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

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

Abstract The aim of the present contribution is to merge bibliographic data for members of a bounded scientific community in order to derive a complete unified archive, with top-international and nationally oriented production, as a new basis to carry out network analysis on a unified co-authorship network. A two-step procedure is used to deal with the identification of duplicate records and the author name disambiguation. Specifically, for the second step we strongly drew inspiration from a well-established unsupervised disambiguation method proposed in the literature following a network-based approach and requiring a restricted set of record attributes. Evidences from Italian academic statisticians were provided by merging data from three bibliographic archives. Non-negligible differences were observed in network results in the comparison of disambiguated and not disambiguated data sets, especially in network measures at individual level.

✉ Vittorio Fuccella
vfuccella@unisa.it

Domenico De Stefano
destefano@units.it

Maria Prosperina Vitale
mvitale@unisa.it

Susanna Zaccarin
susanna.zaccarin@econ.units.it

[1] Department of Informatics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy

[2] Department of Political and Social Sciences, University of Trieste, Piazzale Europa, 1, 34127 Trieste, Italy

[3] Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy

[4] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Piazzale Europa, 1, 34127 Trieste, Italy

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 is "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 produce 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 on the U.S. patent inventors).

In this contribution, we aimed at merging bibliographic data for members of a "bounded" scientific community (i.e., a target population) in order to obtain a complete unified archive, containing both top-international and nationally oriented production, as a new basis to carry out network analysis. A two-step procedure is used to deal with the two aforementioned challenges in order to reach a better quality of co-authorship links. 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. In the second step, we addressed the problem of author disambiguation through an unsupervised method due to the lack of training data. Among the recent unsupervised methods (Cota et al. 2010; de Carvalho et al. 2011; Imran et al. 2013; Wu et al. 2014; Santana et al. 2015), we strongly drew inspiration from the procedure described in Strotmann et al. (2009), because it follows a network-based approach to create a co-authorship network and, in addition, it has the advantage of requiring a restricted set of record attributes (identifier, co-authors, venue). Therefore, it can be easily adapted to our case study given the aims and the information available in the unified archive we obtain after record linkage step.

The usefulness of the adopted procedure is showed 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 all kind of scientific production (De Stefano et al. 2013).

[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).

Besides checking the performance of the modified disambiguation procedure by using basic performance metrics, we mainly compare overall and individual network statistics computed before and after the disambiguation process.

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 "The procedure" provides details on the approach 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 "Results" section, we first discuss the accuracy of the adapted algorithm and then we compare the co-authorship networks constructed after the record linkage and the disambiguation steps. In the "Discussion and Conclusion" section, we provide final remarks and comments.

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, no matches, 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; Santana 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; Wu et al. 2014; Santana et al. 2015) or semi-supervised techniques, typically based on a small amount of labeled data with a large amount of unlabeled 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 part 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

We start from a case study focusing on a target population, i.e. the 792 academic statisticians (henceforth denoted by "statisticians") who have permanent positions in Italian universities, as recorded in the MIUR database at 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 of statisticians in Italy (single-authored vs co-authored and/or writing articles vs books and/or publishing in international vs national journals).

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). For this national source, the list of publications were directly provided by the Cineca.

In studying the influence of database characteristics on the co-authorship patterns of Italian statisticians, De Stefano et al. (2013) and De Stefano and Zaccarin (2016) 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.

For the WoS and CIS international databases, information was retrieved through a web form by specifying one or more parameters (author name, affiliation, publication title, etc.). Common information gathered from both interfaces were authors, title, year and kind of publication. Only for WoS details about subject categories, abstract, authors' affiliation were available.

International databases, usually containing high-impact publications on topics covered by the archive editorial policies, have been often used to study scientific collaboration

[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.

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[a]	5054	70.2

[a] Years of the project

within 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 project managers and members.

The procedure

As reported in the previous section, the three data sources contain 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. The details of the two steps are reported in the following.

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 several 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).

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.

