 2020), and OpenCitations (<https://opencitations.net/>) (Peroni & Shotton, 2020). In addition, we devised and reused a set of bibliographic and citation-based metrics using a top-down process and combining popular metrics (e.g. bibliographic coupling) to others proposed for the purpose of this study to describe the relationship between the candidates and the commission of the NSQ. Thanks to the NSQ data and the metrics adopted, we verify whether using our metrics as input to a series of computational methods produces similar results to the ones of the human evaluation conducted by NSQ experts.

## The National Scientific Qualification: how it works

In 2011, Italian Law of December 30th 2010 n.240 (L. 240/2010, 2011) introduced extensive changes in the organization and evaluation of higher education institutions and scholars. In particular, it made it mandatory to pass the NSQ in order to apply to academic positions. The NSQ consists of two distinct qualification procedures designed to attest two different levels of scientific maturity of a scholar. The first procedure qualifies the candidate for the role of Full Professor (FP); whereas the second qualifies the candidate for the role of Associate Professor (AP). It is worth mentioning that passing the NSQ does not grant a tenure position. Each university is responsible for creating new positions according to financial and administrative requirements and in compliance with local hiring regulations.

Besides, Ministerial Decree of June 14th 2012 (Ministerial Decree 159, 2012) defines a taxonomy of 184 Recruitment Fields (RF) divided in groups and sorted into 14 different Scientific Areas (SA). SAs correspond to vast academic disciplines, whereas RFs correspond to specific scientific fields of study. Each scholar is assigned to a specific RF which belongs to a single SA. In the taxonomy, RFs are identified by an alphanumeric code in the form AA/GF. AA is a number indicating the SA and it ranges from 1 to 14. G is a single letter identifying the group of RFs. F is a digit indicating the RF. For instance, *Neurology*’s code is 06/D5, where 06 indicates the SA *Medicine* and D indicates the group *Specialized Clinical Medicine* (Ministerial Decree 159, 2012). When applying for the NSQ, scholars can choose to be evaluated for more RFs at a time. Since each RF has its own assessment rules, the candidate may pass the qualification in some fields and not in others.

The first two terms of the NSQ took place in 2012 and 2013. Although L. 240/2010 prescribes that the NSQ must be held at least once a year, the next session took place from 2016 to 2018: with 1 term in 2016, 2 terms in 2017 and 2 terms in 2018. Those who fail the qualification cannot apply again to the same RF in the year following the date of

<sup>2</sup> Recruiting Fields correspond to specific scientific fields of study (Ministerial Decree 159, 2012).

submission of the application. Once acquired, the qualification's certificate lasts for nine years. For each RF, the Ministry of University and Research (MIUR) appoints an evaluation committee (a.k.a. a commission) composed of five full professors responsible for assessing applicants for associate and full professorships. In order to apply to the NSQ, candidates have to submit a curriculum vitae with detailed information about their research accomplishments.

In the preliminary phase of the evaluation process, each candidate's academic expertise is assessed using bibliometrics. These metrics vary depending on whether the candidate has applied to a CD or a ND. Candidates applying to CDs are evaluated using:

- CD\_M1: the number of their journal papers;
- CD\_M2: the total number of citations received;
- CD\_M3: their h-index.

Citation-based metrics are used to evaluate predominantly scientific RFs, for which reliable citation databases exist: all RFs in the first nine SAs (01-09), with the exception of the RFs 08/C1, 08/D1, 08/E1, 08/E2, 08/F1 and the four RFs in Psychology (11/E). While, candidates applying to NDs are evaluated using:

- ND\_M1: number of their journal papers and book chapters;
- ND\_M2: number of their papers published on *Class A* journals<sup>3</sup>;
- ND\_M3: number of their published books.

Non-citation-based metrics are applied to predominantly humanistic RFs, for which no sufficiently complete citation database exists: the last five SAs (10-14) with the exceptions described above.

In this first step of the evaluation, candidates' metrics are expected to exceed two out of the three thresholds in their RF. Successively, the candidate's maturity is evaluated based on their CV. The aforementioned metrics are computed for each candidate, taking into consideration only publications that are less than 15 years old for candidates to the role of FP and 10 years old for candidates to the role of AP. This process utilizes data retrieved from Scopus and Web of Science and is conducted by ANVUR. ANVUR also sets thresholds for each metric by RF. Normalization based on the scholars' scientific age (the number of years since the first publication) is used to compute the metrics, for both CDs' and NDs' candidates in all of the considered sessions of the NSQ.

## Methods and materials

This section introduces all the methods and material used for our study. All the data gathered and the software developed for our work are available in (Bologna et al., 2021) and in the GitHub repository <https://github.com/sosgang/bond>.

---

<sup>3</sup> The List of Class A journals is periodically released by the National Agency for the Assessment of Universities and Research (in Italian, Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca, or ANVUR), available at <https://www.anvur.it/attivita/classificazione-delle-riviste/classificazione-delle-riviste-ai-fini-dellabilitazione-scientifica-nazionale/elenchi-di-riviste-scientifiche-e-di-classe-a/> (Last accessed 13 January 2021).

**Table 1** The basic metrics extracted from the list of publications of each candidate

<i>ID</i>	Description
Cand	Overall number of publications authored by the candidate found in the open sources of use
Books	Number of books authored by the candidate
Articles	Number of journal articles authored by the candidate
Other_pubbs	Number of other kinds of publications (e.g. proceedings articles and workshop papers) authored by the candidate
Co-au	Number of publications authored by both the candidate and at least one member of the commission

## Citation Network Analysis and Metrics

The main idea of our work is to calculate some metrics on the NSQ candidates' research production, exclusively from open data, and to study if these metrics could give insights on the peer-review process.

The metrics we take into account are organized in two groups. Table 1 shows the first one. It contains some well-established measures to assess the publication success and academic activity of scholars, thus they do not require further explanation.

The last metric in Table 1 (i.e. *co-au*) deserves some discussion. It is meant to investigate if the relation of co-authorship between the candidates and the commission members has played some role in the evaluation process. The idea of studying the co-authorship network between candidates and evaluators is not new. Bagues et al. (2019) and Martini et al. (2022) even applied it to the Italian NSQ studying how these connections relate to the potential candidates' decision to apply and their success in the NSQ. The authors extracted the co-authorship relations from all the publications submitted to the first round of the NSQ in 2012. Then, they measured the distance between the candidates and the commission members. They also considered two researchers as connected if they are affiliated to the same institution. The conclusion was that "the applicants tend to receive more favorable evaluations from connected evaluators". The authors performed a similar analysis to the Spanish Qualification (Zinovyeva & Bagues, 2015). In this work, they also took the connection *<PhD Student-Advisor>* into account. The findings are very similar to the Italian case and, even there, the "social ties" proved to have a great impact on the peer-review process.

Many other experiments have been performed with similar results in other contexts. Abramo et al. (Abramo et al., 2015a, 2015b) studied career advancements in the Italian academia. They considered the submissions in the Associate Professor competitions in 2008 and performed a statistical analysis to identify which factors have most determined the success of the candidates. They considered the research production (number of publications, citations, journals, etc.) but also the co-authorship and co-affiliation links between candidates and evaluators. The latter (same university) proved to be the most influential parameter. Several research groups also studied if co-authorship and co-affiliation connections increase the possibility of getting articles published in journals (Brogaard et al., 2014; Colussi, 2018; Dondio et al., 2019; Teplitskiy et al., 2018) or being appointed as journal editor (Miniaci & Pezzoni, 2020) or even receiving more research funds (Ebadi & Schiffauerova, 2015).

To the best of our knowledge, this is the first work that investigates citations as connectors between candidates and evaluators. Our other metrics, in fact, are calculated on the

**Table 2** The metrics extracted from the citation networks with a short description

ID	Description
cand_comm	Number of citations going from a candidate’s publication to a publication authored by at least one member of the commission
comm_cand	Number of citations going from a publication authored by at least one member of the commission to a candidate’s publication
BC (bibliographic coupling)	Number of publications cited by both a publication authored by the candidate and a publication authored by at least one member of the commission
CC (co-citation)	Number of publications citing both a publication authored by the candidate and a publication authored by at least one member of the commission
cand_other	Number of other publications (i.e. which are not authored neither by the candidate nor by any member of the commission) cited by a publication authored by the candidate
other_cand	Number of other publications (i.e. which are not authored neither by the candidate nor by any member of the commission) citing a publication authored by the candidate

network of citations across the articles published by the candidates and the commission as reported in Table 2.

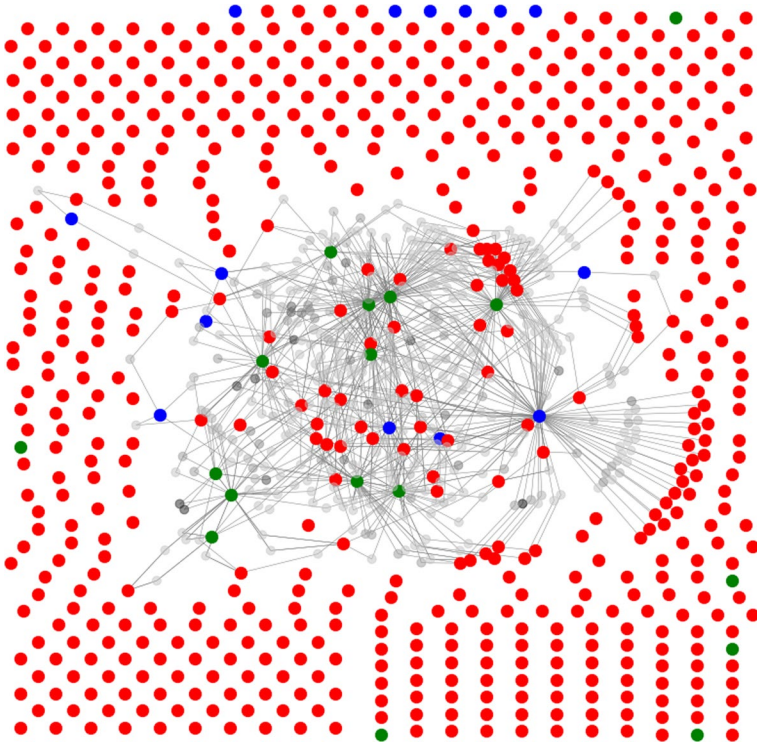
Note that these indicators are not new, but they have never been studied from this perspective. Indeed, citation networks proved to be effective for topic identification and research trends analysis (Chang et al., 2015; Kleminski et al., 2020), as well as for studying interactions among research groups and institutions (Yan & Sugimoto, 2011), for evaluation purposes (Cai et al., 2019) and for visualizing research impact and production (Portenoy et al., 2017).

Here, the presence of citations from/to the publications of the candidates’ and the commission indicates *proximity*. There are actually two forms of proximity embodied by the citations: cognitive and social. From the cognitive perspective, the fact that an author cites another one indicates that they share research interests and activities, they work in the same area and know their respective works. However, there is also a social dimension to take into account: candidates and commission members might work in the same department or university, might collaborate within the same national/international research project, might belong to the same research network, and their citations will probably be influenced by such connections.

This work does not make distinction between these forms of proximity, though some indicators that we took into account can be considered more on the cognitive side—for instance co-citation and bibliographic coupling—while others on the social side—for instance cand\_comm. Section “[Looking at cognitive and social proximity in citations](#)” contains some thoughts about such a distinction and possible directions for discerning between these forces within the citations.

Note also that it might well happen that the citation network of a candidate has a few connections with that of the commission, but the candidate equally gets a positive evaluation in the NSQ. In that case, there is no proximity considering only our metrics but other metrics might have influenced the evaluation.

