Network Science (2023), 1–15
doi:10.1017/nws.2022.40

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Quality issues in co-authorship data of a national scientific community

Domenico De Stefano[1], Vittorio Fuccella[2], Maria Prosperina Vitale[3]* and Susanna Zaccarin[4]

[1]Department of Political and Social Sciences, University of Trieste, Trieste, Italy, [2]Department of Informatics, University of Salerno, Fisciano (SA), Italy, [3]Department of Political and Social Studies, University of Salerno, Fisciano (SA), Italy, and [4]Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy
*Corresponding author. Email: mvitale@unisa.it

Guest Editors (Special Issue on Scientific Networks): Dmitry Zaytsev, Noshir Contactor

Abstract

A stream of research on co-authorship, used as a proxy of scholars' collaborative behavior, focuses on members of a given scientific community defined at discipline and/or national basis for which co-authorship data have to be retrieved. Recent literature pointed out that international digital libraries provide partial coverage of the entire scholar scientific production as well as under-coverage of the scholars in the community. Bias in retrieving co-authorship data of the community of interest can affect network construction and network measures in several ways, providing a partial picture of the real collaboration in writing papers among scholars. In this contribution, we collected bibliographic records of Italian academic statisticians from an online platform (IRIS) available at most universities. Even if it guarantees a high coverage rate of our population and its scientific production, it is necessary to deal with some data quality issues. Thus, a web scraping procedure based on a semi-automatic tool to retrieve publication metadata, as well as data management tools to detect duplicate records and to reconcile authors, is proposed. As a result of our procedure, it emerged that collaboration is an active and increasing practice for Italian academic statisticians with some differences according to the gender, the academic ranking, and the university location of scholars. The heuristic procedure to accomplish data quality issues in the IRIS platform can represent a working case report to adapt to other bibliographic archives with similar characteristics.

Keywords: bibliographic archives; IRIS; web scraping; data management; data quality; Italian statisticians

1. Introduction

Over the last decades, scientific collaboration has been considered an important driver of research progress that supports researchers in the generation of novel ideas (Beaver, 2001; Glanzel & Schubert, 2005). The role of scientific collaboration was emphasized in recent government policies and international exchange programs aimed at stimulating the mobility of researchers and collaborative research (Wuchty et al., 2007; Defazio et al., 2009). Scientific collaboration has been also recognized as a key factor in measuring and evaluating scholars' scientific performance (Ferligoj et al., 2015; De Stefano & Zaccarin, 2016).

Most of the empirical studies on scientific collaboration mainly refer to the analysis of co-authorship, used as a proxy of scholars' collaborative behavior (Ponomariov & Boardman, 2016), especially in the wide variety of studies adopting the network perspective to analyze collaboration ties (Moody, 2004; Newman, 2004; Goyal et al., 2006). Co-authorship is receiving attention not only because bibliometric data are increasingly available but also because this type of relationship often is a tangible output of research collaboration. It seems quite reasonable that—particularly in

social sciences—scholars who co-authored a publication must have collaborated in some informal way by sharing ideas, data collection, and analysis, writing, etc. (Laudel, 2002).

Co-authorship networks, enriched with information on authors and publications that can be extracted by bibliographic databases, represent a good example of *complex networks* (Savić et al., 2019) and can help in the understanding of the structure and the evolution of research collaboration over time (Yan & Guns, 2014).

Two main broad types of co-authorship networks can be distinguished (Savić et al., 2019): (1) *field co-authorship networks*, representing collaboration among scholars working in a scientific field or discipline, such as Physics (Newman, 2004), Economics (Goyal et al., 2006), Sociology (Moody, 2004); and (2) *specific scientific community (or target population) co-authorship networks*, representing collaboration among scholars affiliated with institutions from one country, also referred as national or domestic co-authorship networks (Kronegger et al., 2012; Bellotti et al., 2016), or among scholars tied to a given field within institutions in one country (Digiampietri et al., 2017; Sciabolazza et al., 2017).

In this context, it is well recognized that the most convenient way to obtain a co-authorship network is to retrieve information on publications provided by bibliographic databases or digital libraries. The advantages of using such data sources are that, usually, they are relatively inexpensive and do not impose a burden on informants to provide information (De Stefano et al., 2013).

However, mostly in co-authorship networks on a target population, it was pointed out (Hicks, 1999) that international digital libraries provide partial coverage of all scholar scientific production, as well as under-coverage of authors in the target population. In particular, selection bias in retrieving co-authorship relationships in the community of interest increases as the specificity of the group increases, such as the interdisciplinarity of the research topics that is not easily classified in well-defined subject categories, publications in one's native language—usually restricted in a country—and/or publications in books or book chapters (Abel et al., 2019). Bias in retrieving co-authorship data of the community of interest can affect network construction and network measures in several ways, providing a partial picture of the real collaboration in writing papers among scholars.

For instance, in the Web of Science (WoS) or Scopus, the two bibliographic databases most used in such studies, products other than articles in international journals (books, book chapters, and papers in national journals) are not indexed to the same extent, resulting in a not satisfactory coverage of output, especially in the humanities and the social sciences (Abramo et al., 2019, p. 409; Aksnes et al., 2019, p. 748). Moreover, the selection policies of the products included in these archives can make the choice of a bibliographic data source a challenging task to obtain the completeness and accuracy of the publication data (Visser et al., 2021), since resulting co-authorship patterns can mirror data source characteristics. As shown in De Stefano et al. (2013), an interdisciplinary co-authorship behavior can emerge by the kind of products in WoS while internationalization openness by research topics and publication style can better captured by a thematic archive.

In order to have a thorough coverage of the overall scientific production of a target population, the ideal situation would be to retrieve data from the individual scholars scientific curricula, which are difficult to collect by direct interviews to the scholars and often are not available or not regularly updated on the Internet.

The integration between high-impact journal databases with specialized and local bibliographic archives could be a good compromise to retrieve a large proportion of all the research products of the community under study.

