Scientometrics

Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians

--Manuscript Draft--

Manuscript Number:	
Full Title:	Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians
Article Type:	Manuscript
Keywords:	Bibliographic data source; Record linkage; Author name disambiguation; Scientific collaboration; Co-authorship network
Corresponding Author:	Vittorio Fuccella ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Vittorio Fuccella
First Author Secondary Information:	
Order of Authors:	Vittorio Fuccella
	Domenico De Stefano
	Maria Prosperina Vitale
	Susanna Zaccarin
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	The bibliographic archives used to study scientific collaboration can affect bibliometric indicators as well as co-authorship network structures. In addition, the most used international databases might not be able to cover all kinds of works, especially for those disciplines with also nationally oriented publications. The integration of high-impact journal archives with specialised and local bibliographic ones could be a good compromise to obtain a higher coverage of the scientific work being produced. In this framework, our contribution aims at introducing a two-step procedure based on both record linkage and author name disambiguation in the presence of information retrieved by multiple data sources for a specific population. Evidences from Italian academic statisticians were provided by merging data from three bibliographic archives. The performance of our procedure was assessed by means of classic evaluation metrics in information retrieval and by discussing its implications in the co-authorship network analysis, both of disambiguated data and not disambiguated data.

Cite this article
Fuccella, V., De Stefano, D., Vitale, M.P. et al. Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians. Scientometrics 107, 167–184 (2016). https://doi.org/10.1007/s11192-016-1872-y

Noname manuscript No.
(will be inserted by the editor)

Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians

Received: date / Accepted: date

Abstract The bibliographic archives used to study scientific collaboration can affect bibliometric indicators as well as co-authorship network structures. In addition, the most used international databases might not be able to cover all kinds of works, especially for those disciplines with also nationally oriented publications. The integration of high-impact journal archives with specialised and local bibliographic ones could be a good compromise to obtain a higher coverage of the scientific work being produced. In this framework, our contribution aims at introducing a two-step procedure based on both record linkage and author name disambiguation in the presence of information retrieved by multiple data sources for a specific population. Evidences from Italian academic statisticians were provided by merging data from three bibliographic archives. The performance of our procedure was assessed by means of classic evaluation metrics in information retrieval and by discussing its implications in the co-authorship network analysis, both of disambiguated data and not disambiguated data.

Keywords Bibliographic data source · Record linkage · Author name disambiguation · Scientific collaboration · Co-authorship network

Introduction

The bibliographic archives used to study scientific collaboration can affect bibliometric indicators as well as co-authorship network structures. In addition, the most frequently used international databases might not be able to cover all kinds of products, especially for those disciplines having a more national orientation in their scientific production (Hicks 1999). In this case, the integration of high-impact journal databases with specialised and local bibliographic archives could be a good compromise to obtain a higher coverage of all the research products of a set of scientists involved in a specific field.

In exploiting the usefulness of heterogeneous bibliographic data sources, two main challenges have to be addressed: *1)* how to combine information by identifying and linking duplicate records, i.e. record linkage, and *2)* how to deal with issues related to author name disambiguation, i.e. the resolution of synonyms and polysems.

The record linkage of metadata refers to "the task of identifying records from disparate data sources that refer to the same entity" (Durham et al. 2012, p. 245), and it is often used to define integrated information systems in statistical settings (Fellegi and Sunter 1969; Liseo et al. 2006). Author name disambiguation "occurs when one author can be correctly referred to by multiple name variations (synonyms) or when multiple authors have exactly the same name or share the same name variation (polysems)" (Veloso et al. 2012, p. 680). The correct identification of author identities by name disambiguation tools enables research into co-authorship networks of scholars (see Li et al. 2014 for an application of name disambiguation and network analysis in U.S. patent inventors).

In this contribution, we aim at introducing a two-step procedure to deal with these two challenges in order to reach a better quality of co-authorship networks. We show the usefulness of the proposed procedure within a case study focusing on the scientific community of the 792 Italian academic statisticians (our target population) and their bibliographic data retrieved from three heterogeneous archives[1] to cover both top international products as well as nationally oriented publications (De Stefano et al. 2013).

To obtain a complete unified archive for co-authorship network analysis, we adopted a procedure based on both record linkage (RL), and author name disambiguation (AD) steps. In the first step, a semi-automatic method was adopted to merge in one unique database the three bibliographic archives by matching the sources in pairs. To evaluate the similarity of two records, some distance functions were considered on each of the key fields of authors, title and year of publications.

Due to the lack of training data, in the second step, a modified version of the unsupervised technique described in Strotmann et al. (2009) was applied for author name disambiguation. The algorithm followed a network analysis-based heuristic approach in which a graph-based representation of author occurrences was defined. An edge between two vertices was added if some evidences of belonging to the same identity were present (at least one co-author in common, same publication venue, publication titles share keywords, etc.).

The performance of our procedure was evaluated by first providing the classic evaluation metrics in the field of information retrieval (i.e. Precision, Recall, F_1 metrics), and then comparing overall and individual network statistics computed before and after the disambiguation process.

[1] Two international databases, one general (WoS) and one thematic (Current Index to Statistics, CIS) were considered, together with bibliographic information retrieved from the Italian Ministry of University and Research (MIUR) database of nationally funded research projects (PRIN).