There is another key point to stress on: our analysis is focused on open data. It might happen that these data are limited for some candidates and the list of their publications is underestimated and thus their citation network does have a very limited tie with that



**Fig. 1** Example of the kind of citation network we build for each candidate. The blue dots represent the candidate's publications, the red dots represent the commission's publications and the green dots represent the publications co-authored by the candidate and at least one member of the commission. The edges represent citation links. Gray nodes represent publications by other authors—neither the candidate nor a member of the commission—that connect a candidate's publication with a publication authored by at least one member of the commission via either bibliographic coupling or co-citation. The color of these nodes visually represents the number of publications that both the candidate's and the commission's publications either cite or are cited by: the darker the node, the higher the number of connecting publications. (Color figure online)

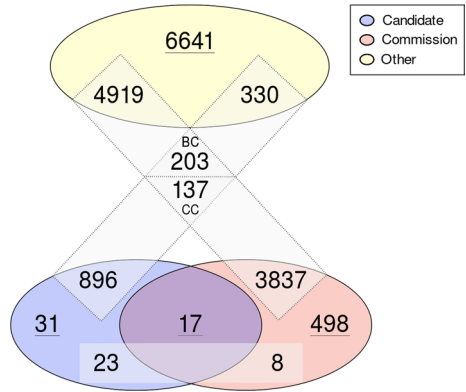
of the commission. In that case, the candidates have probably succeeded in the NSQ even if our data would suggest a different outcome.

To give readers a clearer idea of our approach, Fig. 1 shows one example of the citation network we built for each candidate. A graph is created in which the nodes represent publications and the edges represent citation links. Note that we do not use arrows for the direction of the citations, just to not overload the picture, but we take it into account when computing the metrics. Different colors are used to indicate if a publication is by the candidate (blue) or the commission member (red) or co-authored by the candidate and at least one member of the commission (green).

The gray nodes indicate the publications of other authors—neither the candidate nor any member of the commission—that cite or are cited by a publication of the candidate or a member of the commission. These citations, in fact, contribute to the values of co-citations and bibliographic coupling. To make this evident, we colored these nodes depending on the CC and BC parameters: the more these values are high, the more the color is dark.



**Fig. 2** An alternative visualization for the co-authorship and citation network of Fig. 1. (Color figure online)



In the example in Fig. 1, the candidate has a strong co-authorship connection with the commission members (half of her/his articles are co-authored with at least one member of the commission) but also a lot of candidate’s and commission members’ publications are closely-coupled by citation relationships, as confirmed by the dense network in the center of the diagram. Indeed, the candidate succeeded in the NSQ evaluation.

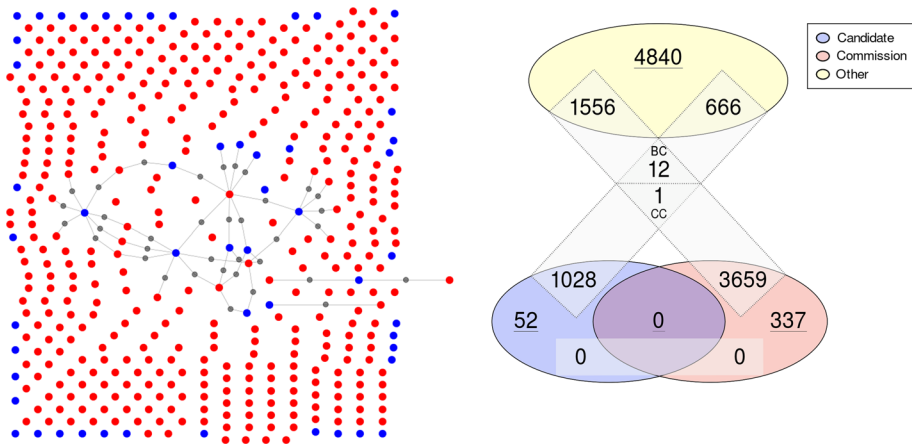
Figure 2 shows the actual values of the metrics, in an alternative visualization loosely-based on Venn diagrams. The three colored sets represent the publications of the candidate (in light blue), of the commission (in pink) and of other authors (in light yellow, useful to calculate co-citation and bibliographic coupling). The set of publications co-authored by the candidate and at least one member of the commission are colored in violet, as the intersection of the two. The number of publications in each set is underlined while the numbers of outgoing/incoming citations among these sets are not; the grey rectangle is used to connect these values from and to each group. In this case, the candidate published 31 articles, including 17 co-authored with commission members, who cumulatively had 498 publications. The candidate’s articles were also cited 8 times by the commission, while she/he cited 23 times articles authored by at least one member of the commission.

Figure 3 shows a different case, with both visualizations. The candidate has co-authored no articles with any of the members of commission, even if she/he has a good research record with 52 publications which received 666 citations. There is no citation from and to the candidate’s articles and the articles authored by at least one member of the commission. Note that the data about the commission are different from the previous diagrams, since the two cases are from two different Recruitment Fields. In this case the candidate failed the NSQ.

We built all the citation networks and computed our bibliographic metrics using exclusively open data, in order to test whether quantitative citation-based indicators can deepen our understanding of peer-reviewed NSQ evaluation. The following subsections describe in detail the data collection process and the sources we used.

**Selection of Recruiting Fields and Data**

For the purposes of this study, we considered the bibliographic data of the candidates and commissions that applied to the Recruiting Fields 10/G1, *Historical and General Linguistics*, and 13/D4, *Mathematical Methods of Economics, Finance and Actuarial Sciences* in the 2016, 2017, and 2018 sessions of the NSQ. In order to choose which 2 RFs we should



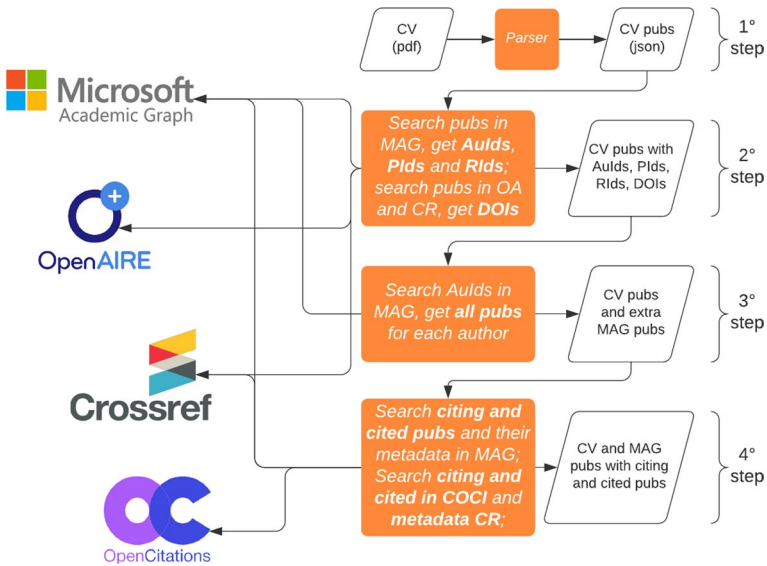
**Fig. 3** The case of a candidate with no direct connection with the commission members and a good research record. (Color figure online)

focus on, we collected all the bibliographic information related to candidates to the role of AP of the 2016 NSQ session and compared the coverage of all the RFs in Microsoft Academic Graph (referred to as MAG). Since the majority of NDs are humanistic disciplines, we decided to base this comparison on the bibliographic data found in MAG. Indeed, previous studies have demonstrated that it has a better coverage of such disciplines (Martín-Martín et al., 2020).

For each candidate, we calculated the percentage of publications found in MAG by dividing the number of all publications in MAG by the number of publications listed in his/her CV submitted to the NSQ. Then, for each RF, we calculated the median percentage. In this comparison, coverage percentages can go well above 100%, as we also considered any additional publication by the same author that we retrieved from MAG but was not included in the candidate’s CV. Candidates can decide not to list all of their academic production on their CV, but only their most recent or relevant publications. However, these remaining ones can be found in MAG and added to the count. We decided to also consider these additional publications since all of a candidate’s bibliographic and citation data is useful for the calculation of our metrics. Among the top 15 RFs in terms of median percentage of publications found in MAG, 10/G1 and 13/D4 were the best covered two RFs that do not belong to the same SA. 13/D4 being the first RF in terms of coverage with 269% publications found in MAG and 10/G1 being the 15th with 156%. All the remaining RFs from the 2nd to the 14th position belonged to SA 13.

**Sources**

In October 2020, we collected open citation data making use of four sources. The first dataset we used is Microsoft Academic Graph (referred to as MAG here), which results from the efforts of the Microsoft Academic Search (MAS) project. This dataset is updated biweekly, is distributed under an open data license for research and commercial applications and is accessible using the Project Academic Knowledge API (<https://docs.microsoft.com/en-us/academic-services/project-academic-knowledge/introduction>) (Wang et al., 2020).



**Fig. 4** Data collection process stages with resources and outcomes. MAG, OA, CR, and COCI stand for Microsoft Academic Graph, OpenAIRE, Crossref and OpenCitations Index of Crossref Open DOI-to-DOI Citations. Auld, Pld, and Rld stand for Author Id, Paper Id, and Reference Id respectively. The first identifies the author, the second identifies the entity and the third is a list of the Paper Ids of all cited entities

The second dataset we used is the OpenAIRE Graph (referred to as OA here), which includes information about objects of the scholarly communication life-cycle (publications, research data, research software, projects, organizations, etc.) and semantic links among them. It is created bi-monthly, and is freely accessible for scholarly communication and research analytics using the OpenAIRE Graph Access API (<http://api.openaire.eu/>) (Rettingberg and Schmidt, 2012).

The third is Crossref (referred to as CR here), which was born as a nonprofit membership association among publishers to promote collaboration to speed research and innovation. The dataset is fully curated and governed by the members as they autonomously provide their publications’ DOIs, webpages, metadata, and update the information whenever it is necessary. The metadata are available through a number of APIs, including REST API (<https://github.com/CrossRef/rest-api-doc>) and OAI-PMH (Hendricks et al., 2020).

Lastly, we also relied on COCI, the OpenCitations Index of Crossref Open DOI-to-DOI.

Citations (Heibi et al., 2019), made available by OpenCitations (Peroni & Shotton, 2020). COCI only stores the citation links between the citing and cited bibliographic entities identified by their DOIs and enables their metadata to be retrieved from other sources using the corresponding REST API (<https://opencitations.net/index/coci/api/v1>).

**Collection process**

In order to calculate citation-based metrics that quantified the overlap between the candidate and the commission’s citation networks, we needed to collect the bibliographic and citation data of both the candidates and the commissions. To achieve this result, we used the four step procedure shown in Fig. 4 and described below.

First, we extracted the publications' metadata. For candidates, we extracted the metadata of each publication listed in the candidates' resumes. In particular, the metadata contained in the resumes available in PDF format were stored into JSON files. In the resumes, the candidate's publications are listed under two sections: "publications" and "publications for indicators". The first are considered in the peer-review phase. Whereas, the second are utilized to compute the metrics used in the preliminary phase of the NSQ. Since there is often a partial overlap between these two lists, publications were disambiguated based on DOI, when present, or title and year correspondence. To ensure matching, space, accents, and punctuation were removed from the publications' titles. For the commissions, we manually compiled a list of publications and searched for their metadata for each member. Since all commissions' members are professors, we searched for their IRIS web pages<sup>4</sup> in the browser. For each member it was possible to find their IRIS publications' page inside their university's IRIS website. The interface allows for the download of the publication's metadata in Excel format. These files were then combined into a single CSV which stored all the publication data of the commissions.

Secondly, we searched for each candidate's Author Ids in MAG and for each article's DOI in OA and CR. In MAG, each author may be assigned with multiple Author Ids due to author disambiguation issues. Therefore, we queried MAG for each article in order to retrieve as many Author Ids as possible. Each original CV article was searched for by title and year or, when present, by DOI, using the *Evaluate* method of the MAG REST API. Academic entities returned by DOI query were considered valid when there was exact lowercase correspondence between the original and found DOI – since DOIs are case insensitive identifiers. In the case of title and year queries, returned academic entities were matched to the original articles based on:

- title correspondence (after having removed punctuation, accents and spaces from the titles);
- publication date (a two year margin is permitted);
- author's surname and author's name (only in case two or more authors shared the same surname).