In dealing with bibliographic archives, several issues must be addressed, such as the identification of duplicate publication records, the treatment of scholar synonyms and homonymies, and the author name disambiguation of co-authors external to the target population.[1] These issues affect the data quality and thus the derived co-authorship network in several ways and are

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

exacerbated if different and heterogeneous archives are combined. As reported in several studies (e.g., Fegley & Torvik, 2013; Kim & Diesner, 2016), network measures were found to be biased by the merging or splitting of ambiguous author names. In particular, Fuccella et al. (2016) highlighted that the splitting identities of the adopted disambiguation procedure reduced network connectivity and affected statistics like the average degree, while the merging reduced the size of network structures, in terms of number of nodes and links. At the author level, splitting and merging mainly affected network measures based on geodesic distance.

In this contribution, we aim at discussing data quality issues in defining the collaboration style and the co-authorship ties among scholars of a specific scientific community. We focus on the group of Italian academic statisticians, as recorded in the Ministry of University and Research (MUR) in 2017. This case study is mainly proposed as a working example on the kind of issues to consider in the construction of co-authorship networks of a target community, even in a small size group, from available archives. For the community of Italian academic statisticians, we collected bibliographic records from an online platform, the Institutional Research Information System (IRIS), recently available at most Italian universities and in a few foreign universities, and including international and national publications. In each university, the IRIS platform is organized as a *people-article-centered* bibliographic database (Savić et al., 2019), with unique identifiers for publications and internal authors within institution.

The extraction of co-authorship networks from this type of bibliographic database poses difficulties in the identification of both authors and publications. In particular, we notice that the platform presents the pros and cons common to digital libraries. Even if it guarantees a high coverage rate of the target population and its scientific production, to retrieve co-authorship ties among scholars, it is necessary to combine the data contained in different platform deployments available at each university. In addition, especially for co-authors external to the target population, data quality is affected by the manual data entry by the authors and with duplicate records of publications that are co-authored by scholars hired at different universities. It is worth to note that similar difficulties have to be managed also in case of a simpler study at intra-institutional level.

To deal with the issues described above, we propose a strategy based on a web scraping procedure with a semi-automatic tool to retrieve publication metadata from this online platform. Then, we adopt data management processes, exploiting useful information about publications to deal with the data quality issues described above (e.g., duplicate records, disambiguation, authors external to the target population).

Finally, we report the main findings to describe the collaboration style among scholars using co-authored scientific products[2] with a focus on differences by gender, academic ranking, and geographic location of the university with which scholars are affiliated to highlight similarities and differences with comparable scientific communities in line with related literature (Abramo et al., 2013, 2019; Aksnes et al., 2019).

The remainder of the paper is organized as follows: Section 2 presents the IRIS platform and the related data quality issues. Web scraping techniques for extracting bibliographic data and the adopted data management tools are described in detail. In Section 3, the main results of the analysis are reported by presenting the characteristics of the target population, the author coverage rate, and the publication data management findings. The main characteristics of the resulting database are provided in Section 4, with a discussion of the co-authorship behavior of the target community. Section 5 reports final remarks.

2. Data retrieval from the online platform

2.1 Institutional Research Information System (IRIS)

The increasing availability of online, either public or private, bibliographic sources is "*extremely important in scientific communities since they give scholars ability to search and discover*

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

publications relevant for their work, p. 193" (Savić et al., 2019). These so-called current research information systems (CRISs) can offer the main data sources for a detailed study of co-authorship styles and their characteristics.

The IRIS[3] recently introduced in Italy seems to provide a unique platform for managing and supporting research in Italian academic and research institutions. Within this system, the Institutional Repository/Open Archive Module (IR/OA) is available as the repository of research products, allowing "*the storage, consultation and enhancement of publication outputs reflecting the various activities of a university*, p. 739" (Bollini et al., 2016). Thanks to this device, universities can access a system able to communicate with the MUR database and international databases for the management and dissemination of scholars' scientific publications. The need for the development of IRIS in Italy can be traced back to the 2009 act by the MUR (Law n.1, January 2009, art. 3-bis) requiring the creation of a national registry of scholars (full and associate professors, and lecturers) containing a list of their scientific publications updated annually. Unfortunately, to date, the registry is not available and renewed attention emerged in the last years. In the meantime, based on CRISs developed in other countries,[4] Italian experts on managing digital libraries agree on the general characteristics the registry must have: openness, accessibility, connection with international databases, flexibility of research product definition, and high data quality standards with respect to duplicate records, missing data, and errors, as well as certified data validation. Further, experts also agree on the potential of the IRIS platform to be used as the registry base (Galimberti, 2019).

In the list of institutions affiliated with the IRIS, 66 Italian universities out of 97 adopted this platform for publication data storage in 2018. At the university level, the IRIS platform is organized as a *people-article-centered* bibliographic database (Savić et al., 2019), with unique identifiers for each publication and each internal author to the institution. For each author affiliated with a given Italian university, people-article-centered organization assures complete longitudinal bibliographic data updated by scholars—the publication coverage is close to individual scientific CV—allowing, in principle, the extraction of all kinds of publications for the university level as well as selected target academic staff. In particular, personal and bibliographic information Italian scholars inserted in their IRIS are directly transferred to the individual scholars' web pages ("sito docente") managed by the MUR to rule the employment relationships (career progression, national funded projects participation, phd boards, etc.) of academic scholars. Unfortunately, the information on scholars' scientific production in the "sito docente" is not freely available, due to privacy policies.

The availability of this archive, in use since the end of 2015, is therefore a promising tool for co-authorship studies of the Italian academic community, as well specific academic groups. IRIS, as similar archives, represents an interesting case report for network science community, particularly when co-authorship networks are extracted by digital libraries accounting for a wide range of scientific production, not limited to the international indexed journals' repositories (e.g., WoS Scopus). Such a wide range production is especially relevant if the interest is to study the collaboration behavior in disciplines where co-authorship is still an emerging practice or where publishing in international journals is not the major goal (e.g., humanities, law). Nevertheless, several issues must be managed in merging data after the collection process from each university-based source to obtain a national-level archive.