The remainder of this paper is organised as follows. In the “Related works” section, we briefly review the main approaches proposed for record linkage and author name disambiguation in bibliographic Digital Libraries (DLs). Section “Data” describes the main characteristics of the data sources used to retrieve bibliographic data on Italian academic statisticians. Section “Two-step procedure” provides details on the procedure we adopted to merge the three data sources in one unique archive (*Record linkage*) and to deal with the author name disambiguation issue (*Author name disambiguation*). In the Section “Results”, we first discuss the algorithm disambiguation results in terms of evaluation metrics and then we compare the co-authorship networks constructed after the record linkage and the disambiguation steps. In the “Conclusion” section, we provide final remarks and future work.

Related works

Record linkage and disambiguation of metadata in DLs are very sensitive issues that involve the processing of person names on the basis of name-internal and/or external features (Kang et al. 2009). Several different computer-oriented record linkage methods are reported in the literature (Domingo-Ferrer and Torra 2003; Dong et al. 2005; Yan et al. 2007; Christen 2012). The methods that are currently in use generally compare record pairs and classify each pair into matches, nonmatches, and possible matches. The main objective of recent methods is to ensure a high efficiency and scalability on large data sets. Several different indexing techniques, aimed at reducing the number of comparisons, have been proposed. A common indexing technique is blocking (Baxter et al. 2003) which groups similar input entities into non-overlapping blocks. Only records that belong to the same block are compared with each other. Another technique, called sorted neighbourhood method (Hernandez and Stolfo 1995), first sorts all records and then iterates on the sorted list, comparing all the records in a sliding window of a fixed size. A technique for adaptively selecting the window size has been described by Yan et al. (2007). A survey and a comparison of indexing techniques is presented in Christen (2012).

A myriad of recent studies are devoted to name disambiguation methods in bibliographic DLs in computer science, sociological and linguistic settings by covering supervised techniques, based on training data sets of pre-labeled citations (Torvik et al. 2005; Veloso et al. 2012; Ventura et al. 2015), unsupervised techniques, based on a learning-free similarity function between two citations (Han et al. 2005; Kang et al. 2009; de Carvalho et al. 2011; Imran et al. 2013), or semi-supervised techniques, typically based on a small amount of labeled data with a large amount of un-labeled data (Smalheiser and Torvik 2009; Criminisi et al. 2012) techniques. A recent survey is presented in Ferreira et al. (2012) along with a hierarchical taxonomy to characterise automatic methods for author name disambiguation. This taxonomy reported the most representative methods proposed in the literature according to the main type of exploited approach to deal with author name references or, alternatively, according to the information (evidence) explored in the disambiguation task, mainly citation attributes and Web information (Ferreira et al. 2012, p. 16).

More formally, given the set of citations $C = \{c_1, c_2, \ldots, c_k\}$, where each citation c_i contains both name-internal and name-external features (such as author names, affiliation, publication title and venue), the name disambiguation task is to define a function to partition the set of citations into n sets $\{a_1, a_2, \ldots, a_n\}$, where each partition a_i contains the citations of i-th author (de Carvalho et al. 2011; Veloso et al. 2012).

Among the minimal set of citation attributes (typically co-authors, publication title and venue), co-authorship was considered to be "the most reliable and decisive from the viewpoint of discriminating the identities of authors, since it implies real-world acquaintances among authors" (Kang et al. 2009, p. 85). By relying exclusively on collaboration patterns between authors, the algorithm described in Strotmann et al. (2009) merged *compatible* occurrences which show some evidence of referring to the same identity. This algorithm can be defined as a "network analysis-based heuristic approach" (Cota et al. 2010).

Data

Our case study focuses on the target population of the 792 academic statisticians (henceforth denoted by "statisticians") who have permanent positions in Italian universities, as recorded in the MIUR database in March 2010[2] and belonging to one of the five subfields established by the governmental official classification: Statistics, Statistics for Experimental and Technological research (E&T), Economic Statistics, Demography, and Social Statistics.

The five subfields differ mainly on the basis of a methodological or an applied research interest in Statistics. Beside scientists' preferences, subfield specialties and community traditions can affect the publication production style (single-authored *vs* co-authored and/or writing articles *vs* books and/or publishing in international *vs* national journals) of statisticians in Italy.

Complete bibliographic information on this scientific community could be collected from publication forms filled in individual scholars' web pages ("sito docente Cineca"), managed by the MIUR and the Cineca consortium. Due to the privacy policy, access to this database is denied to the public.

Since 2000, only partial bibliographic information has been made available by the Cineca consortium regarding selected publications by statisticians involved in nationally funded research projects (PRIN)[3] as national managers or members. We referred to the period 2000-2008 for this study; 2008 was the last available year in the PRIN database collected by De Stefano et al. (2013). In studying the influence of database characteristics on the co-authorship patterns of Italian statisticians, De Stefano et al. (2013) and later De Stefano

[2] At December 2014 the size of population was 722.

[3] Although PRIN funding was launched in 1996, information on funded projects has been released only since the year 2000.

and Zaccarin (2015) retrieved publications from two additional sources: the international database of Web of Science, (WoS) and the thematic archive of Current Index to Statistics (CIS). For statisticians CIS represents the principal available data source containing publications in Statistics and related fields, though it is not regularly updated.