Once the correspondence between the returned and the original academic entity had been validated, the following attributes were retrieved from MAG: Author Id, Paper Id, and Reference Ids. The first identifies the author, the second identifies the entity and the third is a list of the Paper Ids of all cited entities. The remaining publications that were not found in MAG and were not associated with a DOI were searched for in OA by title's keywords, author's surname and publication date. Keywords were chosen by selecting the first six words in the title that were not stopwords. Since OA strictly returns publications that perfectly match with the attributes specified in the query, the first returned entity was considered valid and the corresponding DOI was retrieved. Lastly, publications that had not been found in OA and did not have a DOI were searched for in CR by title's keywords and surname. We considered only the first four results and we calculated three similarity scores for each. Score A was calculated based on how close the publication date was to the original one: 3 points were given if the new publication had the same date of the original; 2 points if the new publication's date was one year apart from the original's; 1 point if the

---

<sup>4</sup> IRIS is an IT solution that supports Italian universities in collecting, storing and managing research data.

**Table 3** Dataset specifics

500 unique applications
433 unique candidates
10 unique commission experts
15,330 candidate publications listed in their CVs
11,617 candidate CV publications found in open data sets
15,753 candidate publications found in open data publications
865 commission publications found in open datasets

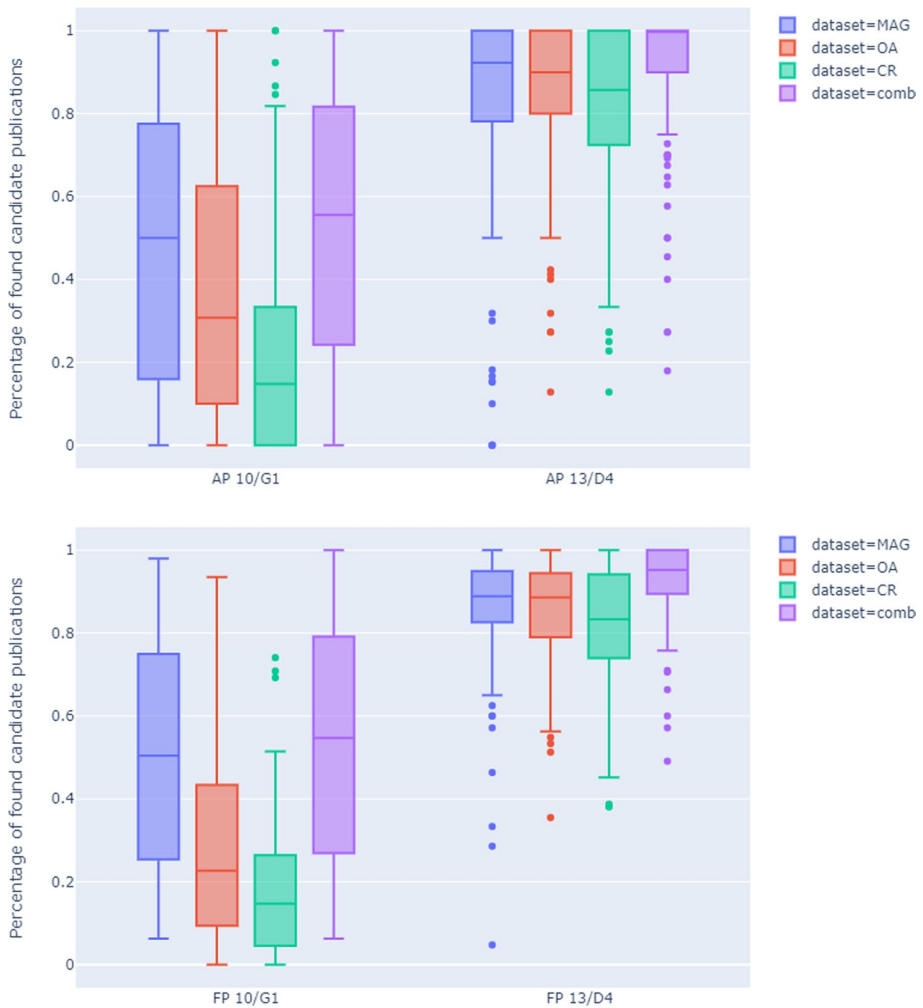
new publication’s date was two years apart from the original’s. Score B was based on the surname and name correspondence: 2 points were given if the new publication’s author had the same name and surname of the original’s; 1 point was given if the new publication’s author had the same name and initial of the original’s. Score C was the Levenshtein distance between returned title and original title. The results were then ranked by C score. If the result with the highest C score has a C score greater than 0.8 and has A and B scores equal to or greater than 1, then the entity is considered valid and the corresponding DOI is retrieved. This matching solution was inspired by Visser, van Eck and Waltman’s own matching process (2020).

Thirdly, MAG was queried by each Author Id and the returned entities that were not already contained in the CV were added. Each returned entity was compared to the CV publications by Paper Id, DOI or title and year. If the new publication matched the present one, the missing information would be added and the new publication would be discarded. In case the new entity did not match any previous publication, it would be included in the JSON file under a different category than the publications originally present in the CV. It would be categorized as an extra publication found in MAG.

In the fourth step of this procedure, the metadata of the citing and cited publications were retrieved. For those publications having at least one Reference Id, the metadata of the entities cited by that publication was collected by querying MAG by Paper Id using each Reference Id. Viceversa, citing entities metadata was returned if querying MAG by Reference Id using the publication Paper Id. To ensure the best coverage of citation data, all the publications associated with a DOI were searched in COCI to collect the DOIs of the citing and cited publications. These were then queried in CR to retrieve their metadata. Given that the same citing or cited publication could be retrieved from both MAG and COCI, the citing and cited articles were disambiguated by DOI to prevent repetitions.

As summarised in Table 3, 500 unique NSQ applications spread over 5 terms from 433 unique candidates were considered. The number of unique applications is greater than the number of unique candidates due to the fact that candidates can re-apply to the NSQ if rejected. Overall we obtained 15,753 publications from MAG, OA, CR and COCI. In particular, we found 11,617 of the original 15,330 publications listed in the candidates CVs.

Finally, applications were divided into three sections, A, B or C, depending on how many of the original CV publications were retrieved from the mentioned open access datasets. Applications were placed in section A when more than 15 of the publications originally listed in the CV were retrieved, or, otherwise, 70% of the CV publications were retrieved. Applications were placed in section B when less than 70% of their CV publications were found but additional publications not listed in the CV were extracted from MAG to reach a total number of retrieved publications comparable to the original one. Applications were placed in section C otherwise. This categorization gives us the ability to decide



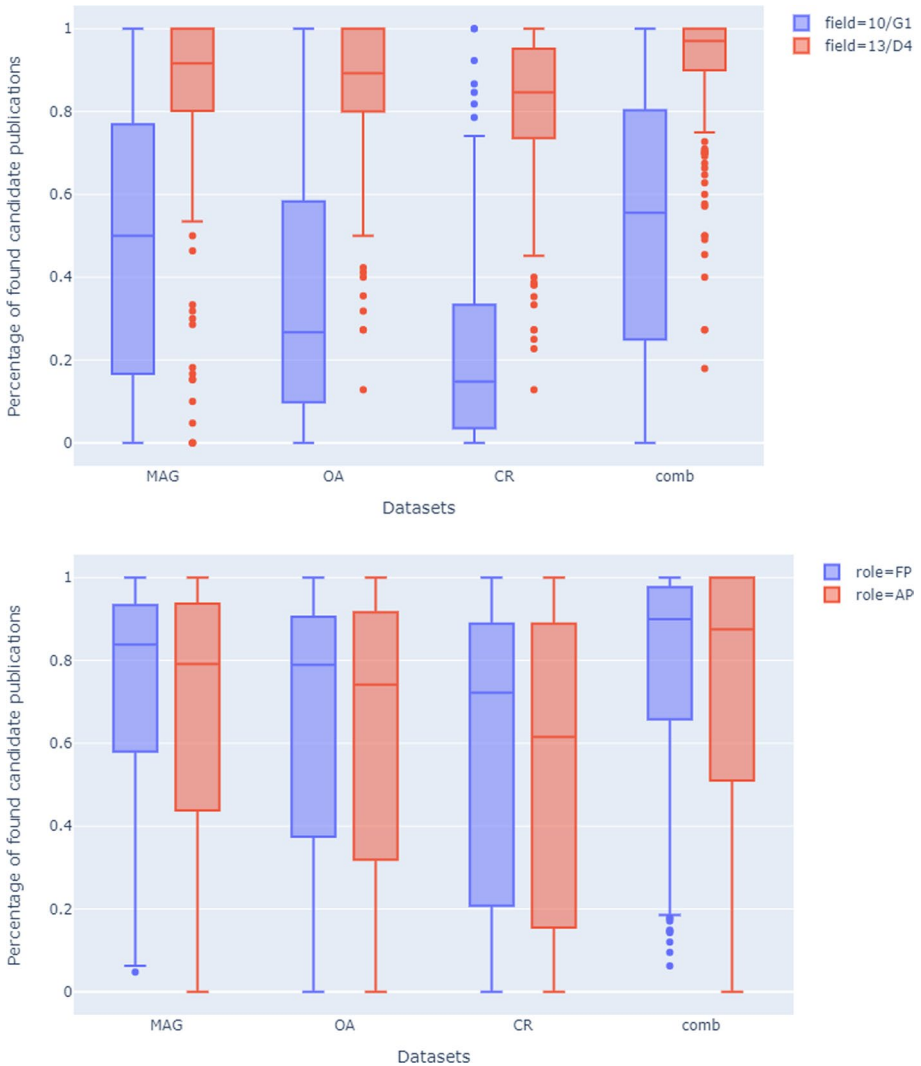
**Fig. 5** Boxplots representing the coverage of each dataset for each role and field. FP and AP stand for Full Professor and Associate Professor respectively. (Color figure online)

whether to include applications that are not greatly covered in our datasets of use in our experiment, described in the following section.

**Coverage in open access datasets: MAG, OpenAIRE and Crossref**

Before running the machine learning classifiers, we examined the coverage of the publications of the candidates for the RFs 10/G1 and 13/D4 in the three datasets we took into account: MAG, OpenAIRE and Crossref. This helped us to better understand our results and to identify some weaknesses of these datasets.

To do so, we queried the datasets for each unique publication listed in the CV of each candidate. We then calculated the percentage of unique CV publications found in



**Fig. 6** Boxplots representing the coverage of each dataset for role and each field. FP and AP stand for Full Professor and Associate Professor respectively. (Color figure online)

each dataset over the total number of unique publications in the CV. As shown in Fig. 5, among the three datasets, the one with the best coverage of both sections and fields is MAG. However, combining the three datasets yields the best result, which means that some bibliographic entities included in other datasets were not included in MAG.

In the case of the RF 13/D4, MAG’s coverage is slightly higher or equal to that of OA and CR. Whereas for 10/G1, MAG’s coverage almost doubles that of OA and more than doubles that of CR. Despite both being categorized as a ND, 10/G1 is more traditionally humanistic than 13/D4: the first is interested in linguistics, while the second

in mathematical models. MAG's better coverage of publications and citations in the Humanities is also confirmed by the literature on the topic (Martín-Martín et al., 2020).

Interestingly, the datasets' coverage varies more greatly across the two RFs than across the two academic roles (i.e. FP and AP), as shown in Fig. 6. Hence, coverage of a candidate's publications is not affected by what role the candidate is applying for, but rather by their field of study. This indicates that the datasets cover scientific fields better than humanistic ones.

It is worth pointing out that these percentages are evidently lower than those presented in Section “[Selection of Recruiting Fields and Data](#)” because any additional publication to those in the CV found in MAG was not taken into consideration in this analysis, in order to ensure a fair comparison between the datasets.

## Machine Learning Classifiers

The aim of the analysis we present is to investigate whether citation-centric metrics can provide insights on the human evaluation in the NSQ of NDs. To this end, we designed an experiment based on computational methods and machine learning techniques. In particular, we employ two different machine learning classifiers: SVM and Decision Trees. These classifiers have been recently used in many bibliometric studies, demonstrating their effectiveness in different contexts. For instance, SVM has been used to predict academic career outcomes (Tregellas et al., 2018) and the results of evaluation procedures of academics (Nuzzolese et al., 2019; Poggi et al., 2019) based on bibliometrics indicators, to perform citation analyses for identifying instrumental citations (Fu et al., 2013), to recognize in advance articles that will become relevant and breakthrough (Savov et al., 2020). Also Decision Trees have been extensively used, for example to measure scientific knowledge flows using citation context analysis (Hassan et al., 2018), to estimate and predict article citation count based on bibliometric data (Fu & Aliferis, 2010) and text analysis techniques (Ibáñez et al., 2009).