Each university institution hosts its own IRIS operating deployment, where only some bibliographic data are fixed and mandatory, with no standard rules for inserting information. For example, manual data insertion affects the names of co-authors not affiliated with the university. Unfortunately, the platform does not provide a procedure for systematic download. Thus, to obtain bibliographic information at national level, web scraping techniques are needed to extract bibliographic data, and careful data management procedures have to be adopted to reconcile publication records and to detect duplicates.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

2.2 Web scraping techniques

Starting from scraping data techniques (Mitchell, 2015), a semi-automated tool is used to retrieve the bibliographic metadata of the target population from the IRIS platform. Each author has an IRIS page from which it is possible to access data about his or her publications. Therefore, bibliographic data are retrieved individually for each member of the target population.

The tool is implemented in Java. In addition to standard Java libraries from which download webpages, the Tagsoup library[5] is used for parsing well-formed or even unstructured and malformed HTML. This tool is programmed with the aim of automatically extracting the data from the system, obtaining good coverage of the author publications and reducing the manual adjustments to manage errors or uncertainty conditions. The input information is a table containing references (first name, surname, and academic institution) of all authors.

First, the URL of the IRIS page is retrieved for each author. Each institution hosts a different deployment of the system; thus, each author is linked to the index page of the IRIS deployment of his or her institution. Then, a query is launched on a specific search by the author interface available in the system. The interface responds by outputting a webpage containing a list of authors indicated by first name and surname, each associated with a link to the author's page. The last name of the author is used as a query string, and the author's first name and surname are considered to match an item in the list. As a result of the query, a complete database of the publication records for each author is available. The author's page contains the list of publications of which the member is a co-author. Each publication in the list is associated with a link to a new page containing the details of the publication.

2.3 Data management issues

The IR/OA module implemented in the IRIS platform provides several fields for each publication record, in particular, the title, list of all authors, publication venue, year of publication, type of product (e.g., article, book chapter, or conference proceedings), and various standard unique identifiers (URL, ISSN/ISBN, DOI, WoS, and/or Scopus codes, and so on).

Among the several characteristics of the IR/OA module, those of interest for our purpose are the following: (1) the IRIS and its IR/OA module were created to be maintained and managed by the individual university library without any restriction of the customization of IRIS functionalities, with the only aim of promoting open access to the university's publications; (2) the focus of the whole platform is on management of the individual scientific activities; therefore, data entry is left to the individual author who manually inserts mandatory as well as optional information for each publication.[6]

These characteristics can affect data quality and the consequent co-authorship network construction in several ways. First, each university library can heavily customize the IR/OA module content and freely choose which bibliographic data—except for the few fixed for every repository—are mandatory. For instance, publication standard unique identifiers are not always set as mandatory and, thus, cannot be used to reconcile different publication records. Given this heavy customization, publication records, and—even more importantly for data management—the available metadata of a publication, co-authored by authors at different universities, can vary for the same scientific product. Unfortunately, there is no automatic and reliable procedure that allows to match the same publication co-authored by authors at different institutions. Therefore, for each co-authored publication, there may be many duplicates of the same work according to the number of co-authors belonging to several universities.

Second, due to the manual data entry, the authors' names—in particular, those who are recognized as not belonging to the university—may be spelled in different ways or even typed incorrectly. Errors of this sort need to be identified and corrected to properly associate network authors with each identity.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

To cope with these issues, an additional data management step is necessary, in which the records corresponding to the same publication are (possibly) automatically reconciled with identities as well as authors' identities. To solve these data quality issues, the following tasks are performed.

2.3.1 Identification of duplicate publication records

There may be several instances of the same publication, each differing even for many fields, because of the manual entry required by the local deployment of the IRIS platform.

In general, we can distinguish two cases of possible duplications: (1) true duplicates, obtained when n scholars from different universities co-authored the same publication. In such a case, we expect to have up to n publications with identical or similar titles (to account for typing errors or misspellings due to manual entry) and same publication year, and (2) apparent duplicates, observed when authors developed an early-stage publication (typically, conference proceedings) into journal article leaving the publication title unchanged. In case of identical publication years but with different unique identifiers (e.g., ISBN) we retain the different occurrences in the dataset.

To recover and reconcile duplicate publications, we compute the edit distance (ED_T) between all publication titles for the retrieved data. The edit distance is a measure of similarity between two strings and is calculated as the minimum number of operations required to transform one string into the other. Operations can be of three types: removal, insertion, or substitution of a character. Then, we identify h different n-tuples ($n \geq 2$) of titles for which $ED_T \leq k$ (with k a specified threshold), obtaining h sets of potential duplicate publications. This simple heuristic allows us to reconcile even records affected by a few misspellings in the title. Finally, we identify as duplicate publications the ones included in the same sets with identical publication years and/or (if available) the same identifier. After this automatic identification procedure, we remove the duplicate publication if the author list is identical. It is possible that the author list may differ. To keep the maximum information about the author names, we select the publication whose author string contains the full first name and surname of the authors.

2.3.2 Detection and reconciliation of internal authors

In a given IRIS deployment, each internal author is marked with several unique identifiers (ORCID and e-mail). However, when an author appears as a co-author in a publication uploaded in another IRIS, we do not have a unique identifier attached to that identity. To reconcile internal authors when they appear as co-authors, we can match only the author name occurrences. In particular, we performed the internal author name disambiguation by matching surnames and names as described in Fuccella et al. (2016, pp. 171–176). In detail, firstly we create candidate identities merging all authors with "similar" surnames and at least the same initial letter of their names. Potentially, these merged identities are shrunk into a single node in the final co-authorship network. In particular, starting from the list of names internal to some IRIS deployment, we decided to consider differences in only one character as the most frequent misspellings in our data involve this case (except for the first letter of the surname).[7] Once these surnames are merged, we form a set of candidate identities, and we manually check if occurrences in the such sets belong to the same identity by means of their IRIS identifiers. We then split and merge identities accordingly to Fuccella et al. (2016).

That procedure showed good performance on a noiser dataset (0.83 and 0.81 precision and recall[8] values, respectively). For a review of some recent author name disambiguation techniques, see the contribution of Hussain & Asghar (2017).

