Social Networks 35 (2013) 370–381

Contents lists available at SciVerse ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

The use of different data sources in the analysis of co-authorship networks and scientific performance

Domenico De Stefano [a], Vittorio Fuccella [b], Maria Prosperina Vitale [c,*], Susanna Zaccarin [a]

[a] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy
[b] Department of Informatics, University of Salerno, Italy
[c] Department of Economics and Statistics, University of Salerno, Italy

ARTICLE INFO

Keywords:
Bibliometric databases
Co-authorship data
Network topology
Scientific performance
h-Index
GEV model

ABSTRACT

Scientific collaboration is usually derived from archival co-authorship data. Several data sources may be examined, but they all have advantages and disadvantages, especially when a specific discipline or community is of interest. The aim of this paper is to explore the effect of the use of three data sources – Web of Science, Current Index to Statistics and nationally funded research projects – on the analysis of co-authorship networks among Italian academic statisticians. Results provide evidence of our hypotheses on distinct collaboration patterns among statisticians, as well as distinct effects of scientist network positions on scientific performance, by both Statistics subfield and data source.

1. Introduction

Collaboration in science is a complex phenomenon which affects scientific productivity in various ways (Lee and Bozeman, 2005), as well as knowledge diffusion within and between disciplines. Collaboration is considered to be a key element in the advancement of knowledge, because scientists in collaboration networks share ideas, use similar techniques, and influence each other's work. By means of collaboration, scientists may benefit by both technological expertises and team work synergy, thus improving the quality and quantity of their research output. As empirical evidence, collaboration among scientists is increasing in all disciplines (e.g., Babchuk et al., 1999; Glanzel and Schubert, 2004; Kronegger et al., 2011).

In this stream of research, Social Network Analysis (SNA) has become the privileged theoretical and statistical approach to study the typical collaboration patterns within disciplines (for instance, see Burt, 1978/1979, and Moody, 2004 for Sociology; Albert and Barabási, 2002, and Newman, 2004 for Physics and Biomedical research; and Goyal et al., 2006 for Economics). 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 scientific collaboration among them. Thanks to the availability of international bibliographic databases, the most frequent way of specifying such networks is to take into account formal research activities, especially co-authorship (i.e., co-production of scientific publications)[1].

The present paper deals with network analysis of co-authorship patterns in Statistics, focusing in particular on the population of academic statisticians in Italy, that is, those scientists classified as belonging to one of the five Statistics subfields: Statistics, Statistics for Experimental and Technological Research, Economic Statistics, Demography, and Social Statistics.

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 Italian academic tradition – it plays a central role in all sciences in view of the importance of statistical methods in everyday applications. As reported by Leti (2000, p. 188): "The new natural science was made possible by the invention and scientific use of instruments which went beyond man's capabilities in their examination of nature. Similarly, Statistics as a method, by superseding human inability to quantify collective phenomena, permitted greater insight into these phenomena (originally those concerning the state and society). The new natural sciences and Statistics followed the same approach, shared a mathematical basis, and pursued both scientific and practical aims". Similar arguments are also reported in Kagan (2009) when he proposed nine dimensions

* Corresponding author. Via Giovanni Paolo II 132, IT 84084 Fisciano (SA), Italy. Tel.: +39 089962211; fax: +39 089962049.
E-mail address: mvitale@unisa.it (M.P. Vitale).

[1] There is a considerable amount of work using SNA applied to citation networks in many domains. In a citation network the "actors" are papers and the (directed) ties between them are citations of one paper by another (e.g., Garfield, 1979; Hummon and Doreian, 1989; Hummon and Carley, 1993).

0378-8733/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2013.04.004

to compare research approach in natural sciences, social sciences and humanities. Furthermore, although social and natural scientists work both in and outside of traditional lab settings, "the rise of large-scale data collection efforts suggests a team-production model" (Moody, 2004, p. 217) similar to the typical one that mainly characterises the scientific output production in natural sciences.

Statistics is also unique with respect to the other social sciences, since several problems in different disciplines may be addressed by its methods (Cox, 1997). Therefore, it is of interest to examine what emerging pattern describes the diffusion of statistical knowledge – although limited to a country level community.