Our experiment is based on a comparison of the behaviors of the metrics introduced in Tables 1 and 2, and those provided by ANVUR, described in Section “[The National Scientific Qualification: how it works](#)”. These metrics are used as features of classification algorithms for automatically discerning between candidates who passed the NSQ and who did not. This is a binary classification problem since we have two mutually exclusive classes (i.e. one composed of candidates who passed the NSQ, and the other of candidates who failed) and, for each individual in the population, we attempt to predict which class he/she belongs to.

We base our analysis on a variable ranking method from feature selection theory (Guyon & Elisseeff, 2003) to determine the metrics that are significant for the current classification problem, and hence are useful to discriminate between candidates passing the NSQ or not, and to identify the metrics that are either redundant or irrelevant – and that can, thus, be removed without incurring much loss of information (Bermingham et al., 2015).

The first step of the experiment consists in splitting the data about the NSQ candidates into a *training set*, composed of the information about 283 candidates from the first four NSQ terms (56.6% of the total), and a *test set*, composed of the information on the remaining 217 candidates from the last term (43.4% of the total). In this way we use around half of the data to train the algorithms, which are tested on the remaining data.

For each academic recruitment field (i.e. 10/G1 and 13/D4) and academic role (i.e. FP and AP), we have two nominal classes to use for the two classification algorithms (SVM



**Table 4** Dataset for the classification experiment

	FP			AP		
	Passed	Failed	Total	Passed	Failed	Total
10/G1 (training)	22	13	35	29	35	64
10/G1 (test)	13	22	35	16	23	39
13/D4 (training)	35	41	76	22	17	39
13/D4 (test)	53	55	108	24	36	60

**Table 5** Total number of classifiers to compute for the experiment

Variable description	Number of instances/counter
Combinations of metrics	16,383
Recruitment fields (i.e. 10/G1 and 13/D4)	2
Academic roles (i.e. FP and AP)	2
Input data coverages (i.e. A, AB and ABC)	3
Classification algorithm configurations (i.e. Decision Trees and SVM with regularization parameter C set to 1.0, 0.5, 0.1 and 0.02)	5
Total number of classifiers to compute	982,980

and Decision Trees) that come from the NSQ procedure, i.e. “Passed” or “Failed”. Table 4 reports the details about the dataset used for experimenting with the classifiers.

We test each possible subset of metrics and find those which minimizes the error rate and lead to good predictions. We start by computing all the possible combinations of metrics that can be obtained considering the eleven metrics based on open datasets and the three indicators provided by ANVUR. The number of combinations can be computed by the following expression:

$$\sum_{k=1}^{14} C_{14,k} = \sum_{k=1}^{14} \frac{n!}{k!(n-k)!} = 16,383.$$

For each of the two academic recruitment fields (i.e. 10/G1 and 13/D4), two academic roles (i.e. FP and AP), and three input data coverages (i.e. A, AB and ABC), all the combinations of metrics have been used as features of the classification algorithms for automatically predicting the results of the NSQ. We consider Decision Trees and SVM with regularization parameter  $C^5$  set to 1.0, 0.5, 0.1 and 0.02, resulting in five classification

<sup>5</sup> SVMs work by finding data points of different classes (i.e. the support-vectors) and drawing boundaries between them (i.e. the hyperplanes).  $C$  behaves as a regularization parameter that trades off the correct classification rate of training examples against the maximization of the decision function’s margin. For larger values of  $C$ , a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower  $C$  will encourage a larger margin for future data at the cost of training accuracy. Since there is no general rule in choosing the  $C$  parameter as this depends on the dataset in use, we decide to test different values ranging from large to smaller ones for  $C$  in our experiment. An interesting discussion with examples on the topic is available at <https://stats.stackexchange.com/questions/31066/what-is-the-influence-of-c-in-svms-with-linear-kernel#answer-159051>.

algorithm configurations. The total number of classifiers computed is 982,980, as summarized in Table 5.

All these classifiers are trained on the data about candidates to the first four terms of the NSQ, and are tested on the candidates of the last term. An oversampling technique with stratification has been applied to the minority class for managing imbalanced classes,<sup>6</sup> i.e. in cases where the distribution of examples across the known classes is not equal (Krawczyk, 2016). In particular, we used the approach described in (Japkowicz, 2000) that prescribes to resample the smaller class (by creating synthetic instances chosen at random from the minority class) until it consisted of as many samples as the majority class. The Python code developed for this experiment is based on the Pandas<sup>7</sup> and Scikit-learn<sup>8</sup> libraries, and is available at <https://github.com/sosgang/bond>.

Of all the computed classifiers, we consider only those with good discrimination abilities (i.e. those whose weighted average F1-score is at least 0.7, and which therefore lead to good classification performances). Since the classifiers were calculated using all possible combinations of metrics, each metric was used as a feature by half of the classifiers. We then count how many times each of the fourteen metrics has been used as a feature by the classifiers with good discrimination abilities, and identify as significant for the current classification problem those that have been used by more than 50% of these classifiers. We also identify the metrics that are not relevant or redundant as those that are used as features by a low fraction (i.e. less than 35%) of the classifiers with high classification performances.

The results of this experiment allow us to understand what metrics provide relevant information that can be used by machine learning algorithms to replicate the decisions of the NSQ commissions. The hypothesis here is that if a subset of metrics is used by most of the classifiers with good performances, it signals that such metrics are important proxies of the human evaluation process and outcome.

## Results

The results of the 982,980 computed classifiers described in the previous section, in terms of precision, recall and F1-score are available in (Bologna et al., 2021) and in the GitHub repository <https://github.com/sosgang/bond>. Of all the computed classifiers, 4,217 have good discrimination abilities, with F1-scores spanning from 0.700 (i.e. the threshold) to 0.807 (i.e. the F1-score of the best classifier). The number of times each of the fourteen metrics has been used as a feature by these classifiers is reported in Table 6, together with their relative percentages. Table 6 also reports the results for each recruitment field and academic role, allowing us to perform a more specific analysis and get insights into the different relevance of metrics in the two disciplines and the two academic roles. It is however

---

<sup>6</sup> Machine learning algorithms are designed to improve accuracy by reducing the overall error. This can be a problem when classes are strongly imbalanced, since in these cases machine learning classifiers tend to be very biased and inaccurate towards the minority class. For instance, if we had two classes A and B and a population of 100 individuals of which 95 belong to class A and the remaining five to class B, a silly classifier predicting always class A would have an overall accuracy of 95%, but an error rate for class B of 100%. This behavior is an issue in many scenarios, for instance anomaly detection, identification of fraudulent transactions in banks, identification of rare diseases, etc., and many techniques have been developed for solving or mitigating this problem.

<sup>7</sup> <https://pandas.pydata.org/>.

<sup>8</sup> <https://scikit-learn.org/>.

**Table 6** The results of the experiment for each recruitment field and academic role

			METRICS BASED ON OPEN DATASETS														ANVUR METRICS		
			CITATION-BASED METRICS							NON-CITATION-BASED METRICS							NON-CITATION-BASED METRICS		
			and_comm	comm_cand	BC	CC	cand_other	other_cand	cand	co-au	books	articles	other_pubbs	ND_M1	ND_M2	ND_M3			
10/G1 FP	feature count	604	604	858	695	384	452	804	512	536	136	600	188	72	487				
	feature percentage	<b>50.04%</b>	<b>50.04%</b>	<b>71.09%</b>	<b>57.58%</b>	31.81%	37.45%	<b>66.61%</b>	42.42%	44.41%	11.27%	49.71%	15.58%	5.97%	40.35%				
10/G1 AP	feature count	1454	1454	1080	2883	883	704	1423	1454	1499	1147	523	758	516	962				
	feature percentage	<b>50.26%</b>	<b>50.26%</b>	37.33%	<b>99.65%</b>	30.52%	24.33%	49.19%	<b>50.26%</b>	<b>51.81%</b>	39.65%	18.08%	26.20%	17.84%	33.25%				
13/D4 FP	feature count	37	37	0	20	32	5	50	42	39	0	53	20	0	35				
	feature percentage	<b>52.86%</b>	<b>52.86%</b>	0.00%	28.57%	45.71%	7.14%	<b>71.43%</b>	<b>60.00%</b>	<b>55.71%</b>	0.00%	<b>75.71%</b>	28.57%	0.00%	<b>50.00%</b>				
13/D4 AP	feature count	24	24	7	40	4	32	11	7	39	19	19	24	40	18				
	feature percentage	<b>51.06%</b>	<b>51.06%</b>	14.89%	<b>85.41%</b>	8.51%	<b>68.09%</b>	23.40%	14.89%	<b>82.98%</b>	40.43%	40.43%	<b>51.06%</b>	<b>85.11%</b>	38.30%				

The metrics with high percentages of use are highlighted in bold and italics

important to underline that in 13/D4 the number of classifiers with good discrimination abilities is very low compared the number of classifiers in 10/G1 (i.e. 117 vs 4100).

Among the citation-based metrics, *cand\_comm* and *comm\_cand* have been used in around 50% of the classifiers for both fields and academic roles ( 0.04% in 10/G1 FP, 50.26% in 10/G1 AP, 52.88% in 13/D4 FP and 51.06% in 13/D4 AP for both metrics). Whereas, *BC* has been used in more than half of the classifiers for 10/G1 FP (71.09%), but in very few ones in 10/G1 AP and 13/D4 AP, and in no classifiers in 13/D4 FP. *CC* has a very high percentage of use in both disciplines (99.65% in 10/G1 and 85.11% in 13/D4) at the AP level. Whereas, at the FP level, it has a percentage above average (i.e. 57.58%) in 10/G1 and a very low percentage (i.e. 28.57) in 13/D4. *cand\_other* has low percentages in both fields and both roles, and *other\_cand* is used in more than 50% of the classifiers only in 13/D4 AP (68.09%).

Of the non-citation-based metrics, *cand* is used in 66.61% and 71.43% of the classifiers in both fields for the FP role, whereas it is used in less than 50% of the classifiers for the AP role. Interestingly, *co-au* have above-average percentages in 10/G1 AP and 13/D4 FP (50.26% and 60% respectively), but low percentages in 10/G1 FP and 13/D4 AP. *books* has above average percentages in both fields and both roles, but 10/G1 FP (51.81% 10/G1 AP, 55.71% 13/D4 FP, 82.98% 13/D4 AP). *articles* has low percentages in both fields and both roles, whereas *other\_pubbs* reaches an above average percentage only in 13/D4 FP (75.71%).

All ANVUR metrics are used in less than half of the classifiers for both roles in 10/G1. However, *ND\_M1* and *ND\_M2* have higher-than average percentages for 13/D4 AP (51.06% and 85.11 respectively), and *ND\_M3* is in half of the classifiers for 13/D4 FP (50%).

## Discussion and conclusions

In this work, we explored whether open citation-based metrics relate to the qualitative peer-review of NDs. Specifically, we grounded our analysis on the data of the candidates and commissions that took part in the 2016, 2017, and 2018 terms of the NSQ for the disciplines *Historical and General Linguistics* and *Mathematical Methods of Economics, Finance and Actuarial Sciences*, having Recruiting Fields 10/G1 and 13/D4. We collected the bibliographic metadata and citation data from MAG, OA, Crossref and OpenCitations. We devised and reused a set of bibliographic and citation-based metrics using a top-down process and combining popular metrics (e.g. bibliographic coupling) to others proposed for the purpose of this study to describe the relationship between the candidates and the commission of the NSQ.

The results of the experiment presented in Section “[Results](#)”, based on computational methods and machine learning techniques, provides insights into which metrics are relevant in the human evaluation of different disciplines and academic levels. Our hypothesis is that metrics used by most of the classifiers with good performances to predict the outcome of the NSQ, such metrics are important proxies of the human evaluation process.