2.3.3 Reconciliation of authors external to the IRIS system

External authors are all the scholars that do not belong to the IRIS system. They can be foreign authors, scientists working in the private sector, retired Italian academics, or belonging to

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

Table 1. Distribution of the 421 Italian academic statisticians in 2017. Source: MUR

Gender	%	Academic ranking	%	University location	%
Female	49.4	Researcher	33.7	North	46.3
Male	50.6	Associate professor	39.2	Center	24.0
		Full professor	27.1	South	29.7

universities not adopting IRIS. When an author is external to the system, it means that we do not have any unique identifiers for her or his identity. The reconciliation of these authors can be performed only if we have identities with their full first names and surnames. Thus, for authors external to the platform, we assume that they belong to the same identity only if they have the same first name and surname. All the other occurrences are treated as different identities. It is worth noting that we use external authors only to measure some characteristics of the internal authors (e.g., propensity toward co-authored publications); however if one is interested in reconstructing the overall co-authorship networks, the role of external authors becomes crucial since they considerably shape the network topology.

3. The target population

The target population under analysis is composed of the 421 academic statisticians who have a position as researcher or associate or full professor at Italian universities as recorded in the MUR database in 2017 and is classified as belonging to the Statistics disciplinary sector.[9]

Table 1 reports the composition of the statisticians by gender, academic ranking, and university location as reported in the MUR database. With respect to the gender composition, the distribution is balanced, a result in line with other contributions (Abramo et al., 2013). The same is noted for academic ranking. More than 40% of statisticians are affiliated with universities located in northern Italy. The proposed web scraping techniques and data management tools were performed to extract publication data for this national community as described below.

3.1 Data extraction and management

The publications were extracted from IRIS at the beginning of 2018. To set up a suitable time frame during which all Italian scholars started to manage the inclusion of their research products in online bibliographic archives following the lines established since 1999 in Italian university evaluation processes, we consider only products stored in the IRIS with a publication year from 2000 to 2017. Thus, the 1,900 papers published before 2000 were not taken into account in the further analysis.

Out of 421 statisticians, 349 were found, resulting in around an 83% author coverage rate. This value is in line with the rate (85.1%) resulting in 2010 from the no longer updated Current Index to Statistics (CIS) (p. 374) (De Stefano et al., 2013), a thematic database collecting worldwide publications on Statistics, sometimes also not in the English language. However, the value is higher than those obtained for the same Statistics subfield using WoS and a national archive (PRIN) based on publications attached to the national funded grants (71.3% and 72.7%, respectively). This first finding confirms the potentiality of the IRIS.

The implemented web scraping tool failed to trace 72 scholars: 59 scholars were affiliated to three universities with IRIS platform but with restricted access, and 13 scholars were affiliated to universities without IRIS archive. Despite the missing scholars, the resulting gender and academic rank distributions are in line with distributions from MUR database.

The average number of all publications found for statisticians in the platform is around 52 (St.Dev. 33.2), with a difference of around 10 publications in favor of male (55.4, with St.Dev. 37.3

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

for male and 46.6, with St.Dev. 27.4 for female), probably due to the gender differences in the academic ranking distribution.

These values are extremely high if compared with those found in 2010 (Table 3, p. 375) (De Stefano et al., 2013). They are similar to the value reported for Physics (52.5) in the COBISS database (Kronegger et al., 2011). This finding provides further evidence of the selection bias affecting the scientific production of target groups of scholars in international digital libraries, pointing out the usefulness of the IRIS for Italian scholars.

The kernel density plots in Figure 1 show the distribution of publications per author by gender, academic ranking, and university location of the statisticians found. Women, researchers, and scholars affiliated at a university located in southern Italy have a relatively smaller average. Positive skewness (presence of a few authors with a high number of publications) mainly characterizes men, full professors, and scholars affiliated with a university located in northern Italy.

The total number of retrieved publications is 14,514 (with duplicates). As described in the proposed data management procedure, to identify duplicates, we construct the edit distance (ED_T) matrix among all the retrieved publication titles. The distribution of the number of titles having ED_T up to 21 is shown in Figure 2. We assume that publications having $ED_T \leq 3$ (4,172 publications) and presenting the same publication year (and when available, the same publication identifiers) can be considered the same record. This threshold was established based on the authors' previous experience with publication data record linkage (Fuccella et al., 2016). By this criterion, we found 2,016 true duplicates. Most of the publications (904) are repeated twice, and only one publication is repeated five times. To test the effectiveness of our procedure and parameters, we performed a manual analysis of 500 random pairs of titles differing of 1 to 8 characters (included). We detected no false positives (no false duplicates were mistakenly detected). We instead detected 12 false negatives (duplicates not identified). These were mainly substantial misspellings in the titles due to very poor data entry by the authors. Duplications give also a rough estimate of the frequency and the size of inter-institutional collaborations among academic statisticians in the target population. In particular, about 14% of the total scientific production is derived from an inter-institutional collaboration, involving statisticians hired mainly at two Italian universities. For the detected duplicate publications, the one with the most informative author list (the full first name/surname of each co-author) is maintained if available; otherwise, it is randomly chosen from the set of duplicates. After the publication data management process, the number of publications retained without duplicates is 13,403 of the 2,016 duplicates we kept in our dataset one instance of the publications with at least two occurrences.

4. Publication characteristics and co-authorship behavior

To store a publication in the IRIS, in addition to the mandatory fields (type of publication, publication year, title, string of authors, journal venue), the authors may fill in several other types of information, such as the number of authors, language of product, various unique identifiers (ISSN, ISBN, DOI, ISI-WoS or Scopus). As most could not be mandatory in the specific IRIS university customization, they can provide, if any, useful insights into archive data quality, as well as on author behavior and style of publication.

By considering the 13,403 publications we retrieved after the data management process, we observed the following results.

