

# Co-authorship Patterns of Italian Statisticians by Combining Different Data Sources

## *Rete di Collaborazione degli Statistici Italiani attraverso Integrazione di Fonti Diverse*

Domenico De Stefano, Vittorio Fuccella, Maria Prosperina Vitale and Susanna Zaccarin

**Abstract** The present paper aims at describing the scientific collaboration patterns of the Italian academic statisticians by exploiting the information retrieved through different bibliographic archives (ISI-WoS, Current Index to Statistics), and the bibliographic information of the database of nationally funded research projects. A data merging procedure is adopted to obtain a unified dataset containing both top-international as well as nationally oriented production. The unified dataset will be used as basis for network analysis.

**Abstract** *Obiettivo del presente lavoro è descrivere la struttura della collaborazione scientifica tra gli statistici accademici italiani attraverso l'utilizzo di diversi archivi bibliografici (ISI-WoS, Current Index to Statistics) e le informazioni bibliografiche presentate nei progetti PRIN finanziati. Una procedura di fusione è adottata al fine di ottenere un'unica base dati contenente la produzione scientifica sia internazionale che nazionale. La rete di collaborazione così derivata è studiata mediante le tecniche di analisi di rete.*

**Key words:** Bibliographic archive, Author disambiguation, Scientific collaboration, Social Network Analysis

---

Domenico De Stefano  
Dept. of Economics, University Ca' Foscari Venezia,  
e-mail: domenico.destefano@unive.it

Vittorio Fuccella  
Dept. of Informatics, University of Salerno,  
e-mail: vfuccella@unisa.it

Maria Prosperina Vitale  
Dept. of Economics and Statistics, University of Salerno,  
e-mail: mvitale@unisa.it

Susanna Zaccarin  
Dept. of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste,  
e-mail: susanna.zaccarin@deams.units.it

## 1 Introduction

Among the widespread applications of Social Network Analysis (SNA), the study of scientific collaboration within disciplines is one of the topic that most benefited from this approach in the last decades. It is straightforward to think about collaboration among scientists as a network, in which the actors are scholars and ties may be represented by various forms of collaboration among them. The interest in collaboration networks lies in the fact that long-term trends in scientific work as well as scientist's productivity might closely depend on the topological features of disciplinary networks. Mainly thanks to the availability of international bibliographic databases, several seminal studies in various fields focused on co-authorships as a proxy of scholars' collaborative skills (e.g. [2] for Physics, Mathematics and Neurosciences).

The present paper aims at analyzing co-authorship patterns in Statistics, focusing on Italian academic statisticians belonging to one of the five Statistics subfields, as recorded in the Italian Ministry of University and Research (MIUR) database in March 2010. Attention to this community derives from several motivations. Unlike other disciplines, co-authorship behaviour in Statistics has not yet been investigated. The field of Statistics presents some characteristics common to natural sciences as well as social sciences. Even if it is usually considered in the stream of social sciences - especially in the Italian academic tradition - it plays a central role in all sciences in view of the importance of statistical methods in everyday applications [5]. Therefore, it is of interest to examine what emerging pattern describes the diffusion of statistical knowledge –although limited to a country level community.

To the best of our knowledge, only few studies have been specifically devoted to the Statistics field. Baccini *et al.* [1] explore the structural properties of the network generated by the interlocking editorships of editorial boards around the 81 statistical journals included in the category “Statistics & Probability” of Web of Science (WoS). Evidences of a very compact network are found. This is interpreted as the result of a common perspective about the appropriate methods for investigating the problems and constructing the theories in the domain of Statistics. De Battisti and Salini [6] investigate the publication style of Italian academic statisticians from several data sources (WoS, Scopus, Current Index to Statistics –CIS–, and Google Scholar) through robust clustering techniques. The authors recognize that the use of a single data source can led to biased and partial results. Lastly, co-authorship networks of Italian statisticians derived from three bibliographic archives (WoS, CIS, and bibliographic information retrieved from the database of nationally funded research projects, PRIN) were extensively analysed in [7] with the specific aim to explore the effect of different data sources on the resulting network patterns. Results provide evidence of distinct collaboration patterns among statisticians, as well as distinct effects of scientist network positions on scientific performance, by both Statistics subfield and data source.