First of all, we remark that citation-based metrics register high percentages of use in classifiers with good discrimination abilities in both disciplines and academic roles, i.e. *cand\_comm*, *comm\_cand*, *BC* and *CC*. The relevance of citation-based metrics in both disciplines is the first element supporting our hypothesis.

## Analysis of Citation-Based-Metrics

Regarding our citation-based metrics, *cand\_comm* and *comm\_cand* show an average percentage of use in all disciplines and levels. Since these metrics describe the strength of the citational relationship between the candidate and the commission, their high percentage of use in the classifiers appears to indicate that this relationship does play a role in the peer-review phase of the NSQ. Therefore, it suggests a positive answer to the second part of our research question. Additionally, both *CC* heightened relevance at the AP level for both disciplines and *other\_cand*'s high percentage in 13/D4 AP suggest that being recognized as a member of the same scientific community of the commission, or as a trust-worthy researcher by a third party weighs heavily in the evaluation of junior scholars. Indeed, *CC* is the number of publications citing both the candidate and the commission, and *other\_cand* is the number of citations from others to publications of the candidate. Furthermore, in 10/G1 FP, *BC* has a high percentage of use, and *CC* has an above-average percentage. This suggests that the evaluation of candidates applying to this discipline and academic role is based on the strength of their relationship with the commission and their connections within their scientific community. In order to further investigate *CC* relevance for the academic role AP, we analyzed the structures of the computed Decision Trees.

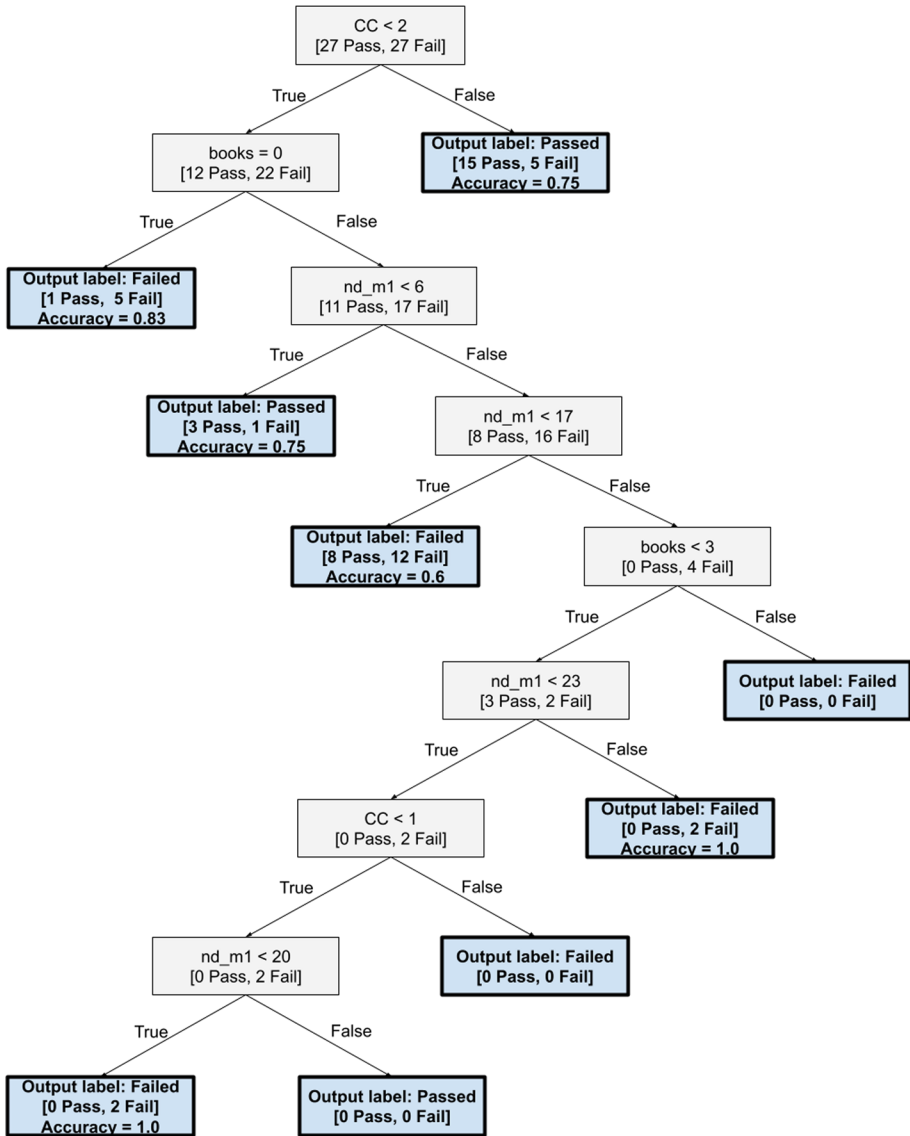
Decision Trees are flowchart-like structures composed of nodes (represented by rectangles) and branches (represented as arrows connecting nodes) showing the flow from question to answer. Each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after evaluating all tests). For instance, Fig. 7 is the Decision Tree generated using the four features *CC*, *books* and *nd\_m1*, and applied to the test set data having coverage A or B. Internal nodes are depicted as gray rectangles, and report the tests and the number of individuals in the test set belonging to each class (i.e. "pass" for candidates who passed the NSQ and "fail" for those who did not) between square brackets. Leaf nodes are depicted as light blue rectangles, and report the output class labels, the actual number of individuals in the test set belonging to each class and the node accuracy.

By analyzing the Decision Tree we observe that the first test splits individuals having *CC* greater than one (right branch) from those having *CC* lower or equal to one (left branch). The former flow into a leaf node (i.e. no other condition is tested), and are labeled as candidates who did not pass the NSQ resulting in an accuracy score of 0.75. The latter flow in the left branch, requiring a series of tests on the metrics *books*, *nd\_m1* and *CC* to discern different cases, and leading to high accuracy values in all of them.

The Decision Trees generated for this sector share similar structures, which emphasizes strong significance for the *CC* metric. In fact, with just one test about the metric *CC* it is possible to correctly label with good accuracy almost half of the individuals in the test set.

## Analysis of Non-Citation-Based Metrics

Non-citation-based metrics also seem relevant in the evaluation process of both disciplines and academic levels. Specifically, *cand* has a high or above-average percentage of use at the FP level in both disciplines, and *books* has high or above-average percentage at the AP level in both disciplines. In 10/G1 AP, *co-au* has an above-average percentage as well. These results suggest that having published a book or having co-authored a publication



**Fig. 7** The decision Tree computed for the discipline 10/G1 and academic level AP for candidates with coverage A or B using the metrics CC, books and nd\_m1

with at least one member of the commission is a discriminating criterion in the evaluation of scholars applying to the role of AP. Whereas, candidates to the academic role of FP are evaluated based on their combined publishing efforts. It is also interesting to notice that our non-citation-based metrics appear to weigh more heavily in the assessment of candidates to 13/D4 than candidates to 10/G1. In 13/D4, not only do *cand* and *books* have higher percentages of use than in 10/G1, but also *co-au*, *books* and *other\_pubbs* have high or above-average percentages at the FP level, while having very low values in 10/G1 FP.

## Analysis of Anvur Metrics

ANVUR non-citation-based metrics are also relevant in the evaluation process of 13/D4. *ND\_M3* (number of books), has an average percentage at the FP level. *ND\_M1* (number of journal papers and book chapters) and *ND\_M2* (number of papers published on *Class A* journals) respectively have a slightly above-average and a high percentage of use at the AP level. Consequently, these values combined with the ones described just above indicate that all aspects of a candidate's academic production are taken into consideration in the evaluation of 13/D4 applicants, especially at the FP level. In 13/D4 FP, *cand* maps the overall number of publications, *co-au* records publications co-authored with at least one member of the commission, *other\_pubbs* tracks the publications that are not articles or books, and *books* and *ND\_M3* measure the number of books in the CV and the overall number of books respectively. Similarly, in 13/D4 AP, *books* measure the number of books, *ND\_M1* maps journal papers and book chapters and *ND\_M2* tracks the number of *Class A* papers.

An additional interesting characteristic of 13/D4 is the very low number of classifiers with an F-measure above 0.700. This could indicate the great heterogeneity of the evaluation process for this specific discipline and role. Moreover, this could mean that the evaluation of the commission concerns aspects that are not captured by bibliometric indicators. Typical examples are the scientific responsibility of international projects, the editorship of international journals, academic awards and fellowships: all these examples attest the maturity of academics, but are inherently non-bibliometric aspects.

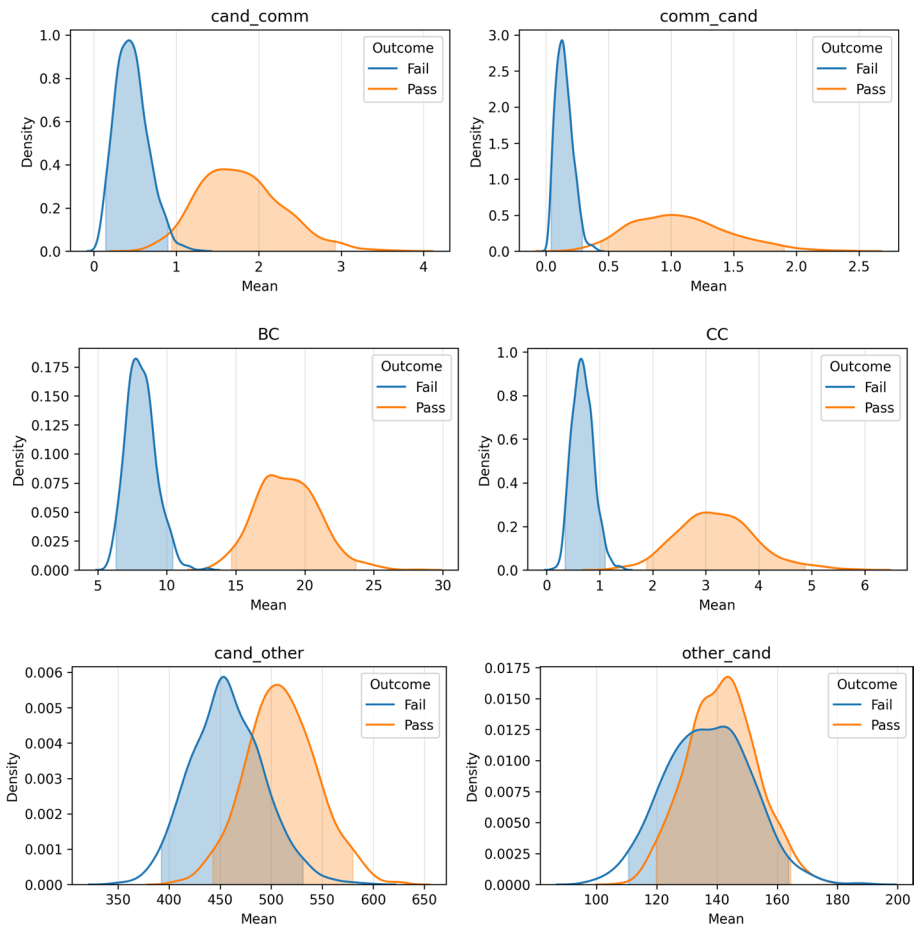
Candidates to the academic role of FP are mostly assessed on their good-standing within their scientific community and overall publishing efforts. Whereas, the evaluation of applicants to the academic role of AP heavily rests on whether a third party confirms their belonging to their scientific community, they authored a book, published a paper in a major journal, and co-authored a publication with the commission. Moreover, although the candidate's publication history is considered also in the evaluation of applicants to 10/G1, it is particularly important in the assessment of 13/D4 applicants.

## Comparing machine learning classifiers with statistical tests

To further support our analysis of the relationship between our metrics and the result of the evaluation, we compare the 95% confidence intervals (CIs) of the bootstrapped means of each of our metrics between candidates that pass the evaluation and candidates that fail the evaluation. Specifically, we split our sample depending on the outcome of the evaluation, bootstrap the means of each bibliographic or citation-based metric, and calculate the confidence intervals of the distributions of means. Visually comparing the CIs provides us with insights regarding which metrics are similar between the candidates that pass and those that fail (and thus the distributions overlap), and which are different (therefore the distributions do not overlap), as shown in Figs. 8, 9, 10.