International databases, usually containing high-impact publications on topics covered by the archive editorial policies, have been often used to study scientific collaboration inside disciplines (see, among others, Albert and Barabási 2002; Moody 2004; Newman 2004; Goyal et al. 2006). The main problem with these databases in gathering co-authorship data for a specific target population –as in our case– is the uncoverage of those works published at the national level (Hicks 1999).

As discussed by De Stefano et al. (2013), the specific features of the three data sources on publications of Italian statisticians affected the retrieved number of publications and the author coverage rate (i.e. the percentage of statisticians found in a data source out of the total of 792). The highest number of publications was collected through the PRIN database, followed by CIS and WoS (see Table 1). As expected, this result reflects the different kinds of publications collected in the three databases with a higher inclusion of nationally oriented production in PRIN (e.g. national conference proceedings, papers in Italian journal and books).

WoS showed the lowest author coverage rate (60.7%) (see Table 1) with substantial subfield differences (De Stefano et al. 2013, Table 2, p. 374): Statistics for E&T research was quite well-represented (86.7%) whereas only 40.0% of scientists were found in Demography. Statistics and Economic Statistics were well covered within CIS (85.1% and 65.0%, respectively), while authors in Demography and Social Statistics appeared more frequently in PRIN (81.1% and 67.1%, respectively). The lowest author coverage rates in WoS and CIS for subfields oriented to Social Sciences applications may be due to the partial inclusion of publications focusing on the specific research topics of these subfields, and a higher tendency to produce publications at a national level. The total percentage of authors not found in the three databases was 13%.

The highest percentage of co-authored publications was found in WoS (about 85% on average) and the lowest value in CIS (55.3%) with PRIN exhibiting an intermediate value (71.2%) (De Stefano et al. 2013, p. 374). Furthermore, WoS appeared as the data source in which the average number of co-authors for each statistician was extremely high, due to the presence of few statisticians with a large number of co-authors (mainly from not statistical disciplines).

Resulting co-authorship patterns also mirrored data source characteristics (De Stefano et al. 2013, p. 380). Patterns consistent with well-established network structures were found in the CIS database. In particular, CIS captured internationalisation openness by research topics and publication style, while WoS mainly captured the tendency towards an interdisciplinary behaviour. Finally, PRIN combined some of both CIS and WoS characteristics, although it referred only to the selected publications by projects managers and members.

Table 1 Number of publications and author coverage rate in the three bibliographic archives.

	Years	# of publications	Author coverage rate
WoS	1989–2010	2289	60.7%
CIS	1975–2010	3459	73.4%
PRIN projects*	*2000–2008**	5054	70.2%

* Years of the project.

Two-step procedure

As reported in the previous section, the three data sources presented only partially overlapping information. To take advantage of this heterogeneity in order to obtain a better quality of co-authorship data for our target population, two main challenges have to be addressed: *1)* how to combine information from heterogeneous sources by identifying and linking duplicate records, and *2)* how to deal with issues related to author name disambiguation. To this purpose, we adopt a two-step procedure to merge the three bibliographic archives in one unique archive, through record linkage (RL), and to cope with the author disambiguation (AD) issue.

Record linkage

Given the relatively small number of records in the three data sources (see Table 1), we opted for a semi-automatic method, which requires human intervention to resolve situations of uncertainty. We adopted this procedure because of the presence of errors and omissions in the original datasets (e.g. misspellings in the names of authors and titles, discrepancies in the name of the venue, lack or inaccuracy in the year of publication), especially in PRIN.

In order to perform the linkage of the three data sources, we proceeded with the commonly used approach of matching the sources in pairs and then performing a reconciliation of possible discrepancies (Sadinle et al. 2011).

In order to evaluate the similarity of two records, we used the following distance functions on each of the key fields:

- **Authors**: the *Jaccard* distance between the set of surnames of the authors of the two records (d_A).
- **Title**: the error rate measure derived from the edit distance between the two compared strings t_1 and t_2. In particular, we defined the distance as:

  $$d_T = Ld(t_1, t_2)/max(|t_1|, |t_2|)$$

  where the numerator is the *Levenshtein* distance between t_1 and t_2, and the denominator is the maximum length of the two compared titles.
- **Year**: the absolute value of the difference between the years of publication (d_Y).

All strings were lower-cased before any comparison. The overall distance was defined as a 3-tuple (d_A, d_T, d_Y), where each element was the distance calculated as described above on the three key fields. We established a threshold for the distance on each element and automatically linked the pairs whose distances were below the following thresholds: the couples having $d_T < 10\%$, $d_A = 0$ and $d_Y = 0$ were marked as "matches". The couples having $d_T < 20\%$ and $d_A \leq 1$ (except for those already automatically linked) were marked as "possible matches" and left for further manual processing.

We first looked for matching records in pairs of sources. The count of "matched" and "possibly matched" pairs of records are reported in Table 2. The "possible matches" were manually inspected, resulting in a number of "total links found" reported in the last column of the table.

Table 2 Number of linked records in the pairs of sources before reconciliation. At the end of the process, the number of linked records slightly changed.