The type of publication was missing in 26.3% of the records. Although this field was filled in, the content was not usable in another 13% of publications. The missing data in this mandatory field are due to the migration from previous research archives used by universities before the IRIS was installed where these metadata were not available or not yet correctly recovered at the time of our extraction. We expect that in the future, such missing data will be recovered mainly because both internal and external research evaluation exercises retrieve publication from the IRIS platform and such evaluation is crucial for career progression, national-funded projects

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

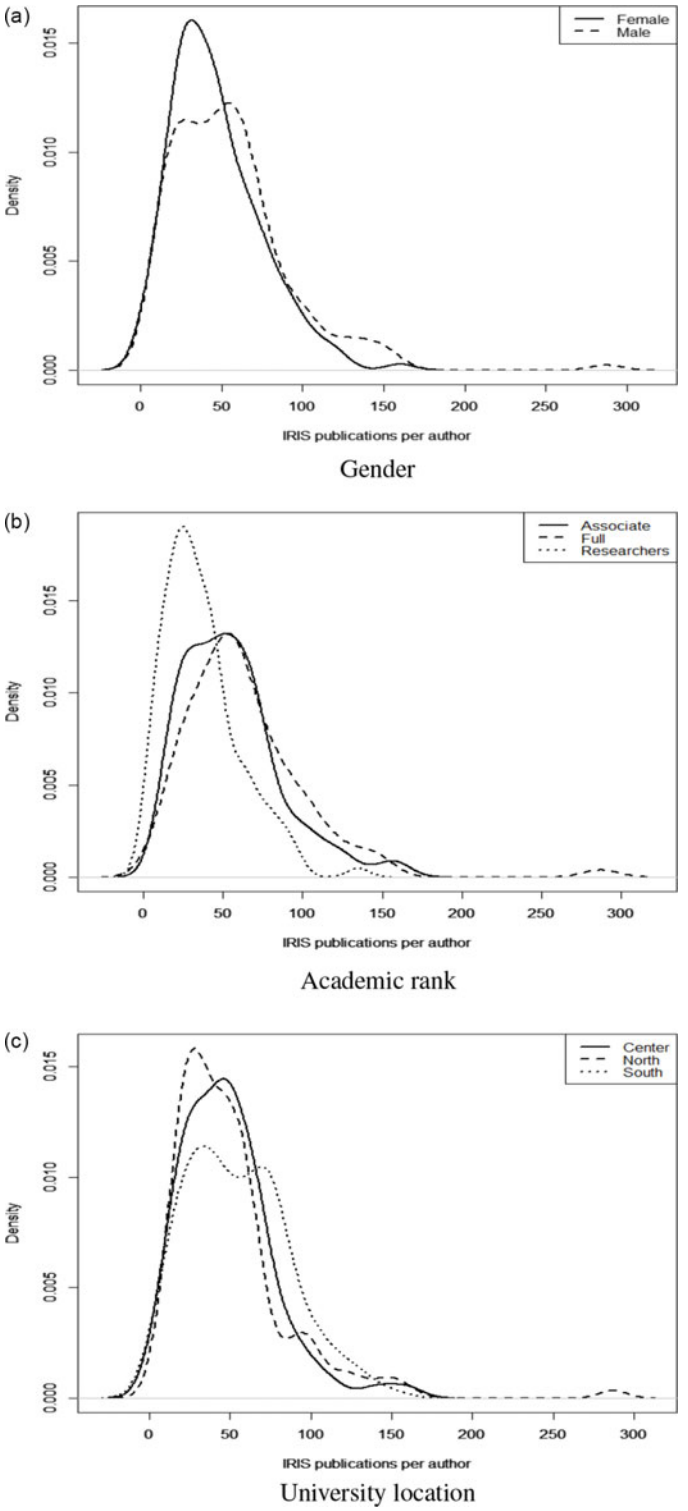

Figure 1. Kernel density plot of the distribution of IRIS publications per statisticians by gender (panel a), academic rank (panel b), and university location (panel c).

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

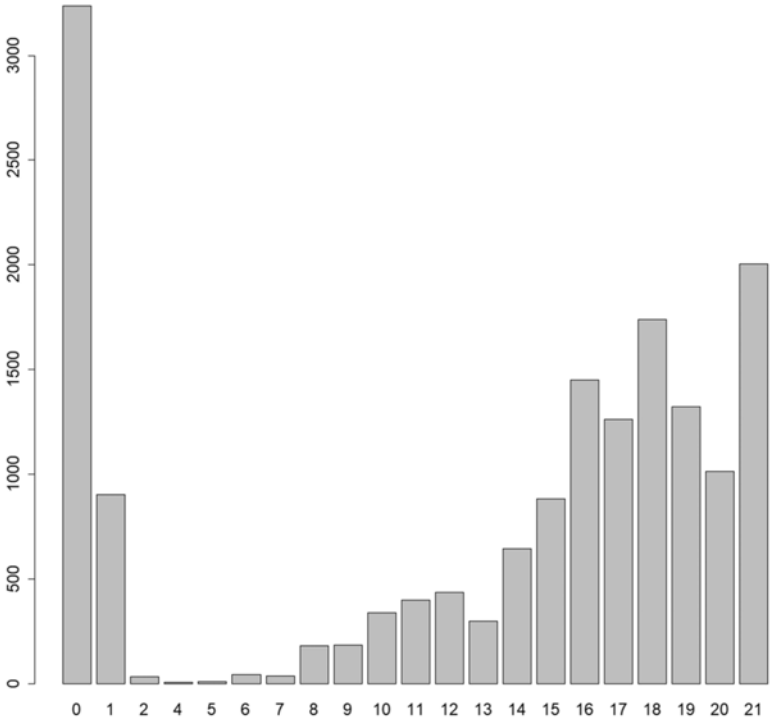

Figure 2. Distribution of the edit distance (ED_T) up to 21 among IRIS publication titles.

participation, phd boards, and so on. In the remaining publications (representing about 60% of the products), it appears that 38.3% of papers is published as articles in international and national journals, 22.9% as conference proceedings, 19.6% as chapters in books, 3.4% as monographs, and 15.7% as "other" kind of publications (e.g., conference abstracts, patents, posters, technical papers, teaching materials).

The language of the publication, a non-mandatory field, was missing in 15.8% of cases. If present, this information revealed a high tendency to internationalization in the writing style, with a percentage of 82.6% of the papers written in English (0.5% in other languages), and only 16.9% in Italian.