Using this visual analysis, we notice that the CIs of bootstrapped means of candidates that fail the evaluation do not overlap with CIs of candidates that pass for the following citation-based metrics: *cand\_comm* (97.5% CI Fail=0.90, 2.5% CI Pass=0.92), *comm\_cand* (97.5% CI Fail=0.29, 2.5% CI Pass=0.43), *BC* (97.5% CI Fail=10.43, 2.5% CI Pass=14.65), and *CC* (97.5% CI Fail=1.10, 2.5% CI Pass=1.87) (see Fig. 8).

The CIs overlap for all non-citation-based metrics, as shown in Fig. 9. However, the shape of the two distributions for *co-au* is different than for the other non-citation-based

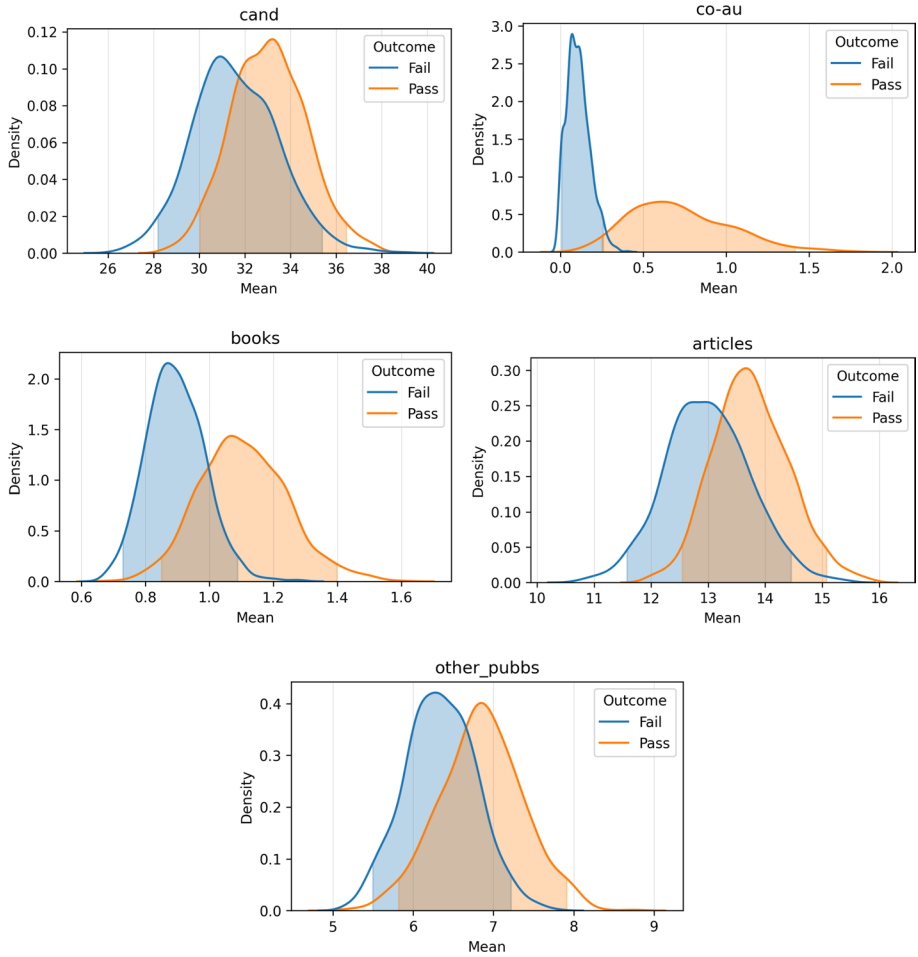


**Fig. 8** Confidence intervals of the bootstrapped means for our citation-based metrics. (Color figure online)

metrics. It resembles more closely to the distributions of the means of the citation-based metrics for which we do not find an overlap. In addition, the CIs for *co-au* barely overlap (97.5% CI Fail=0.26, 2.5% CI Pass=0.24) compared to other metrics. To either support or dismiss that *co-au* values are different for candidates that pass the evaluation and those who fail, we conduct further statistical testing. The results of said tests are reported later in this section. Among ANVUR’s metrics, the CIs do not overlap only in the case of *ND\_M2* (97.5% CI Fail=5.78, 2.5% CI Pass=6.33) (see Fig. 10).

We carry out further statistical analyzes in order to confirm or dismiss that *cand\_comm*, *comm\_cand*, *BC*, *CC*, *ND\_M2*, and potentially *co-au*, are statistically different between candidates that pass and candidates that fail the evaluation. Specifically, we carry out a permutation test and a Mann–Whitney U test for each of our metrics. Using these tests, we can determine whether there is a statistically significant difference between the sample of bibliographic and citation-based metrics of candidates that pass the evaluation and the sample of metrics of candidates that fail the evaluation. We choose these tests because they are both non-parametric and thus do not assume that the samples are normally distributed.

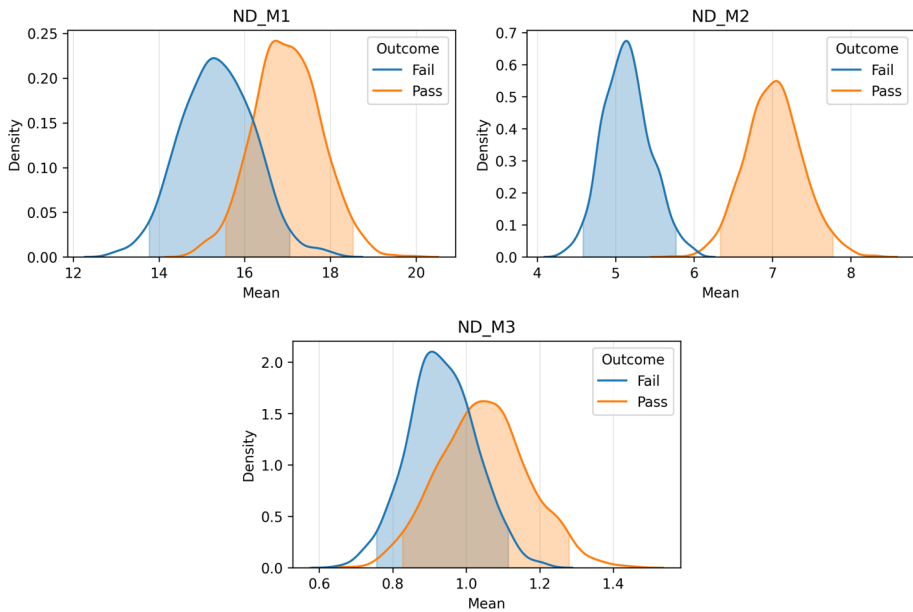




**Fig. 9** Confidence intervals of the bootstrapped means for our non-citation-based metrics. (Color figure online)

Table 7 reports the metrics that are statistically different between our two samples (the candidates that passed the evaluation and the candidates that did not).

In line with our ML experiment’s findings, we find that *cand\_comm*, *comm\_cand*, *BC*, *CC* are statistically different between the two samples. Candidates who pass the NSQ tend to have a higher number of citations to the commission and from the commission (*cand\_comm* and *comm\_cand*), and to share a higher number of citations with the commission (*BC* and *CC*). This further supports our findings regarding the importance of citation-based metrics for the evaluation of both disciplines and academic roles. However, these tests don’t find a statistically meaningful difference between the two samples for *other\_cand*, the number of citations from other scholars to the candidate. Of our non-citation-based metrics, the tests only identify *co-au* as statistically different between the two samples, as candidates who pass the NSQ have a higher number of publications they co-authored with at least one member of the commission. Whereas, our machine learning analysis also identifies the



**Fig. 10** Confidence intervals of the bootstrapped means for ANVUR’s metrics. (Color figure online)

relevance of the number of publications (*cand*) and of books (*books*) authored by the candidate in the evaluation process. Among the ANVUR metrics, *ND\_M2* is the only one that reaches statistical significance between the two samples. After all, *ND\_M2* has the highest percentage among the ANVUR metrics in the results of our analysis.

### Looking at cognitive and social proximity in citations

One of the aspects we discussed at the beginning of this paper is the mixture of cognitive and social aspects in the use of the citations. Citations are complex tools and an author can cite another one for multiple reasons, so that a citation can express different kinds of proximity: *cognitive* (common research interests, application of the same approach, common background, etc.) or *social* (same affiliation, same research group, past collaborations, etc.) or a combination of the two. Our work does not make any assumption on the nature of the citations and the indicators we calculated on the citation network. Though, it is possible to make some distinction. We consider BC (bibliographic coupling) and CC (co-citation) primarily as “cognitive connectors” between the candidates and the commissioners: the fact that a third party is involved—citing or being cited by both—makes it unlikely that the citation derives from a social connection only. The characterization of *cand\_comm* and *cond\_camm* indicators, on the other hand, is much less evident: the presence of a citation from/to the articles of candidates and commissioners might be the result of different forces, either cognitive and social.

Discerning between these forces is a very complex task, since there are no clear boundaries. We run a preliminary experiment on our dataset to start looking at these differences.

**Table 7** Permutation test and Mann–Whitney *U* statistics and their associated *p*-values

Metric	Permutation	Mann–Whitney <i>U</i>
cand_comm	Difference: -1.47 ( <i>p</i> -value: 0.0)	<i>U</i> statistic: 26,915.0 ( <i>p</i> -value: 7.43e-06)
comm_cand	Difference: -0.96 ( <i>p</i> -value: 0.001)	<i>U</i> statistic: 27,580.5 ( <i>p</i> -value: 1.90e-05)
BC (bibliographic coupling)	Difference: -11.86 ( <i>p</i> -value: 0.0)	<i>U</i> statistic: 22,245.0 ( <i>p</i> -value: 2.87e-08)
CC (co-citation)	Difference: -3.00 ( <i>p</i> -value: 0.0)	<i>U</i> statistic: 23,652.5 ( <i>p</i> -value: 1.34e-09)
cand_other	Difference: -57.44 ( <i>p</i> -value: 0.102)	<i>U</i> statistic: 27,587.0 ( <i>p</i> -value: 0.028)
other_cand	Difference: -5.69 ( <i>p</i> -value: 0.372)	<i>U</i> statistic: 26,912.5 ( <i>p</i> -value: 0.009)
cand	Difference: -2.35 ( <i>p</i> -value: 0.166)	<i>U</i> statistic: 27,025.0 ( <i>p</i> -value: 0.011)
co-au	Difference: -0.66 ( <i>p</i> -value: 0.001)	<i>U</i> statistic: 28,821.0 ( <i>p</i> -value: 0.001)
books	Difference: -0.15 ( <i>p</i> -value: 0.16)	<i>U</i> statistic: 30,181.5 ( <i>p</i> -value: 0.523)
articles	Difference: -0.74 ( <i>p</i> -value: 0.245)	<i>U</i> statistic: 26,338.5 ( <i>p</i> -value: 0.003)
other_pubbs	Difference: -0.76 ( <i>p</i> -value: 0.141)	<i>U</i> statistic: 30,135.0 ( <i>p</i> -value: 0.533)
ND_M1	Difference: -1.59 ( <i>p</i> -value: 0.86)	<i>U</i> statistic: 23,054.5 ( <i>p</i> -value: 0.012)
ND_M2	Difference: -1.88 ( <i>p</i> -value: 0.0)	<i>U</i> statistic: 19,304.0 ( <i>p</i> -value: 2.59e-07)
ND_M3	Difference: -0.11 ( <i>p</i> -value: 0.246)	<i>U</i> statistic: 26,575.5 ( <i>p</i> -value: 0.939)

For the permutation tests, the observed difference is calculated subtracting the mean of the sample of candidates who passed from the mean of the sample of candidates who did not pass

This is not meant to be exhaustive—and we believe that such a distinction should be investigated in a dedicated article—but gives us a good starting point for future research.