Sources	Matches	Possible matches	Total links found
(WOS, CIS)	782	71	827
(CIS, PRIN)	729	209	917
(PRIN, WOS)	612	166	756

Lastly, we performed the reconciliation step, which allowed us to find a small number of discrepancies. These were manually resolved, resulting in a unified archive containing 8735 publications, whose composition is shown in Fig. 1. In the figure, we use a Venn diagram to summarise the result of the record linkage process. The cardinality of the sets and of their intersections is reported on the curves. The number of overlapping publications retrieved in all the three data sources was rather small. They represented only 5.0% in the combined archive. Considering only couples of databases, we found very similar percentages. 43.7% of publications were retrieved only from PRIN, followed by 24.6% of the publications from CIS and 13.0% from WoS. These results confirm the high heterogeneity of scientific production among Italian statisticians.

Author name disambiguation procedure

The archive resulting after RL contained 8735 publications authored by 677 statisticians and their co-authors, most of them foreigners, for a total of 7332 authors.

We addressed the problem of author disambiguation through an unsupervised method due to the lack of training data. In particular, we strongly drew inspiration from the method described in Strotmann et al. (2009) because it follows a network-based heuristic approach and it has the advantage of requiring the availability of a restricted set of record attributes (identifier, co-authors, venue). Therefore, it is well suited to our needs, considered our aims and the information available in the unified archive.

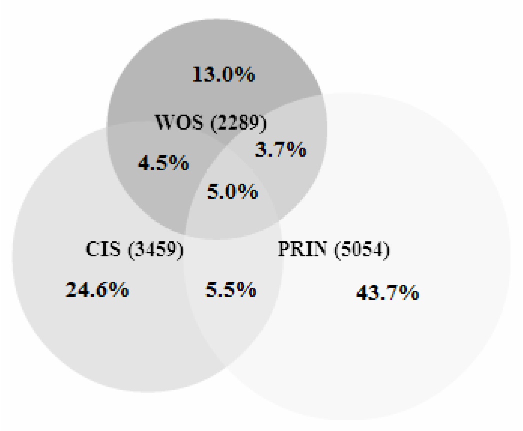


Fig. 1 The number and the percentage of publications in the unified archive after record linkage by data sources (The circle's size is proportional to the number of publications in each data source).

Limitations of this method include both the lack of misprint handling in name compatibility checking, and the pessimistic behaviour in merging the identities, i.e. it returns more identities than there are. Furthermore, the use of the publication identifier from PubMed proposed in the Strotmann algorithm (Strotmann et al. 2009) is not applicable to our case.

We improved the original method by adding misprint handling and an enhanced use of record data to merge identities. In particular, we considered the title of the publication, which conveys important information on the subject of the research, and the identifier of the query, with which the record was retrieved in our database, from one of the three data sources.

Similarly to the above described algorithm, we used a graph-based representation of author occurrences, each of which is associated to a graph vertex. We added an edge between the two vertices every time their associated occurrences showed some evidence of belonging to the same identity. The output identities were obtained by calculating the connected components of the graph, each connected component being a different identity.

Evidences were only checked on compatible names, i.e. names that may refer to the same identity. Some occurrences had a full first name (expanded), others only had the initials (abbreviated). A normalisation of the names is executed before compatibility checks by removing diacritics and by converting the author names into lower case. In order to cope with misspellings, we also considered as compatible surnames that differed by a single character (except for the first letter of the surname). For instance, $Vittadini, G$ and $Vittadin, G$ were considered to be compatible, and not $Martini, C$ and $Sartini, C$. Lastly, we considered the case of authors with more than one first name or surname. In particular, we relaxed the checks by considering as compatible two entries sharing at least one surname, or one first name initial. For instance, the following couples of occurrences were all considered to be compatible: $Aureli\ Cutillo, E$ and $Aureli, E$; $Monti, AC$ and $Monti, A$; $Arboretti\ Giancristofaro, R$ and $Arboretti, GR$. The set of vertices was initially partitioned into two parts: those of abbreviated occurrences and those of expanded ones. Then, edges were added between vertices in three consecutive steps:

- *Step 1*: pairing of occurrences having compatible expanded names (e.g. Vittadini, Giorgio; Vittadin, Giorgio)
- *Step 2*: pairing of abbreviated to compatible expanded names (e.g. Vittadini, G.; Vittadini, Giorgio)
- *Step 3*: pairing of occurrences having compatible abbreviated names (e.g. Vittadini, G.; Vittadin, G.)

An edge was added between two vertices if their associated occurrences were compatible and showed at least one of the following evidences, based on the attributes of their respective publication records:

- at least one co-author in common;
- same publication venue;
- the two records were retrieved in the same query;
- the titles shared at least one keyword.

For each checked occurrence, the associated vertex is only connected to the vertex with the highest evidence. We calculated an evidence measure as:

$$E = w_a \times e_a + w_v \times e_v + w_q \times e_q + w_t \times e_t$$

where E has real values in the range $[0, 1]$; w_a, w_v, w_q and w_t are the weights for the functions e_a, e_v, e_q and e_t, respectively. Table 3 reports, for each function, the attributes used to calculate them, the function domain, and how it was defined. The similarity between titles (function e_t) is determined by the commonality of keywords. The weight of the keywords was established through TF-IDF statistic, which assigns greater weight to infrequent words and penalises those that are particularly common. Thus, it is used for stopwords filtering. In this application, we set the weights to 0.25 in order to give the same weight to each of the four functions.

Table 3 Functions to evaluate the evidence measure E.