Because each data source presents peculiar characteristics affecting network results, in this study we aim at merging bibliographic data of the three databases to obtain a complete unified archive containing both top-international as well as nation-

ally oriented scientific production and to use it as a new basis for network analysis. Specifically, two main challenges are managed to obtain a single co-authorship network: how to combine information from heterogeneous sources by identifying and linking duplicate records (i.e. record linkage), and how to deal with issues related to authors *synonyms* and *homonimies* issues (i.e. name disambiguation).

## 2 Co-authorship networks from different bibliographic archives

In the previous study [7], the aim was to analyze the scientific collaboration of Italian academic statisticians by comparing network results obtained using three bibliographic archives in the definition of co-authorship relations. Before obtaining co-authorship data, a data cleansing step was carried out within the bibliographic data gathered through the data sources in order to identify clean publication data set for each statistician in the target population. Specifically, the removal of publication records erroneously attributed to authors due to the homonymy problem, was especially present in WoS.

Formally, for each data source the co-authorship network was derived from the matrix product  $\mathbf{Y} = \mathbf{A}\mathbf{A}'$ , where  $\mathbf{A}$  is a  $n \times p$  affiliation matrix, with elements  $a_{ik}$  assuming the value 1 if  $i \in \mathcal{N}$  (the set of  $n$  authors) authored the publication  $k \in \mathcal{P}$  (the set of  $p$  scientific publications observed on the  $n$  authors), 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  $\mathcal{P}$ , 0 otherwise. The binary version of  $\mathbf{Y}$ , setting all entries in the valued adjacency matrix greater than zero to 1, was used in the analysis.

The collaboration patterns of Italian statisticians were influenced to some extent by the features of each data source. First, a different authors coverage rate was obtained for all statisticians, and in particular for some subfields and by considering the academic ranking. The lowest value is observed in the international database WoS (61.0%), especially for the subfields of Demography (40.0%), Economic Statistics (42.5%) and Social Statistics (50.0%). The full professor coverage rate is also lower in WoS database (65.0%) with respect to the other two sources (respectively 90.0% in CIS and 83.0% in PRIN).

WoS appeared as the data source in which the average number of co-authors for each statistician is extremely high, being affected by the presence of few statisticians with a large number of co-authors.

Patterns consistent with well-established network structures are found out in CIS database. CIS captured internationalisation openness by research topics and publication style, while WoS mainly captured the tendency towards an interdisciplinary behavior. Finally, PRIN combined some of CIS and WoS characteristics, although referred only to the selected publications by project's managers and members.

In summary, the choice of bibliographic archives to be considered in order to study scientific collaboration through co-authorship relations must be carefully examined according to the aims of the analysis.

### 3 Towards a combined co-authorship network

To get a co-authorship network by exploiting the usefulness of the three data sources, two main challenges have to be addressed: 1) how to combine information from heterogeneous sources by identifying and linking duplicate records (i.e. record linkage), and 2) how to deal with issues related to authors *synonyms* and *homonymies* (i.e. name disambiguation). To this aim, we first adopt a record linkage procedure to merge the three bibliographic archives in one unique database then we cope with authors disambiguation to guarantee the quality of the data for network analysis.

#### 3.1 Data sources linkage

The record linkage of metadata in Digital Libraries (DLs) is a very sensitive issue. It refers to “the task of identifying records from disparate data sources that refer to the same entity” [10, p. 245], often used to define integrated information systems in statistical setting [13]. In recent years computer-oriented record linkages methods, which ensure a high efficiency and scalability on large data sets, are reported in the literature [8, 9, 17, 4].

Given the relatively small number of records in the three data sources, 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 sources, we proceeded with the commonly used approach of matching the sources in pairs and then perform a reconciliation of possible discrepancies [14]. In particular, we used the following distance functions on each of the key fields:

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

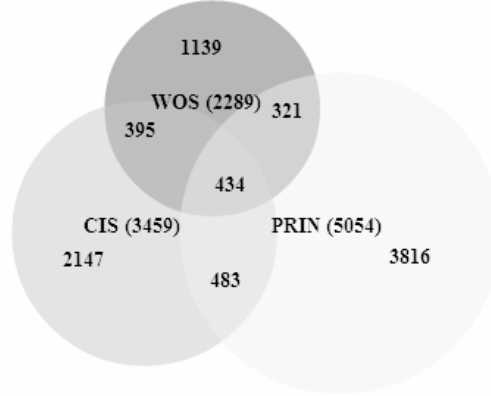
$$d_T = Ld(t_1, t_2) / \max(|t_1|, |t_2|)$$

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:** absolute value of the difference between the years of publication ( $d_Y$ ).