In particular, we studied the co-affiliation network of the candidates and commissioners in our dataset, with the objective of measuring the correlation between the number of years a candidate worked in the same institution of at least one commissioner and the parameters *cand\_comm* and *cond\_camm* indicators; we also studied the correlation between the co-affiliation and the final outcome in the NSQ. The affiliations were taken from “Cerca Università” (<https://cercauniversita.cineca.it/>) a publicly available service to retrieve information about the Italian Universities, including data about departments, people and courses. The service contains one record for each year in which a given person was affiliated to an Italian University. The same person can then have multiple records, with different affiliations, for her/his career.

Table 8 summarizes the data we were able to collect. We considered all 500 applications of our experiment and 11 commissioners. For 352 applications (about 70%) we were able to find at least one affiliation in an Italian University. The other candidates are missing for two reasons. First of all, it might happen that a candidate applied for the NSQ even if she/

**Table 8** Data about co-affiliations

500 unique applications
11 commissioners (all with Italian affiliation)
352 applications of candidates with Italian affiliations
33 candidates/commissioners in the same department (0.09%)
53 candidates/commissioners in the same University (0.14%)

**Table 9** Normality assessment of co-affiliation indices

Index	Shapiro–Wilk test (W)	<i>p</i> -value
years_co-affiliation_university – successful candidates	0.340	1.193e <sup>-25</sup>
years_co-affiliation_university – unsuccessful candidates	0.193	3.542e <sup>-26</sup>
years_co-affiliation_department – successful candidates	0.475	2.591e <sup>-23</sup>
years_co-affiliation_department – unsuccessful candidates	0.302	1.105e <sup>-24</sup>

he works abroad; in that case there is no record in “Cerca Università” but we can safely conclude that the candidate and the commission member have never worked in the same institution. The second case is a bit more tricky: “Cerca Università” only records people with a tenure track or a permanent position; PhD students, postdocs and adjunct professors are then not listed and considered in our analysis. This is indeed a limitation of the experiment. Nonetheless, the connection between these candidates and the commissioners is weaker than the others: a PhD student, for instance, might work in the same department of a commissioner even without knowing each other; the same might apply for adjunct professors.

A complete analysis of social proximity should also consider these cases but, at this stage, we only focus on permanent positions and co-affiliations and consider 352 candidates for which we have reliable data.

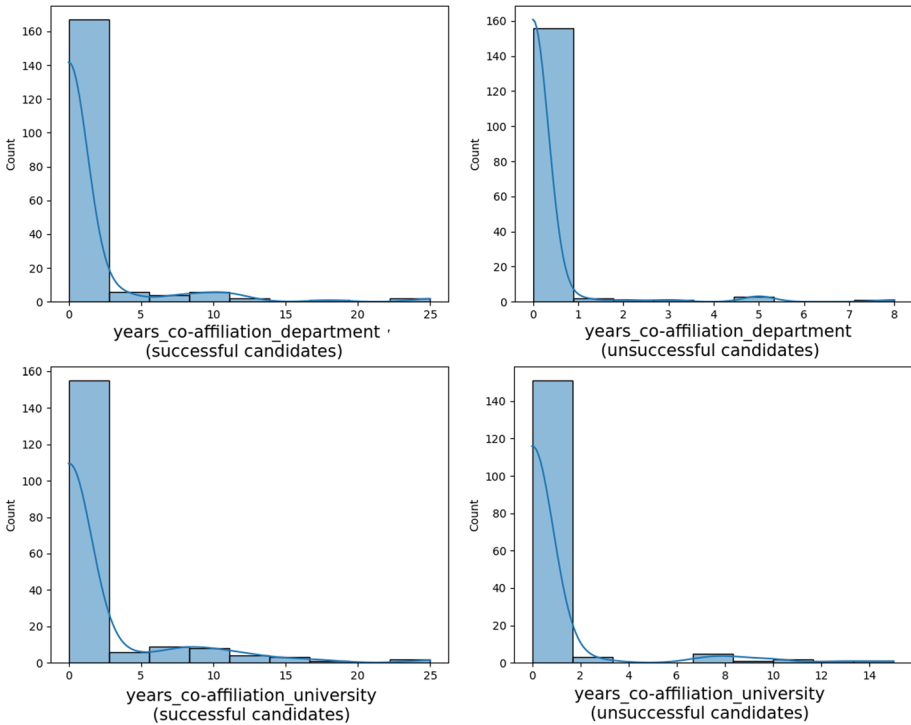
We then searched for co-affiliations, counting the number of years in which each candidate worked in the same institution of a commission member. We considered two cases: (a) the exact same department and (b) different departments but in the same University.

We found that only 33 candidates were affiliated to the same department of at least one commission member (9%), and the number increases to 52 if we consider co-affiliation at University level (14%). These data, though limited, indicate that the influence of such a proximity is limited.

We then measured the percentage of these candidates who achieved a positive outcome in the NSQ, so as to compare that percentage to the same value over the full dataset. As reported before, in total, 48,3% of candidates passed the NSQ in the 10/G1 field and 48,7% in 13/D4.

The result is interesting: these percentages increase significantly when restricting the analysis to the candidates with co-affiliations only. In fact, for 10/G1 the percentage increases up to 66,66% (same department) and 65,21% (same university) against the overall 48,3%. Similar figures were found for the 13/D4 field: the percentage goes up to 77,77% (same department) and 65,52% (same university) against 48,7% (overall).

We also investigated whether the aforementioned co-affiliation indices are related to the results of the NSQ. The approach to answer this question is to compare the NSQ successful



**Fig. 11** Histograms of the two co-affiliation indices for successful and unsuccessful NSQ candidates. (Color figure online)

and unsuccessful candidates through co-affiliation indices. Since the co-affiliation indices do not follow a normal distribution (as shown in Table 9 and Fig. 11), we chose the Wilcoxon–Mann–Whitney test applying the continuity correction in the normal approximation for the *p*-value.

The analysis was carried out on all candidates, regardless of discipline, because the aim of this step is to assess whether the co-affiliation indices differ significantly between the two groups, showing them to be a priori potential explanatory factors for the outcome of the NSQ. Table 10 shows the results of Wilcoxon–Mann–Whitney test performed on each co-affiliation. Both indices tested were significantly different in the two groups, with averages always higher in the successful candidates’ group. These results support the idea of using co-affiliation and other indices measuring the social proximity as explanatory variables of NSQ.

This seems to imply that the co-affiliation relates to the outcome of the evaluation, even if it is not considered explicitly. More experiments are needed to confirm this hypothesis, considering the small amount of data used for this experiment. Note also that other social connections could be in place—for instance, the participation in the same research project or consortium or the co-organization of events or co-editing of journals—and are worth being included in a more exhaustive analysis of social proximity that we plan to integrate with the current work.

**Table 10** Wilcoxon–Mann–Whitney tests for the co-affiliation indices between the NSQ successful and unsuccessful candidates groups. Indices are computed using the data from all candidates regardless of discipline

Index	Successful candidates mean	Unsuccessful candidates mean	W	<i>p</i>
Years_co-affiliation_university	1.106	0.183	16,680	7.803e10 <sup>-3</sup>
Years_co-affiliation_department	1.809	0.646	17,022	7.652e10 <sup>-3</sup>

### Drawing conclusions and future directions

The results of our experiments with ML classifiers draw two orthogonal conclusions. On the one hand, they seem to highlight the existence of a possible bias between applicants to the NSQ and the related committee members observed by analyzing the citational relationships between the researchers, even in those contexts where such citations are not explicitly taken into account into the NSQ process. Indeed, we found that the connections between the citation networks of the candidates and the commissioners are relevant for the peer-review process even for NDs disciplines.

We believe that this is necessarily not a direct causal-effect connection with a negative connotation—i.e. a candidate is supported by a commission member because they have some proximity, being cognitive or social. Rather, this is a consequence of the use of the citations: commissioners are well-known in the community and recognized experts in their field; citing their works—or being cited by these works—also implies being well-connected to the same community and experts in that area.

The second conclusion is about the ML process we took in place. The fact that the evaluation process can be studied by looking at the results of machine learning techniques suggests new applications for these techniques: not just predicting but also analyzing a posteriori the human-based evaluation. In fact, we do not advocate the adoption of machine learning and other artificial intelligence technologies as a mean to replace human judgement in research assessment exercises. Rather, we believe that these technologies enable looking and analysing the current assessment practices from a new perspective by highlighting possible hidden (conscious and unconscious) relations among the different actors involved in the process. For instance, such tools might be used for scanning a process (such as a research assessment exercise) to identify emerging issues within it—for instance influenced by proximity—thus enabling humans to look closely at these issues and, eventually, to find workarounds to mitigate them and improve the process itself.

In future work, we will expand this study by taking into consideration all NDs, and including all CDs as well. Indeed, we intend to continue this investigation by proceeding into two separate, yet related, research directions. On one hand, we will analyze the coverage of both CDs and NDs by open datasets. We will draw comparisons across different NDs and between CDs and NDs. This would give us the opportunity to analyze more thoroughly how coverage of the candidates’ publications vary across disciplines and research topics. On the other hand, we plan to study the relevance of our bibliographic and

citation-based metrics in the qualitative evaluation of NDs and CDs. Thus, we would be able to determine whether citation-based metrics have a larger impact on the qualitative peer-review of CDs in contrast to NDs.

In addition, we also plan to extend the analysis of social proximity we started in this work, by further investigating how cognitive and social proximity correlate on NSQ data, as done by (Boyack & Klavans, 2010) and (Yan & Ding, 2012), who provided very accurate comparisons between these social indicators as tools for research trend analysis and evaluation. We will also investigate community detection techniques, like the Leiden algorithm (Traag et al., 2019) that proved to outperform the well-known Louvain algorithm both in terms of performance and reliable identification of connected (sub-)communities, as well as metrics to measure network centrality (Stephenson & Zelen, 1989) and their application to research collaborations measurement (Zhang et al., 2014).

**Acknowledgements** This research has been supported by the University fund for Research 2020 (FAR) of the University of Modena and Reggio Emilia, and by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017452. A special thank goes to the second anonymous reviewer who gave us very useful comments and helped us to improve a lot the quality of the paper.

## Declarations

**Conflict of interest** The authors did not receive support from any organization for the submitted work. The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2021). *Data and code of the article "Do open citations give insights the qualitative peer-review evaluation in research assessments? An analysis of the Italian National Scientific Qualification."* <https://doi.org/10.5281/ZENODO.4570782>
- Abramo, G., D'Angelo, C. A., & Rosati, F. (2015a). The determinants of academic career advancement: Evidence from Italy. *Science and Public Policy*. <https://doi.org/10.1093/scipol/scu086>
- Abramo, G., D'Angelo, C. A., & Rosati, F. (2015b). Selection committees for academic recruitment: Does gender matter? *Research Evaluation*, 24(4), 392–404. <https://doi.org/10.1093/reseval/rvv019>
- Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2019). Connections in scientific committees and applicants' self-selection: Evidence from a natural randomized experiment. *Labour Economics*, 58, 81–97. <https://doi.org/10.1016/j.labeco.2019.04.005>
- Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review*, 63(6), 829–846. <https://doi.org/10.2307/2657504>
- Bedogni, L., Cabri, G., Martoglia, R., & Poggi, F. (2022). Does the venue of scientific conferences leverage their impact? A large scale study on computer science conferences. *Library Hi Tech*. <https://doi.org/10.1108/LHT-09-2021-0305>
- Birmingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*. <https://doi.org/10.1038/srep10312>