Function	Data	Values	Definition
e_a	Co-authors	[0,1]	Jaccard coefficient
e_v	Venue	{0,1}	1 = same venue, 0 otherwise
e_q	Query Id	{0,1}	1 = same query, 0 otherwise
e_t	Title	[0,1]	TF-IDF similarity between titles

Results

The performance of our procedure was evaluated by first providing the classic evaluation metrics in the field of information retrieval for checking authors' identities, and then comparing overall network structures and individual network statistics derived before and after the disambiguation process.

Evaluation of the AD procedure

As a consequence of the name disambiguation procedure, the true authors' identity could be compromised for two reasons: "a given individual may be identified as two or more authors (splitting), or two or more individuals may be identified as a single author (merging)" (Milojević 2013, p. 767).

Hence, it is necessary to evaluate the accuracy of the results provided by a given disambiguation method by using performance measures. In the presence of a list of individuals already correctly assigned, it is possible to identify the number of right identities returned by the algorithm, i.e. the true positive (TP), and the number of incorrect identities obtained by merging separate authors, i.e. the false positive (FP) or by splitting unique author, i.e. the false negative (FN). Three measures of performance (see Table 5) were typically defined according to these quantities, precision (P), recall (R) and the harmonic mean of P and R metrics F_1 (Kang et al. 2009; Gurney et al. 2011; Cuxac et al. 2013; Imran et al. 2013).

Automatically establishing if the results of the name disambiguation algorithm are correct is a difficult task. To this end, two approaches are usually followed: to evaluate the accuracy over a simulated dataset in which the true author's identity is known (Milojević 2013) or to manually check a (small) randomly selected sample and comparing it with the dataset obtained by the disambiguation algorithm (Strotmann et al. 2009; Imran et al. 2013; Wu and Ding 2013).

Since we were mainly focusing on our target population, we adapted the latter approach to compute the three evaluation measures on the disambiguated data:

1. Starting from the list of statisticians, we matched the surnames and initials of the authors included in the target population with the identities returned by the algorithm. In this way, we obtained the set of authors with one identity per author (TP), the set of authors with merged identities (FP) and the set of authors with separated identities (FN). The size of the two FP and FN sets could be considered as an upper bound of errors without a manual check.
2. A sample of authors was extracted from the list of statisticians in order to improve the accuracy of the computed metrics by furnishing the exact number of FP and FN in the sample, thanks to the manual check for the correct author identity.

The disambiguation procedure returned a total of 7230 identities.

By matching the surnames and initials of the statisticians with the disambiguated identities, we found 808 identities possibly associated to the statisticians. More specifically, 489 authors were rightly assigned by the AD procedure

(TP), while the identities of 102 statisticians were merged (FP) and 112 were separated in two or more identities (FN). A fine-grained control on our target population showed that the merging and splitting of identity assignment was mainly due to the presence of authors with double surnames names, compound surnames and double/multiple first names with or without an apostrophe[4]. Table 4 reports some examples of authors presenting these features, showing the algorithm results and the identity assignments in terms of TP, FP and FN.

A 5% random sample of authors was selected from the list of statisticians found after the record linkage step. The total sample size ($n = 34$) was subdivided according to the proportion of the three sets of identities returned by the AD algorithm. The final sample consisted of 24 TP, five FP and five FN authors. After a manual check, we identified two FPs and five FNs in the sample.

The values of the three evaluation metrics computed for the population of statisticians and for the extracted sample of statisticians (Table 5) were similar to the results reported by other authors (Kang et al. 2009; Strotmann et al. 2009; Wu and Ding 2013) showing a good performance of our procedure. In particular, it is worthy to note that in the case of the population, the reported values of around 0.80 represent the lower bound of the evaluation metrics that arise to 0.90 in the sample results.

Beyond the identities of statisticians, the AD procedure found 6422 identities related to external authors. We noticed that the algorithm returned 5880 unique identities (TP); it failed in assigning 285 authors separated in two or three identities (FP) and 261 authors merged in one identity (FN). The three evaluation measures presented very high values (see Table 5) showing a very good performance of the adopted disambiguation method in the case of external authors.

Network results comparison

In the following, we describe how we used the AD procedure output to construct the co-authorship networks[5] (AD$_{NET}$) of all authors (7230 nodes) and of statisticians (808 nodes). In order to assess how the AD procedure may affect network outputs, we also considered the co-authorship networks built on author identities –7332 authors and 677 statisticians– resulting from the record linkage step (RL$_{NET}$).

[4] A total of 14 authors with double surnames, 48 and 88 authors with compound surnames and double/multiple first names with or without an apostrophe, respectively, was observed in our target population.

[5] A co-authorship network is derived from the matrix product $\mathbf{Y} = \mathbf{A}\mathbf{A}'$, where $\mathbf{A}$ is a $n \times p$ affiliation matrix, with elements $a_{ik} = 1$ if $i \in \mathcal{N}$ authored the publication $k \in \mathcal{P}$, 0 otherwise. The matrix $\mathbf{Y}$ is the undirected and valued $n \times n$ adjacency matrix with element y_{ij} greater than 0 if $i, j \in \mathcal{N}$ co-authored one or more publications in $\mathcal{P}$, and otherwise 0. The binary version of $\mathbf{Y}$, setting all entries in the valued adjacency matrix greater than zero to 1, was used in our analysis.