Only 23.7% and 29.7% of publications reported identifier codes of the WoS and Scopus databases of, respectively, the two main academic literature collections, including high-quality peer-reviewed scientific production (Aghaei Chadegani et al., 2013). It is worth to note that such international repository identifiers, although not mandatory in several IRIS installations, can be used in the online filling. In this case, the user does not manually insert the publication metadata but they are directly retrieved from the external—WoS or Scopus—source.

Looking at co-authorship, the percentage of co-authored publications was around 83%. The percentage is in line with values (about 85% on average) found in 2010 from publications in the WoS for the same population and at national level for scholars in the scientific area of Economics and Statistics (Abramo et al., 2013). This was likely due to the statisticians' attitude toward working with external co-authors involved in other disciplines in which the practice of collaboration is well established. The average number of authors per publication was around 3 (St.Dev. 4.4), as reported for the CIS and PRIN databases in 2010 for the same target population. This value was in line with the findings discussed by other groups of scientists. For instance, Kronegger et al. (2011) reported similar values for Slovenian mathematicians (2.8) and sociologists (3.7), whereas physicists and biotechnologists show a higher value (both 4.6) in the COBISS database for Slovenian scholars.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

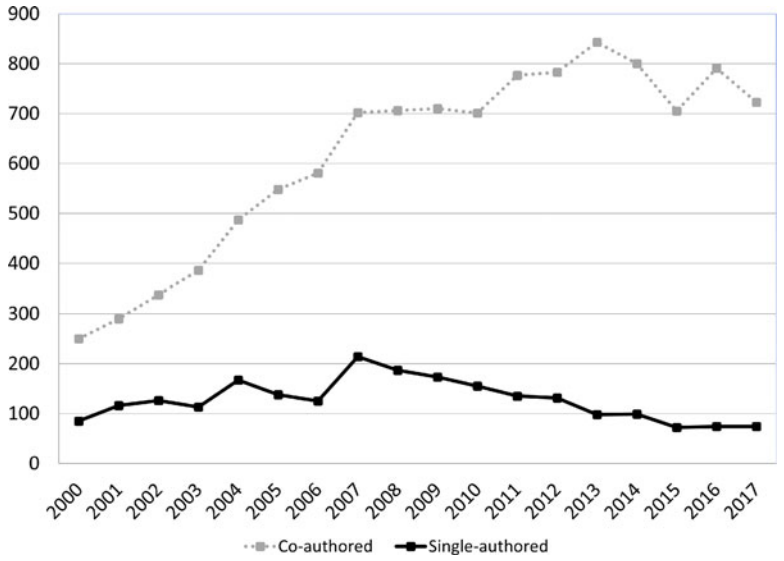

Figure 3. Trend of IRIS publications per statisticians, years 2000–2017.

The number of co-authored publications by Italian academic statisticians was growing faster than the number of single-authored publications, as observed in other scientific disciplines (Moody, 2004). We observed a significant increase in the proportion of co-authored publications especially from 2000 to 2007 (results are provided in Figure 3). This finding confirmed the tendency shown in the literature related to the global increase in collaboration as from the early 1990s (Kronegger et al., 2011). Specifically, in the target population, the mid-2000s were crucial years for scientific collaboration, probably in view of the increasing awareness of the central role played by statistics in all sciences, and in everyday applications.

This general trend was confirmed across the three main characteristics available for the target population, gender, academic ranking, and university location, with the percentage of co-authored publications always higher than that of single-authored publications (see Figure 4). Despite these trends in co-authorship style, a low but stable propensity in all groups to publish some publications as single authors was noted. This tendency was revealed as Italian statisticians need to construct their academic reputation through independent scientific production, as shown in other studies (MCDowell et al., 2006).

5. Conclusions

The present contribution discusses issues in the bibliographic data collection process when publications for a specific community are retrieved by using online platforms. To reconstruct the scientific collaboration style of Italian academic statisticians, the IRIS archive available at most Italian universities was adopted. Although it has guaranteed a high coverage rate of the target population and its scientific production with respect to other digital sources, many aspects undermining data quality were managed during the process of collecting data from this bibliographic source.

After the web scraping step to retrieve publication information, data management tools were used to detect duplicate records and to reconcile authors. Given that each institution hosts its own IRIS deployment, the systems are heterogeneous, with only a few fields fixed and mandatory, and with no standard rules for inserting information. As a result, many missing data were obtained, even for expected mandatory fields such as the type of publication. Product and author