As for "matches", we decided to relax the equality on d_T because we noticed that there were significant differences between titles referring to the same publication in different archives. Sources of differences were special characters, mathematical notations, data entry errors, etc. The 10 % threshold was chosen as a good compromise between the amount of records left for manual processing and false positives. This threshold, combined with the equality checks on d_A and d_Y, allowed us to successfully identify and reject cases of pairs of different records with similar titles (for instance, x and "a note on x").

The result of the matching step is reported in Table 2. The first two columns of the table report the count of "matched" and "possibly matched" pairs of records, respectively. The last column reports the amount of "total links found" after both the automatic matching and the manual check of "possible matches".

We used a specially designed interface to manually reconcile "possible matches". The manual analysis lasted a working day (about 8 h): a small amount of work due to the

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

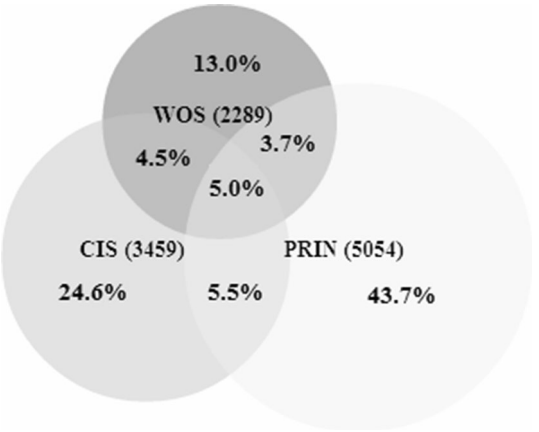

Fig. 1 The number and the percentage of publications in the unified archive after record linkage by data sources (the circle's size is proportional to the number of publications in each data source)

manageable size of our dataset. The feasibility of manual controls must be evaluated carefully with very large datasets. An important role is played by the interface, which can significantly affect work times. Some interfaces designed for RL have been described for instance in Christen (2008).

Lastly, we performed the reconciliation step, which allowed us to find a small number of discrepancies. Again, 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

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.

Starting from our main purpose of reconstructing a unique co-authorship network by using information available in different bibliographic archives, we addressed the problem of author disambiguation through an unsupervised method due to the lack of any training data. The method considered here is an adaptation of the well-established procedure proposed by Strotmann et al. (2009) in dealing with a peculiar disambiguation problem,

that is, reconstructing a disambiguated co-authorship network around "core" actors representing members of a "bounded" scientific community. We chose this procedure because it has the advantage of requiring a restricted set of record attributes (identifier, co-authors, venue) and has good success rates with respect to other results discussed in the related literature. Their algorithm uses a graph-based representation of author occurrences, each of which is associated to a graph node. An edge is added between two nodes every time their associated occurrences show some evidence of belonging to the same identity. The output identities are obtained by calculating the connected components of the graph, each connected component being a different identity. The addition of edges is performed in two different phases. The second phase was added as the authors realized that their algorithm had a *pessimistic behavior*, i.e., it separated known authors into several "individuals". In "Phase 2" a further fusion of occurrences is performed on the basis of collaboration with a third author.

A limitation of this method is the lack of misprint handling in name compatibility checking. Furthermore, the use of PubMed's *MeSH code* as an evidence to merge identities is only applicable in a limited number of cases and is not suitable to our case.

We improved the original approach first by handling misspellings and double names/surnames. Misspellings are one of the main sources of ambiguity in bibliographic archives and many disambiguation approaches implement some mechanism to automatically handle errors (Han et al. 2004).[4] Furthermore, we enhanced the use of record data to merge identities. To this aim, 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. These two additional evidences solved the pessimistic behavior and made superfluous the running of the second phase.

In Strotmann et al.'s algorithm, edges can be added only between nodes representing identities with compatible names, i.e. names that may refer to the same identity. Some occurrences have a full first name (expanded), others only have 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. Given this assumption, for instance, *Vittadini*, *G* and *Vittadin*, *G* were considered to be compatible, and not *Martini*, *C* and *Sartini*, *C*.