Table 4 Examples from the target population with double surnames [DLN], compound surnames & an apostrophe [CLN/A], and compound surnames, double first names & an apostrophe [CN/A], algorithm results and identity assignment.

Target population	Algorithm results	Identity assignment
DLN		
ARBORETTI GIANCRISTOFARO Rosa	Giancristofaro, Rosa Arboretti (RA) = 7 Giancristofaro, Arboretti (A) = 1, Arboretti Giancristofaro, (R) = 21, Arboretti, Rosa (R)=5	**FP**
BERTOLI BARSOTTI Lucio	Bertoli Barsotti, (L) = 3, Bertoli-barsotti (L) = 2, Barsotti, (L)=13, Barsotti, (LB) = 1	**FN**
BERTOLI BARSOTTI Lucio	Bertoli Barsotti, (L) = 1	**FN**
BERNARDINI PAPALIA Rosa	Bernardini Papalia, (R) = 1	**FN**
BERNARDINI PAPALIA Rosa	Bernardini Papalia, (R) = 8	**FN**
BUSCEMI CUCCIOLITO Silvana	Buscemi, (S) = 1	**TP**
DLN/A		
DALLA ZUANNA Gianpiero	Dalla-zuanna, (G) = 3, Dalla Zuanna, (G)=30 Zuanna, (GD) = 3	**FP**
DE CANTIS Stefano	De Cantis, Stefano (S) = 25	**FN**
DE CANTIS Stefano	De Cantis, (S) = 1	**FN**
D AGOSTINO Antonella	D'agostino, Antonella (A) = 9	**TP**
CN/A		
ALTAVILLA Anna Maria	Altavilla, (A) = 11	**TP**
AREZZO Maria Felice	Arezzo, (MF) = 1	**FN**
AREZZO Maria Felice	Arezzo, (MF) = 1	**FN**
BARBIERI Maria Maddalena	Barbieri, Maria Maddalena (MM) = 27 Barbieri, (M) = 3	**TN**
BILLARI Francesco Candeloro	Billari, Francesco (F) = 2, Billari, (FC) = 60	**FN**
BILLARI Francesco Candeloro	Billari, (FRANCESCO) = 1	**FN**
D AGATA Rosario Giuseppe	D'agata, (R) = 1	**FN**
D AGATA Rosario Giuseppe	D'agata, (R) = 2	**FN**

Table 5 Performance measures: formula and computed values for all statisticians, for the sample of statisticians, and for external authors.

Metrics	Formula	Statisticians	Sample of Stats.	External authors
Precision (P)	$\frac{TP}{TP+FP}$	.83	.93	.95
Recall (R)	$\frac{TP}{TP+FN}$	.81	.85	.96
F_1	$\frac{2\times P\times R}{P+R}$	.82	.89	.96

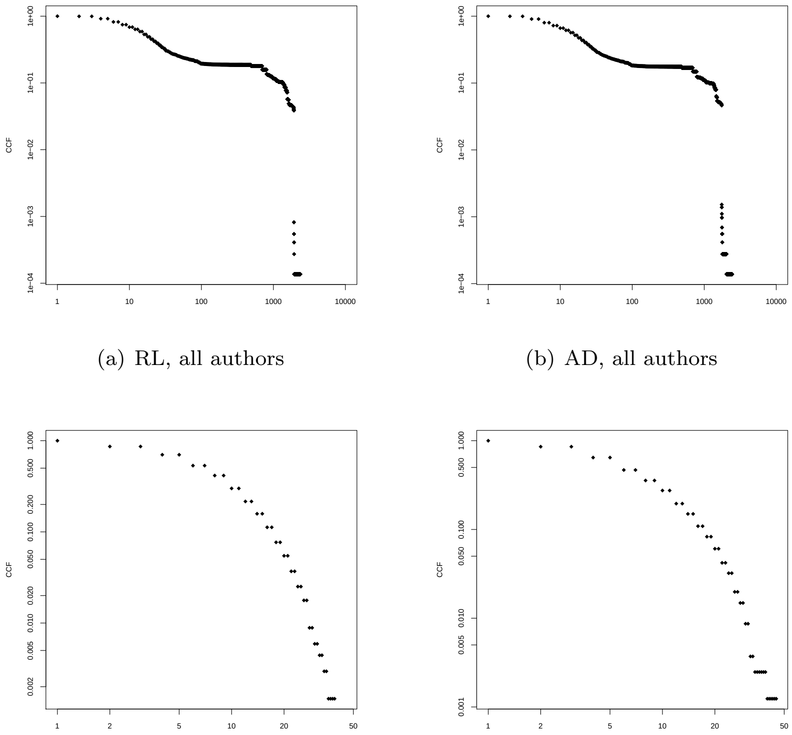
Table 6 RL and AD network statistics for all authors and for statisticians only.