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

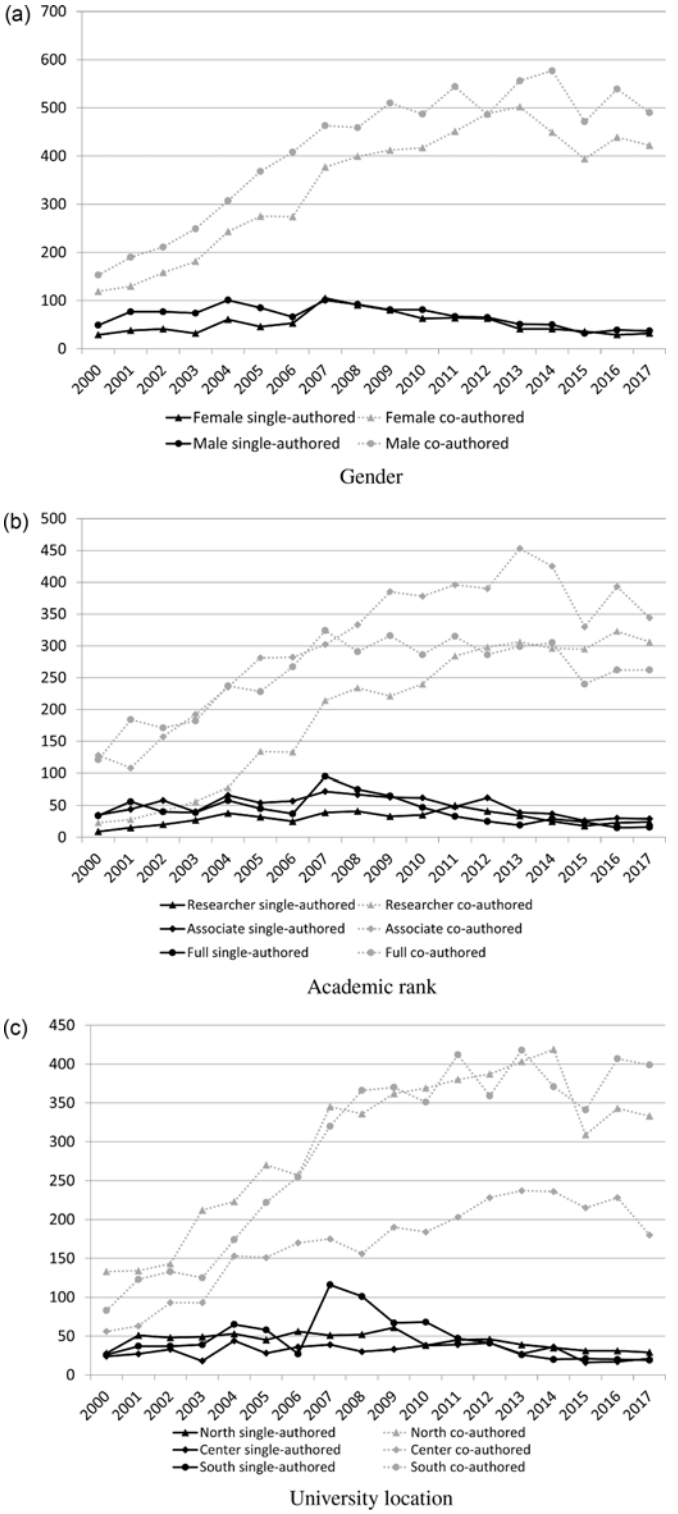

Figure 4. Trend of IRIS publications per statisticians by gender (panel a), academic rank (panel b), and university location (panel c), years 2000–2017.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

name duplications should be addressed before the co-authorship network is constructed. After data cleaning to reconcile publication records detecting duplicates, as proposed, the recognition of internal and external authors of the same publications must be considered by author disambiguation tools, to obtain good co-authorship data quality among scholars affiliated with different universities and with other institutions in Italy or abroad.

Nevertheless, interesting insights into the publication and collaboration style of Italian academic statisticians are derived. Collaboration is an active and increasing practice among scholars in this specific group, although the number of single-authored publications remains low but stable. Beyond the valuable production published with other authors (sometimes from different scientific fields given the wide interdisciplinarity of statistical methods and their applications), this can be motivated by the need to show one's own academic reputation, perhaps fairly appreciated in career progression. Most of the publications were written in English, allowing the access of results of Italian research for the international scientific community. With respect to the total number of products, as shown in other scientific fields and countries, women as a whole produce less publications than men. Although limited to a small group of Italian scientists, these results provide evidence of the high potential of the IRIS to become the bibliographic source for scientific production and co-authorship analyses in Italy. With respect to very recent results on the same population observed at the end of 2020 and composed by 455 statisticians, Bacci et al. (2021) were able to trace the 97.8% with at least one publication in the Scopus database. Despite this high coverage, they retrieved an average number of around 27 publications against the average number of around 52 publications found in IRIS. The number of Universities affiliated to IRIS is increasing (they were 77 at June 2020), and currently no Universities are limiting the access to their data. Therefore, the IRIS platform – accounting for the complete author production – allows a deeper analysis of research collaboration of Italian academic scholars than international databases, (e.g., international vs national collaboration, disciplinary vs interdisciplinary collaboration, different co-authors related to different types of production, etc.). Moreover, since IRIS automatically links WoS and Scopus publication identifiers, it is possible to recognize the different types of production and compare results.

Furthermore, the data quality issues affecting IRIS data source are common to similar studies where the purpose is to account for the complete co-authorship ties of a target population, using or combining the available digital archives. To adequately fulfill CRIS requirements of national-level archives, the drawbacks of the university-based source must be managed. To this aim, common insertion rules of the author name must be defined, and, of particular importance for the study of scientific collaboration, both authors and their affiliations must be indexed, including the country as a validated record. To guarantee complete and comparable coverage of all (scientific and scholarly) publication output (papers, books, edited volumes, conference series, etc.), publication information must be mandatory. To reduce the manual author burden, the system could automatically link the available data on editor and conference websites, in the same way of the current link to the WoS and Scopus databases. If reaching conditions are achieved, the IRIS platform could ensure complete, verifiable, and structured data for bibliometric and co-authorship analysis in Italy, as guaranteed in similar repositories active in European and non-European countries.

We would like to outline that the considerations on data quality are made at the overall two-mode network level which is the standard starting point of a co-authorship network analysis. In particular, we highlighted the issues that an online bibliographic retrieval can influence the structure of the two sets, authors by papers. The inconsistencies and the measurement errors at the level of author and paper sets also affect dyadic-dependent measures of the one-mode projection network, author-by-author. For instance, the node splitting or merging because of name inconsistencies can artificially alter substructures count (likewise dyadic and triadic motifs) thus obtaining misleading network results.

Competing interests. None.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

Notes

1 A useful example of data integration and data quality issues from different archives to reconstruct publication histories of Nobel prize winners, exploiting both manual inspection and algorithmic disambiguation procedures, is provided in Li et al. (2019).

2 We note that it could be also of interest to analyze collaboration by considering team works (Contractor, 2013; Jones et al., 2008) and participation in research projects (Bellotti et al., 2016).

3 The platform was developed by the Cineca consortium https://wiki.u-gov.it/confluence/pages/releaseview.action?pageId=51810588.

4 For instance, see the Brasilian Plataforma Lattes, http://lattes.cnpq.br/ currently in use since 1999; the US StarMetrics, https://www.starmetrics.nih.gov/; the European CERIF, http://www.eurocris.org/Index.php?page=featuresCERIF; the Norvegian CRISTIN https://www.cristin.no/english/, or the Slovenian COBISS, https://www.cobiss.si/en/.

5 For details, see https://hackage.haskell.org/package/tagsoup.

6 In the case of papers published on indexed journals, a new feature of the platform was recently introduced that consists of retrieving bibliographic information from international databases (WoS or Scopus).