We decided to consider differences in only one character as the most frequent misspellings in our data involve this case. It is worth noting that our algorithm uses compatibility "transitively", allowing the detection of misspellings of more than one character. An example from our database is that of "Daria Mendol": "Mendol" and "Mendola" were merged at one step; "Mendola" and "Mendiola" were merged at a second step. As a result, also "Mendol" and "Mendiola" were merged, even though they differ of two characters. It is worth noting that although there may be false positives in the detection of compatibility between names in our procedure, these do not necessarily result in false positives in the process of disambiguation. In fact, to merge identities, further evidences must be present.

An author could appear in the publications under different names (synonymity), when an author has more than one first name (there are 89 of them out of the 792 in our population) or surname (14 cases with double surnames and 50 with compound surname with or without an apostrophe). In these cases, in the algorithm we relaxed the checks by considering as compatible two entries sharing at least one surname, or one first name

[4] For deepening on this problem, the reader can refer to Bilenko et al. (2003).

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*.

Lastly, we chose not to handle the case of completely different surnames, as it is rather infrequent. For instance, changes in the surname for women after the marriage is uncommon in the Italian system. In contrast, our method handles double-barreled surnames, which in many countries are used for married women.

The set of nodes was initially partitioned into two parts: those of abbreviated occurrences and those of expanded ones. Then, edges were added between nodes 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 nodes 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.

The data on which we worked were obtained by performing queries using author names on two of the three datasets (CIS and WOS), and publications not attributable to the queried authors were manually removed. Thus, the "Query Id" provides a very strong evidence that two compatible names refer to the same identity. Common keywords in the title are also a strong evidence, as they characterize the content of a publication. We used the well-established *Inverse Document Frequency* (IDF) metric to distinguish keywords (most important terms) from stop-words (Robertson 2004). In particular, we proceeded as follows:

- We used the whole set of titles to calculate IDFs of all terms;
- We established a threshold for filtering stop-words using a small subset of the titles.

Unfortunately, other textual data (e.g., the abstract) were neglected due to their unavailability in two data sources (PRIN and CIS). As abstracts contain longer text than titles, their availability could help to uncover more bonds than simply using titles. Furthermore, we didn't have the possibility to add evidences based on other attributes, e.g. the sub-field of specialization of the researcher, because these additional details were available only for the Italian statisticians.

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 established through TF-IDF statistic, which assigns greater

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

importance to infrequent words and penalises those that are particularly common and depends on the length of the titles.

The weights in the above formula were set to 0.25, in order to give the same weight to each of the four functions. Uniform weights were chosen for illustrating the general case and moreover in order to reproduce the lack of additional information on some of the four features with respect to the others. Whenever a researcher intends to emphasize some of the used features, she can set different weights.[5]

Results

The performance of our procedure was evaluated by first providing the traditional 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).

Since we slightly modified the original algorithm, we compare the accuracy of the adapted disambiguation approach with the success rates reported by Strotmann et al. (2009). Indeed, given the list of individuals already correctly assigned, we computed 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). The three measures of performance (see Table 5) defined according to these quantities were 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).

Two approaches are usually followed to derive these measures: (1) to evaluate the accuracy over a simulated dataset in which the true author's identity is known (Milojević 2013) or (2) 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). In our case, focusing on our target population, we adapted the latter approach as follows:

[5] For instance, Lee et al. (2005) and Santana et al. (2015) supposed different weights according to the discriminative capability of the attributes.