All strings were lower-cased before any comparison. The overall distance is defined as a 3-tuple  $(d_A, d_T, d_Y)$ , where each element is 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  $d_Y = 0$ . The couples having  $d_T < 20\%$  and  $d_A \leq 1$  (except those already automatically linked) were manually inspected to establish whether to link them. The resulting unified archive contains 8735 publications, and its composition is shown in Fig.1.



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

### 3.2 Author disambiguation

Once the unique database is defined, the problem of author's disambiguation should be addressed by different methods. It “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)” [16, p. 680]. More formally [3, 16], given the set of citations  $C = \{c_1, c_2, \dots, c_k\}$ , where each citation  $c_i$  contains some author's attributes (such as author names, affiliation, work title and publication venue title), the name disambiguation task is to define a function to partition the set of citations into  $n$  sets  $\{a_1, a_2, \dots, a_n\}$ , where each partition  $a_i$  contains the citations of  $i$ -th author.

A myriad of recent studies are devoted to name disambiguation methods in bibliographic DLs (for a recent survey see [11]) in computer science, sociological and linguistic setting by covering supervised, unsupervised or semi-supervised techniques. Here due to the lack of training data, we opt for an unsupervised method and in particular for the techniques described in [12, 15]. These methods have the advantage of operating directly on a collaboration network and of requiring a restricted set of data attributes (co-author names, work title, and publication venue).

Therefore, they are very well suited to our needs given the information available in the unified archive. Basically, we follow a two phase procedure. In the first phase, a clustering step is executed: the method identifies the set of name occurrences which are likely to refer to a single individual; in the second phase, the method uses collaboration links to further merge names belonging to connected components into single identities.

The combined co-authorship network will then be used to analyse collaborative behaviour among Italian statisticians in a comparative way with respect to the findings in [7] referred to the three data sources.

## References

1. Baccini, A., Barabesi, L., Marcheselli, M.: How are statistical journals linked? A network analysis. *Chance* **22**, 35–45 (2009)
2. Albert, R., Barabási, A.-L.: Statistical Mechanics of Complex Networks. *Rev. of Mod. Phys.* **74**, 47–97 (2002)
3. Carvalho, A.P., Ferreira, A.A., Laender, A.H.F., Goncalves, M.A.: Incremental unsupervised name disambiguation in cleaned digital libraries. *Journal of Information and Data Management* **2**, 289–304 (2011)
4. Christen, P.: A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering* **24**, 1537–1555 (2012)
5. Cox, D.R.: The Current Position of Statistics: A Personal View. *International Statistical Review* **65**, 261–276 (1997)
6. De Battisti, F., Salini, S.: Robust analysis of bibliometric data. *Statistical Methods & Applications* **22**, 269–283 (2013)
7. De Stefano, D., Fuccella, V., Vitale, M.P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks* **35**, 370–381 (2013)
8. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing* **13**, 343–354 (2003)
9. Dong, X., Halevy, A., Madhavan, J.: Reference Reconciliation in Complex Information Spaces. *Proceedings of SIGMOD '05*, 85–96 (2005)
10. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion* **13**, 245–259 (2012)
11. Ferreira, A.A., Goncalves, M.A., Laender, A.H.F.: A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record* **41**: 15–26 (2012)
12. Imran, M., Gillani, S.Z.H., Marchese, M.: A Real-time Heuristic-based Unsupervised Method for Name Disambiguation in Digital Libraries. *D-Lib Magazine*, **19** (2013)
13. Liseo, B., Montanari, G.E., Torelli, N. (a cura di): *Metodi statistici per l'integrazione di dati da fonti diverse*. Franco Angeli, Milano (2006)
14. Sadinle, M., Hall, R., and Fienberg, S.E.: Approaches to multiple record linkage. *Proceedings of International Statistical Institute* **260** (2011)
15. Strotmann, A., Zhao, D. and Bubela, T.: Author Name Disambiguation for Collaboration Network Analysis and Visualization. *Proc. Am. Soc. Info. Sci. Tech.* **46**, 1–20 (2009)
16. Veloso, A., Ferreira, A.A., Goncalves, M.A., Laender, A.H.F., Meira Jr., W.: Cost-effective on-demand associative author name disambiguation. *Information Processing and Management* **48**, 680–697 (2012)
17. Yan, S., Lee, D., Kan, M.-Y., Giles, L.C.: Adaptive Sorted Neighborhood Methods for Efficient Record Linkage. *Proceedings of JCDL '07*, 185–194 (2007)