7 As reported in Fuccella et al. (2016), the procedure uses compatibility "transitively," allowing the detection of misspellings of more than one character. An example is that of the surname "Mendoly," "Mendol" and "Mendola" were merged at one step; "Mendola" and "Mendiola" at a second step. As a result, also "Mendol" and "yMendiola" were merged, even though they differ of two characters.

8 In the context of author disambiguation, precision is the fraction of correctly disambiguated author instances among all disambiguated instances, while recall is the fraction of correctly disambiguated author instances among all author instances in the dataset.

9 For research and teaching reasons, at Italian universities each scholar is classified in one, and only one, of the 370 academic fields named "scientific disciplinary sectors," https://www.cun.it/uploads/storico/settori_scientifico_disciplinari_english.pdf. Academic scholars in the Statistics discipline are subdivided in five scientific disciplinary sectors: Statistics with 421 scholars in 2017, Statistics or Experimental and Technological Research with 20 scholars, Economic Statistics with 145 scholars, Demography with 70 scholars, and Social Statistics with 65 scholars.

References

Abel, G. J., Muttarak, R., Bordone, V., & Zaghen, E. (2019). Bowling together: Scientific collaboration networks of demographers at European Population Conferences. *European Journal of Population*, *35*(3), 543–562.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics*, *7*(4), 811–822.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2019). A gender analysis of top scientists' collaboration behavior: Evidence from Italy. *Scientometrics*, *120*(2), 405–418.

Aghaei Chadegani, A., Salehi, H., Yunus, M., Farhadi, H., Fooladi, M., Farhadi, M., & Ale Ebrahim, N. (2013). A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Social Science*, *9*(5), 18–26.

Aksnes, D. W., Piro, F. N., & Rorstad, K. (2019). Gender gaps in international research collaboration: A bibliometric approach. *Scientometrics*, *120*(2), 747–774.

Bacci, S., Bertaccini, B., & Petrucci, A. (2021). The co-authorship network of Italian academic statisticians: New evidences? In *5th European Conference on Social Network EUSN 2021*, 6–10 September 2021, Naples, Italy. Oral presentation.

Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, *52*(3), 365–377.

Bellotti, E., Kronegger, L., & Guadalupi, L. (2016). The evolution of research collaboration within and across disciplines in Italian Academia. *Scientometrics*, *109*(2), 783–811.

Bollini, A., Mennielli, M., Mornati, S., & Palmer, D. T. (2016). IRIS: Supporting & managing the research life-cycle. *Universal Journal of Educational Research*, *4*(4), 738–743.

Contractor, N. (2013). Some assembly required: Leveraging Web science to understand and enable team assembly. *Philosophical Transactions of the Royal Society A*, *371*, 20120385. doi: 10.1098/rsta.2012.0385.

Defazio, D., Lockett, A., & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy*, *38*(2), 293–305.

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, *35*(3), 370–381.

De Stefano, D., & Zaccarin, S. (2016). Co-authorship networks and scientific performance: An empirical analysis using the generalized extreme value distribution. *Journal of Applied Statistics*, *43*(1), 262–279.

Digiampietri, L., Rego, L., de Souza, F. C.OSTA, Ospina, R., & Mena-Chalco, J. (2017). Brazilian network of Phds working with probability and statistics. *Brazilian Journal of Probability and Statistics*, *32*(4), 755–782.

https://doi.org/10.1017/nws.2022.40 Published online by Cambridge University Press

Fegley, B. D., & Torvik, V. I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLOS ONE*, *8*(7), e70299. doi: 10.1371/journal.pone.0070299.
Ferligoj, A., Kronegger, L., Mali, F., Snijders, T. A. B., & Doreian, P. (2015). Scientific collaboration dynamics in a national scientific system. *Scientometrics*, *104*(3), 985–1012.
Fuccella, V., De Stefano, D., Vitale, M. P., & Zaccarin, S. (2016). Improving co-authorship network structures by combining multiple data sources: Evidence from Italian academic statisticians. *Scientometrics*, *107*(1), 167–184.
Galimberti, P. (2019). Anagrafe nazionale della ricerca (ANPREPS): A cosa serve e gli errori da non fare. https://www.roars.it/online/?p+68968
Glanzel, W., & Schubert, A. (2005). Analysing scientific networks through co-authorship. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257–276). Dordrecht, Netherlands: Springer.
Goyal, S., Van Der Leij, M. J., & Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy*, *114*(2), 403–412.
Hicks, D. (1999). The difficulty of achieving full coverage of International Social Science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215.
Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review*, *32*. doi: 10.1017/S0269888917000182.
Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, *11*(1), 3–15.
Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019). A dataset of publication records for Nobel laureates. *Scientific Data*, *6*(1), 1–10.
Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, *322*(5905), 1259–1262.
Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology*, *67*(6), 1446–1461.
Kronegger, L., Ferligoj, A., & Doreian, P. (2011). On the dynamics of national scientific systems. *Quality & Quantity*, *45*(5), 989–1015.
Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, *90*(2), 631–647.
McDowell, J. M., Larry, D., Singell, Jr., & Stater, M. (2006). Two to tango? Gender differences in the decisions to publish and coauthor. *Economic Inquiry*, *44*(1), 153–168.
Mitchell, R. (2015). *Web scraping with Python: Collecting data from the modern web*. Birmingham, UK: Packt Publishing.
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.
Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5200–5205.
Ponomariov, B., & Boardman, C. (2016). What is co-authorship? *Scientometrics*, *109*(3), 1939–1963.
Savić, M., Ivanović, M., & Jain, L. C. (2019). *Complex networks in software, knowledge, and social systems*. Cham, Switzerland: Springer International Publishing.
Sciabolazza, V. L., Vacca, R., Okraku, T. K., & McCarty, T. (2017). Detecting and analyzing research communities in longitudinal scientific networks. *PLOS ONE*, *12*(8), e0182516.
Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, *8*(2), 295–309.
Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41.
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Cite this article: De Stefano D., Fuccella V., Vitale M. P. and Zaccarin S. Quality issues in co-authorship data of a national scientific community. *Network Science*. https://doi.org/10.1017/nws.2022.40