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 providing 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 correctly 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. 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 random stratified sample of 34 authors was selected from the list of statisticians found after the record linkage step (i.e., the 5 % of 677 statisticians retrieved after this step). The total sample size 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 of our adapted approach (Table 5) were in line to the success rates reported from the original algorithm (Strotmann et al. 2009) and quite comparable to the best results others have reported in the recent literature (Kang et al. 2009; Wu et al. 2014; Santana et al. 2015). In particular, in the case of the population, the values of around 0.80 represent the lower bound 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[6] (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

[6] 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 [DLS], compound surnames and an apostrophe [CLS/A], and compound surnames, double first names and an apostrophe [CN/A], algorithm results and identity assignment

Target population	Algorithm results	Identity assignment
DLS		
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
CLS/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)	$\frac{TP}{TP+FP}$	0.83	0.93	0.95
Recall (R)	$\frac{TP}{TP+FN}$	0.81	0.85	0.96
F_1	$\frac{2\times P\times R}{P+R}$	0.82	0.89	0.96

the co-authorship networks built on author identities—7332 authors and 677 statisticians—resulting from the record linkage step (RL_{NET}).

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

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	474.478	424.545	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

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}.

The changes detected are explained by the fact that merging/splitting occurrences is not equally distributed between statisticians and external authors. Merging affects especially external co-authors and (looking at the degree distribution graphs) especially those in large co-authored publications, therefore links are merged too (union not sum) and the overall average degree will drop. However splitting mostly affects statisticians which are of course still the "core" of the network producing an increasing number of components.

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 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} (Fig. 2a, c) disappear in AD_{NET} (Fig. 2b, d).

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

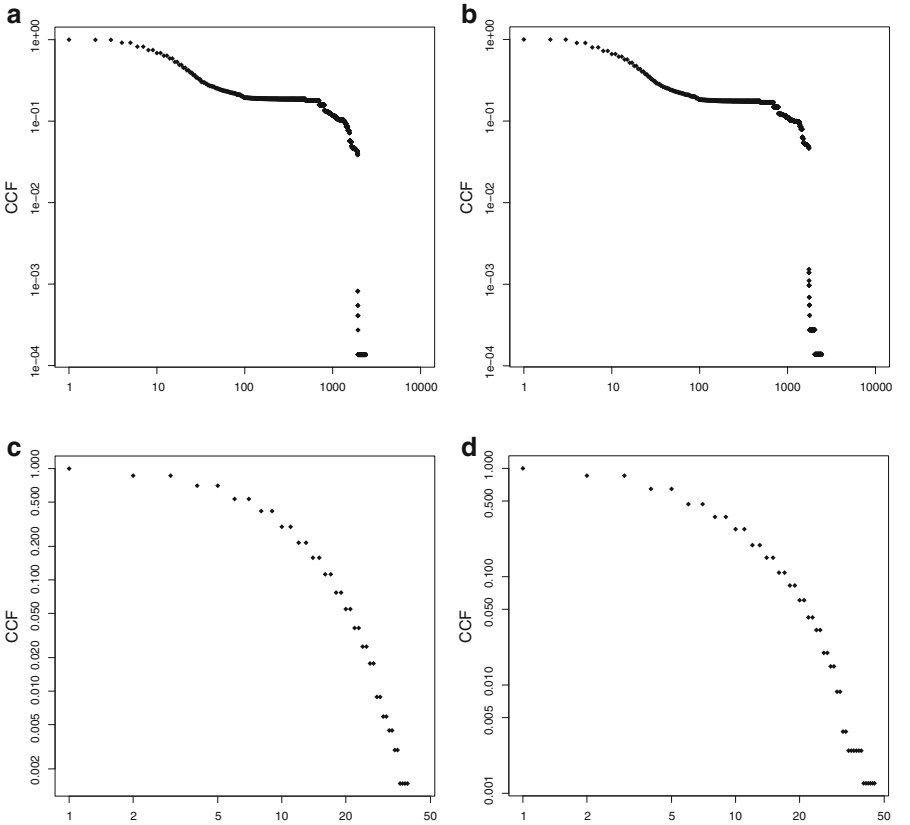

Fig. 2 **a** RL, all authors; **b** AD, all authors; **c** RL, only statisticians; **d** AD, only statisticians. Observed complementary cumulative degree distribution of authors and of statisticians only in the RL and AD co-authorship networks. *Horizontal axes* values of degree k; *vertical axes* complementary cumulative function (CCF) describing the proportion of authors with degree greater than k

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, 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.