	RL	**AD**		**RL**	**AD**
All authors					
# authors	7332	7230	Largest distance	14	16
# isolated	42	31	Average Path Length	5.29	5.17
# edges	474478	424545	Clustering Coeff.	0.88	0.91
Density	0.018	0.008	# of components (> 1 node)	35	58
Average degree	129.43	117.44	Giant component (%)	97.64	95.59
Statisticians					
# authors	677	808	Largest distance	13	14
# isolated	92	116	Average Path Length	5.46	5.53
# edges	1197	1346	Clustering Coeff.	0.26	0.24
Density	0.005	0.003	# of components (> 1 node)	16	15
Average degree	3.54	3.33	Giant component (%)	81.24	81.68

Table 6 reports the RL and AD network level statistics for all authors and considering only the subset of statisticians. In the case of all authors, the AD and the RL network structures are quite similar. The main differences can be noted on the number of isolates, the number of edges, the average degree (i.e. the average number of co-authors), and the number of disconnected components. The corresponding values are lower in the AD$_{NET}$ if compared with RL$_{NET}$, except the number of components, which is higher in AD$_{NET}$ than in RL$_{NET}$.

Basically, two main interacting effects are at work in shaping the network structures: merging and splitting of identities. In particular, for all authors, the merging affects the overall number of authors and links which are both lower in the case of AD (a drop of about 100 authors and 50,000 links in the AD$_{NET}$). The merge especially concerns some external authors, since the number of statisticians detected by the AD procedure is larger than the one registered in the RL output. The splitting jointly produces a reduction of the number of isolates and an increasing number of components.

Looking at the co-authorship networks among statisticians, merging and splitting act in opposite way. In this case, the splitting effect seems to play the most important role in shaping AD$_{NET}$ with respect to RL$_{NET}$ producing a higher number of nodes and edges, but also an increase in the number of isolates. Here, the splitting of the statistician identities is also enhanced by the exclusion of external authors who cannot connect couples of statisticians anymore. In addition, the splitting also produces a drop in the average degree in both networks; because some prominent authors are separated into different



(a) RL, all authors

(b) AD, all authors

(c) RL, only statisticians

(d) AD, only statisticians

Fig. 2 Observed Complementary Cumulative Degree Distribution of authors and of statisticians only in the RL and AD co-authorship networks. Horizontal axes: values of degree k; vertical axes: complementary cumulative function (CCF) describing the proportion of authors with degree greater than k.

identities, the splitting also reduces the presence of authors with high degree in both networks. In fact, upon inspecting the tail of the degree distribution, in Fig. 2, it can be noted that some outliers observed in the RL$_{NET}$ disappear in AD$_{NET}$.

Moving from network-level to node-level analysis and focusing only on the position of the statisticians, some changes occurred in RL$_{NET}$ and AD$_{NET}$. In Table 7, we report the rankings of the 10 prominent statisticians according to three centrality indices: degree, closeness, and betweenness. The degree ranking is slightly affected by the procedure. Degree values are basically lower in the AD step due to the splitting process, as already discussed. In fact, the ranking of betweenness and closeness – indices based on the geodesic distance – are largely affected by our procedure. In the AD$_{NET}$, only two statisticians maintained their position in the top 10 for betweenness, and only one for closeness. As noticed at the network level, including these two centrality measures,

Table 7 Top 10 statisticians ranking by centrality indices in the overall RL_{NET} and AD_{NET}. Capitalised names indicate statisticians present in top 10 ranking of both networks. Lower case names indicate statisticians present in the top 10 of only one network (if bolded they are only present in the top 10 of the AD_{NET}). The symbols ↑ and ↓ besides names indicates if statisticians increase or decrease their rank in the AD_{NET}, respectively.

		RL		**AND**	
Statistics	Rank[a]	Name	Value	Name	Value
Degree					
	1	POSTIGLIONE F	967	POSTIGLIONE F	878
	2	SANTAMARIA L	742	SANTAMARIA L	710
	3	BONETTI M	464	BONETTI M	448
	4	BIGGERI A	424	BIGGERI A	362
	5	ROMUALDI C	191	ROMUALDI C	187
	6	ROSATO R	183	ROSATO R	181
	7	CAVRINI G	141	VIGOTTI MA ↑	152
	8	MIGLIO R	124	CAVRINI G ↓	138
	9	VIGOTTI MA	112	MIGLIO R ↓	119
	10	SALMASO L	91	SALMASO L	89
Betweenness					
	1	BIGGERI A	0.207	BIGGERI A	0.166
	2	Mealli F	0.072	**Betti G**	0.057
	3	ROMUALDI C	0.050	SALMASO L ↑	0.050
	4	ROSATO R	0.047	ROMUALDI C ↓	0.049
	5	Bonetti M	0.044	ROSATO R ↓	0.037
	6	SALMASO L	0.040	MIGLIO R ↑	0.034
	7	MIGLIO R	0.039	**Grassia MG**	0.033
	8	CAVRINI G	0.033	CAVRINI G	0.032
	9	Muliere P	0.032	**Chiogna M**	0.030
	10	Zirilli A	0.032	**Billari FC**	0.029
Closeness[b]					
	1	BIGGERI A	0.256	BIGGERI A	0.305
	2	MEALLI F	0.252	**Betti G**	0.280
	3	Trivellato U	0.251	MIGLIO R ↑	0.268
	4	Lovison G	0.249	**Vigotti MA**	0.264
	5	MIGLIO R	0.247	**Muggeo V**	0.264
	6	Torelli N	0.246	**Lagazio C**	0.263
	7	Chiogna M	0.245	**Romualdi C**	0.262
	8	Bini M	0.244	**Rosato R**	0.261
	9	Rosina A	0.243	MEALLI F ↓	0.261
	10	Chiandotto B	0.242	**Postiglione F**	0.261

[a] Ranking is made only on statisticians.