- Bologna, F., Di Iorio, A., Peroni, S., & Poggi, F. (2022). *Open bibliographic data and the Italian National Scientific Qualification: Measuring coverage of academic fields*. In press. [https://doi.org/10.1162/qss\\_a\\_00203](https://doi.org/10.1162/qss_a_00203)
- Bornmann, L. (2016). Scientific revolution in scientometrics: The broadening of impact from citation to societal. *Theories of Informetrics and Scholarly Communication*. <https://doi.org/10.1515/9783110308464-020>
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404. <https://doi.org/10.1002/asi.21419>
- Brogaard, J., Engelberg, J., & Parsons, C. A. (2014). Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics*, 111(1), 251–270. <https://doi.org/10.1016/j.jfineco.2013.10.006>
- Cai, L., Tian, J., Liu, J., Bai, X., Lee, I., Kong, X., & Xia, F. (2019). Scholarly impact assessment: A survey of citation weighting solutions. *Scientometrics*, 118(2), 453–478. <https://doi.org/10.1007/s11192-018-2973-6>
- Chang, Y.-W., Huang, M.-H., & Lin, C.-W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics*, 105(3), 2071–2087. <https://doi.org/10.1007/s11192-015-1762-8>
- Cheberkus, D., & Nazarovets, S. (2019). Ukrainian open index maps local citations. *Nature*, 575(7784), 596–596. <https://doi.org/10.1038/d41586-019-03662-6>
- Chudlarský, T., & Dvořák, J. (2020). Can crossref citations replace web of science for research evaluation? The share of open citations. *Journal of Data and Information Science*. <https://doi.org/10.2478/jdis-2020-0037>
- Colussi, T. (2018). Social ties in academia: A friend is a treasure. *The Review of Economics and Statistics*, 100(1), 45–50. [https://doi.org/10.1162/REST\\_a\\_00666](https://doi.org/10.1162/REST_a_00666)
- Ministerial Decree 159 (2012). Redefinition of scientific disciplines (Rideterminazione dei settori concorsuali)(Prot. n. 159), Gazzetta Ufficiale Serie Generale n.137 del 14/06/2012—Suppl. Ordinario n.119. Retrieved from <https://www.gazzettaufficiale.it/eli/id/2012/06/14/12A06786/sg>
- Di Iorio, A., Poggi, F., & Peroni, S. (2019). *Open data to evaluate academic researchers: An experiment with the Italian Scientific Habilitation*. 17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings, 2, pp. 2133 - 21442. Scopus. Retrived from <http://arxiv.org/abs/1902.03287>
- Dondio, P., Casnici, N., Grimaldo, F., Gilbert, N., & Squazzoni, F. (2019). The “invisible hand” of peer review: The implications of author-referee networks on peer review in a scholarly journal. *Journal of Informetrics*, 13(2), 708–716. <https://doi.org/10.1016/j.joi.2019.03.018>
- Ebadi, A., & Schiffauerova, A. (2015). How to receive more funding for your research? Get connected to the right people! *PLoS ONE*, 10(7), e0133061. <https://doi.org/10.1371/journal.pone.0133061>
- Else, H. (2018). How unpaywall is transforming open science. *Nature*, 560(7718), 290–291. <https://doi.org/10.1038/d41586-018-05968-3>
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*. <https://doi.org/10.1126/science.aao0185>
- Fox, M. F. (2020). Gender, science, and academic rank: Key issues and approaches. *Quantitative Science Studies*, 1(3), 1001–1006. [https://doi.org/10.1162/qss\\_a\\_00057](https://doi.org/10.1162/qss_a_00057)
- Frenken, K. (2020). Geography of scientific knowledge: A proximity approach. *Quantitative Science Studies*, 1(3), 1007–1016. [https://doi.org/10.1162/qss\\_a\\_00058](https://doi.org/10.1162/qss_a_00058)
- Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, 85(1), 257–270. <https://doi.org/10.1007/s11192-010-0160-5>
- Fu, L. D., Aphinyanaphongs, Y., & Aliferis, C. F. (2013). Computer models for identifying instrumental citations in the biomedical literature. *Scientometrics*, 97(3), 871–882. <https://doi.org/10.1007/s11192-013-0983-y>
- Garfield, E. (1962). Can citation indexing be automated? *Essays of an Information Scientist*, 1, 84–90.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, 116(2), 973–996. <https://doi.org/10.1007/s11192-018-2767-x>
- Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the opencitations index of cross-ref open DOI-to-DOI citations. *Scientometrics*, 121(2), 1213–1228. <https://doi.org/10.1007/s11192-019-03217-6>
- Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. [https://doi.org/10.1162/qss\\_a\\_00022](https://doi.org/10.1162/qss_a_00022)



- Huang, C. K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2019). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *BioRxiv*. <https://doi.org/10.1101/750075>
- Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24), 3303–3309. <https://doi.org/10.1093/bioinformatics/btp585>
- Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *proceedings of the 2000 international conference on artificial intelligence (ICAI)*, 111–117.
- Kleminski, R., Kazienko, P., & Kajdanowicz, T. (2020). Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification. *Journal of Information Science*. <https://doi.org/10.1177/0165551520962775>
- Knorr-Cetina, K. (1981). *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Pergamon Press.
- Knorr-Cetina, K. (1991). Merton sociology of science: The first and the last sociology of science. *Contemporary Sociology*, 20(4), 522–526.
- Kranakis, E., & Leydesdorff, L. (1989). Teletraffic conferences: Studying a field of engineering science. *Scientometrics*, 15(5–6), 563–591. <https://doi.org/10.1007/BF02017071>
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* 5, 221–232. <https://doi.org/10.1007/s13748-016-0094-0>
- L. 240/2010. (2011). Rules concerning the organization of the universities, academic employees and recruitment procedures, empowering the government to foster the quality and efficiency of the university system (Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonché delega al Governo per incentivare la qualità e l'efficienza del sistema universitario), Gazzetta Ufficiale Serie Generale n.10 del 14/01/2011—Suppl. Ordinario n.11. Retried from <https://www.gazzettaufficiale.it/eli/id/2011/01/14/011G0009/sg>
- Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. SAGE library of social research. SAGE Publications, Inc.
- Leydesdorff, L. (1989). The relations between qualitative theory and scientometric methods in science and technology studies: Introduction to the topical issue. *Scientometrics*, 15(5–6), 333–347. <https://doi.org/10.1007/BF02017058>
- Leydesdorff, L., Råfols, I., & Milojević, S. (2020). Bridging the divide between qualitative and quantitative science studies (Issue). *Quantitative Science Studies*, 1(3), 918–1348. [https://doi.org/10.1162/qss\\_e\\_00061](https://doi.org/10.1162/qss_e_00061)
- Luukkonen, T. (1989). Publish in a visible journal or perish? Assessing citation performance of nordic cancer research. *Scientometrics*, 15(5–6), 349–367. <https://doi.org/10.1007/BF02017059>
- Marres, N., & de Rijcke, S. (2020). From indicators to indicating interdisciplinarity: A participatory mapping methodology for research communities in-the-making. *Quantitative Science Studies*, 1(3), 1041–1055. [https://doi.org/10.1162/qss\\_a\\_00062](https://doi.org/10.1162/qss_a_00062)
- Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*. <https://doi.org/10.1007/s11192-020-03690-4>
- Martini, M. C., Pelle, E., Poggi, F., & Sciandra, A. (2022). The role of citation networks to explain academic promotions. An empirical analysis of the Italian national scientific qualification. *Scientometrics*, 127, 5633–5659. <https://doi.org/10.1007/s11192-022-04485-5>
- Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.
- Milojević, S., Sugimoto, C. R., Larivière, V., Thelwall, M., & Ding, Y. (2014). The role of handbooks in knowledge creation and diffusion: A case of science and technology studies. *Journal of Informetrics*, 8(3), 693–709. <https://doi.org/10.1016/j.joi.2014.06.003>
- Miniaci, R., & Pezzoni, M. (2020). Social connections and editorship in economics. *Canadian Journal of Economics/revue Canadienne D'économique*, 53(3), 1292–1317. <https://doi.org/10.1111/caje.12460>
- Nuzzolese, A. G., Ciancarini, P., Gangemi, A., Peroni, S., Poggi, F., & Presutti, V. (2019). Do altmetrics work for assessing research quality? *Scientometrics*, 118(2), 539–562. <https://doi.org/10.1007/s11192-018-2988-z>
- Peroni, S., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Poggi, F., & Presutti, V. (2020). The practice of self-citations: A longitudinal study. *Scientometrics*, 123(1), 253–282. <https://doi.org/10.1007/s11192-020-03397-6>
- Peroni, S., & Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1), 428–444. [https://doi.org/10.1162/qss\\_a\\_00023](https://doi.org/10.1162/qss_a_00023)

- Poggi, F., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Peroni, S., & Presutti, V. (2019). Predicting the results of evaluation procedures of academics. *PeerJ Computer Science*, 5, e199. <https://doi.org/10.7717/peerj-cs.199>
- Portenoy, J., Hullman, J., & West, J. D. (2017). Leveraging citation networks to visualize scholarly influence over time. *Frontiers in Research Metrics and Analytics*, 2, 8. <https://doi.org/10.3389/frma.2017.00008>
- Rettberg, N., & Schmidt, B. (2012). OpenAIRE - building a collaborative open access infrastructure for European researchers. *LIBER Quarterly*, 22(3), 160. <https://doi.org/10.18352/lq.8110>
- Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management*, 57(2), 102168. <https://doi.org/10.1016/j.ipm.2019.102168>
- Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics*, 60(1), 71–79. <https://doi.org/10.1023/B:SCIE.0000027310.68393.bc>
- Stephenson, K., & Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1), 1–37. [https://doi.org/10.1016/0378-8733\(89\)90016-6](https://doi.org/10.1016/0378-8733(89)90016-6)
- Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K., & Evans, J. (2018). The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*, 47(9), 1825–1841. <https://doi.org/10.1016/j.respol.2018.06.014>
- Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 1–12. <https://doi.org/10.1038/s41598-019-41695-z>
- Tregellas, J. R., Smucny, J., Rojas, D. C., & Leggett, K. T. (2018). Predicting academic career outcomes by predoctoral publication record. *PeerJ*, 6, e5707. <https://doi.org/10.7717/peerj.5707>
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
- Visser, M., van Eck, N. J., & Waltman, L. (2020). Large-scale comparison of bibliographic data sources: Scopus, web of Science, dimensions, crossref, and microsoft academic. Retrieved from ArXiv:2005.10732 [Cs]. <http://arxiv.org/abs/2005.10732>
- Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1), 396–413. [https://doi.org/10.1162/qss\\_a\\_00021](https://doi.org/10.1162/qss_a_00021)
- White, H. D. (2004). Reward, persuasion, and the Sokal Hoax: A study in citation identities. *Scientometrics*, 60(1), 93–120. <https://doi.org/10.1023/B:SCIE.0000027313.91401.9b>
- Wyatt, S., Milojević, S., Park, H. W., & Leydesdorff, L. (2017). The intellectual and practical contributions of scientometrics. In U. Felt, R. Fouché, & L. Smith-Doerr (Eds.), *Handbook of science and technology studies* (pp. 87–112). MIT Press.
- Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and cword networks relate to each other. *Journal of the American Society for Information Science and Technology*, 63(7), 1313–1326. <https://doi.org/10.1002/asi.22680>
- Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology*, 62(8), 1498–1514. <https://doi.org/10.1002/asi.21556>
- Zhang, G., Liu, L., Feng, Y., Shao, Z., & Li, Y. (2014). Cext-N index: A network node centrality measure for collaborative relationship distribution. *Scientometrics*, 101(1), 291–307. <https://doi.org/10.1007/s11192-014-1358-8>
- Zhu, Y., Yan, E., Peroni, S., & Che, C. (2019). Nine million book items and eleven million citations: A study of book-based scholarly communication using opencitations. *Scientometrics*, 122(2), 1097–1112. <https://doi.org/10.1007/s11192-019-03311-9>
- Zinovyeva, N., & Bagues, M. (2015). The role of connections in academic promotions. *American Economic Journal: Applied Economics*, 7(2), 264–292. <https://doi.org/10.1257/app.20120337>

## Authors and Affiliations

Federica Bologna<sup>1</sup>  · Angelo Di Iorio<sup>2</sup>  · Silvio Peroni<sup>3</sup>  · Francesco Poggi<sup>4</sup> 

Federica Bologna  
fb265@cornell.edu

Angelo Di Iorio  
angelo.diorio@unibo.it

Silvio Peroni  
silvio.peroni@unibo.it

- <sup>1</sup> Information Science, Cornell University, Ithaca, USA
- <sup>2</sup> Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
- <sup>3</sup> Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
- <sup>4</sup> Institute of Cognitive Sciences and Technologies (ISTC), Italian National Research Council (CNR), Rome, Italy