Although these results cannot be generalized to other cases, mainly because most disambiguation procedure results are not usually evaluated also on network measures, this information intends to show the effect of the adopted procedure on the individual centrality of the members of our target population. Of course also such changes are related to the splitting/merging effects but how these effects work on each node is not predictable, being dependent on both the position and the eventually disambiguated identity.

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

Statistics	Rank[a]	$\mathbf{RL_{NET}}$		$\mathbf{AD_{NET}}$	
		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

Discussion and conclusions

We have proposed a procedure able to merge bibliographic data for members of a target population in order to obtain a unified archive as a new basis to carry out network analysis. In particular, we adapted the unsupervised approach for author disambiguation task on the

basis of Strotmann et al.'s (2009) algorithm. We checked the accuracy of our modified version of the procedure using classic performance as well as by comparing the co-authorship networks before and after the disambiguation step.

The adapted approach was tested within a case study focusing 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 scientific community under analysis.

Although our approach is evaluated through a relatively small "bounded" scientific community in a narrow field and its generalizability is limited, the results demonstrate that our adapted algorithm obtains similar results in terms of effectiveness to the best results others have reported in the literature, and then the viability of our approach could be tested in other fields starting with any given databases. Furthermore, in line with the original algorithm, the author disambiguation approach was adopted to create specifically a co-authorship network for a research study on scientific collaboration.

As a general result, if the purpose is to use network analysis tools to describe the derived co-authorship relations, the AD 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 weights in the evidence function we defined to connect nodes with the highest evidence.

Hence, although we are aware that any disambiguation procedure has more general aims than network construction, it could be an added value to provide, in the algorithm evaluation phase, some information about the behaviour of the resulting network.

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

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47.

Baxter, R., Christen, P., & Churches, T. (2003). A comparison of fast blocking methods for record linkage. In *ACM KDD Workshops* (Vol. 3, pp. 25–27).

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *5*, 16–23.

Christen, P. (2008). Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1065–1068. ACM.

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

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & 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., & 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., & 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., & 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., & 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.

Domingo-Ferrer, J., & 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., & Madhavan, J. (2005). Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international conference on management of data*, pp. 85–96. ACM.

Durham, E., Xue, Y., Kantarcioglu, M., & 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., & 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., & 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., & Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy*, *114*(2), 403–412.

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

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Digital Libraries, 2004. Proceedings of the 2004 joint ACM/IEEE conference on*, pp. 296–305. IEEE.

Han, H., Zha, H., & 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*, pp. 334–343. IEEE.

Hernandez, M. A., & 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., & 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., et al. (2009). On co-authorship for author disambiguation. *Information Processing and Management*, *45*(1), 84–97.

Lee, D., On, B.-W., Kang, J., & Park, S. (2005). Effective and scalable solutions for mixed and split citation problems in digital libraries. In *Proceedings of the 2nd international workshop on Information quality in information systems*, pp. 69–76. ACM.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., et al. (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., & Torelli, N. (2006). *Metodi statistici per l'integrazione di dati da fonti diverse* (Vol. 412). Milan: 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.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.

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

Santana, A. F., Gonçalves, M. A., Laender, A. H., & Ferreira, A. A. (2015). On the combination of domain-specific heuristics for author name disambiguation: The nearest cluster method. *International Journal on Digital Libraries*, *16*(3–4), 229–246.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1–43.

Strotmann, A., Zhao, D., & 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., & 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., & Meira, W. (2012). Cost-effective on-demand associative author name disambiguation. *Information Processing and Management*, *48*(4), 680–697.

Ventura, S. L., Nugent, R., & 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*, *44*(9), 1672–1701.

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, *101*(3), 1955–1972.

Wu, J., & 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., & 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* (pp. 185–194). ACM.