[b] Closeness is computed on giant component.

the re-allocation of statisticians in different identities together with the exclusion of external authors mainly drives the pattern of relations found in the AD step.

Conclusions

In order to reach a better quality of co-authorship data, in the present contribution we adopted a two-step procedure by linking data retrieved from different bibliographic archives and by dealing with the name disambiguation issue.

Specifically, we focused on a target population composed of the Italian academic statisticians. The bibliographic data we used came from three archives covering different kinds of production authored by scientists and published in international as well as national journals and books. To obtain a complete unified co-authorship network, first a record linkage procedure was adopted. Therefore, particular attention was devoted to author name disambiguation to obtain correct identification of the statisticians included in the target population.

Even if the author name disambiguation is considered to be an open issue, the modified version of the procedure described in Strotmann et al. (2009) provided promising results for AD. We checked the accuracy of the results using classic performance measures as well as by comparing the co-authorship networks before and after the disambiguation step. Evaluation metrics showed a good performance of the adopted method.

However, if the purpose is to use network analysis to study co-authorship, the results may be carefully interpreted. Although in several applications author disambiguation is usually not applied (Wu and Ding 2013), the analysis on both RL and AD co-authorship networks for all authors and statisticians only, highlighted that the splitting and merging identities in our AD algorithm produced some non-negligible differences in network results, especially at individual level. The splitting can reduce network connectivity and affect statistics like the average degree. On the other hand, the merging can reduce the variety of network structures, thereby reducing the number of nodes and links. At individual level, besides the lowering of the degree values, splitting and merging mainly affect index values based on geodesic distance, such as closeness and betweenness. In general, the amount of splitting and merging effects– with their implications on network results –can be related to the values of the weight parameters in the evidence function we defined to connect vertices with the highest evidence. In this application, we assigned the same weight to the four available attributes to perform our AD algorithm, however other choices may be used depending on the purpose and/or on nature of the retrieved data.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Baxter, R., Christen, P., and Churches, T. (2003). A comparison of fast blocking methods for record linkage. In *ACM SIGKDD*, volume 3, pages 25–27.

Citeseer.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *Knowledge and Data Engineering, IEEE Transactions on*, 24(9):1537–1555.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., and Laender, A. H. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, 61(9):1853–1870.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Cuxac, P., Lamirel, J.-C., and Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 97(1):47–58.

de Carvalho, A. P., Ferreira, A. A., Laender, A. H., and Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management*, 2(3):289.

De Stefano, D., Fuccella, V., Vitale, M. P., and 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. and Zaccarin, S. (2015). Co-authorship networks and scientific performance: an empirical analysis using the generalized extreme value distribution. *Journal of Applied Statistics*, (ahead-of-print):1–18.

Domingo-Ferrer, J. and Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13(4):343–354.

Dong, X., Halevy, A., and Madhavan, J. (2005). Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 85–96. ACM.

Durham, E., Xue, Y., Kantarcioglu, M., and Malin, B. (2012). Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, 41(2):15–26.

Goyal, S., Van Der Leij, M. J., and Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of political economy*, 114(2):403–412.

Gurney, T., Horlings, E., and Van Den Besselaar, P. (2011). Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449.

Han, H., Zha, H., and Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference* *on*, pages 334–343. IEEE.

Hernandez, M. A. and Stolfo, S. J. (1995). The merge/purge problem for large databases. *Acm Sigmod Record*, 24(2):127–138.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2):193–215.

Imran, M., Gillani, S., and Marchese, M. (2013). A real-time heuristic-based unsupervised method for name disambiguation in digital libraries. *D-Lib Magazine*, 19(9):1.

Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., Sung, W.-K., and Lee, J.-H. (2009). On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Amy, Z. Y., and Fleming, L. (2014). Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). *Research Policy*, 43(6):941–955.

Liseo, B., Montanari, G. E., and Torelli, N. (2006). *Metodi statistici per l'integrazione di dati da fonti diverse*, volume 412. FrancoAngeli.

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

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*, 101(suppl 1):5200–5205.

Sadinle, M., Hall, R., and Fienberg, S. E. (2011). Approaches to multiple record linkage. In *Proceedings of International Statistical Institute*, volume 260.

Smalheiser, N. R. and Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, 43(1):1–43.

Strotmann, A., Zhao, D., and Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–20.

Torvik, V. I., Weeber, M., Swanson, D. R., and Smalheiser, N. R. (2005). A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology*, 56(2):140–158.

Veloso, A., Ferreira, A. A., Gonçalves, M. A., Laender, A. H., and Meira, W. (2012). Cost-effective on-demand associative author name disambiguation. *Information Processing & Management*, 48(4):680–697.

Ventura, S. L., Nugent, R., and Fuchs, E. R. (2015). Seeing the non-stars:(some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy.*

Wu, J. and Ding, X.-H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics*, 96(3):683–697.

Yan, S., Lee, D., Kan, M.-Y., and Giles, L. C. (2007). Adaptive sorted neighborhood methods for efficient record linkage. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 185–194. ACM.

Acknowledgements

The authors would like to thank Andreas Strotmann for providing details on the algorithm code adopted in Strotmann et al., 2009.