It is relevant to trace this specific target population in high-impact journal international databases and to reveal the influence on the resulting co-authorship patterns related to distinct data sources. For these purposes, two international databases, one general (Web of Science, WoS) and one thematic (Current Index to Statistics, CIS) are examined here, together with bibliographic information retrieved from the Italian Ministry of University and Research (MIUR) database of nationally funded research projects (PRIN).

We provide several research hypotheses on the resulting collaboration patterns of Italian academic statisticians, regarded as a whole group, and also taking into account the five subfields into which the group is organised. Following seminal papers on co-authorship analysis (in particular, Albert and Barabási, 2002; Moody, 2004; Newman, 2004; Goyal et al., 2006) to allow comparisons, this study adds some substantial elements:

- it analyses a target population (Italian academic statisticians) involved in a discipline (Statistics) which is not yet fully explored in terms of its scientific collaboration behaviour. In addition, the specialised subfields within the whole discipline may be described by several cooperative patterns, depending on the level of interdisciplinarity characterising scientists' activities;
- it considers three data sources. In general, we assume that the collaboration structure, and hence knowledge flows, in scientific communities depends to a great extent on the kinds of publications pertaining to the various archives considered for network construction;
- it explores the effects of authors' network positions on scientific performance as measured by the *h*-index. For this aim, a generalised extreme value distribution (GEV) is fitted, to take into account the particular distribution of this index, which is usually highly skewed and heavy-tailed.

The paper is organised as follows: Section 2 presents the framework linking network structures to the diffusion of knowledge in scientific communities, and reports the main empirical results related to network topologies observed in several disciplines. After a description of the data sources used to collect co-authorship data on Italian academic statisticians, Section 3 describes data retrieval and cleansing in detail. Authors' coverage rates and publication characteristics in the three data sources are presented. Section 4 illustrates our research hypotheses on scientific collaboration patterns and their influence on scientific performance. In Section 5, the co-authorship trend and networks of Italian academic statisticians are analysed and results on highly connected statisticians are given. The relationship between authors' *h*-index and their network positions is modeled. Section 6 concludes, with a discussion and final remarks.

2. Co-authorship networks and patterns of collaboration in scientific communities

Scientific collaboration is a mix of informal mechanisms (e.g., advices, face-to-face contacts, exchange of personal knowledge), and formal activities (e.g., writing papers, participating in research projects) among scientists involved in producing knowledge, as suggested in Lievrouw et al. (1987), Liberman and Wolf (1997), and Liberman and Wolf (1998). Direct interviews can be very useful to gain insights on informal collaboration,[2] while archive data can provide good information on several kinds of formal collaboration. Although data in on-line archives have not been collected for network studies, they represent a common way of retrieving information on co-authorship. Co-authorship is a partial indicator of scientific collaboration (Katz and Martin, 1997), but it describes one aspect of major formal intellectual cooperation (e.g., Melin and Persson, 1996; Glanzel and Schubert, 2004).

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} 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. Let G be the network described by the adjacency matrix $\mathbf{Y}$.

The interest in analysis of co-authorship networks lies in the fact that collaborative behaviour within a scientific community closely depends on the topological features of G. In particular, a frequent finding in co-authorship networks is that they are consistent with some theoretical network models with well-defined topological and relational properties, which have a meaningful interpretation in terms of knowledge diffusion.

Simplest network models start from the idea that the connections between actors occur at random, as in the Erdos–Renyi random graphs (*ER*s), a family of networks in which the probability of a tie between actors' pairs is π.[3] *ER*s represent the baseline model to assess evidence of non-random behaviours in the observed networks.

Empirical evidence shows that co-authorship networks are usually non-random, because they tend to exhibit distinctive statistical properties deriving from the peculiar mechanisms which generate ties. In particular, small-world (Watts and Strogatz, 1998) and scale-free (Albert and Barabási, 2002) configurations are the theoretical non-random models most frequently emerging in co-authorship.

Networks consistent with a small-world configuration have high node connectivity with low average distance among regions of the network – i.e., the average path length, $\ell(G)$, is not greater than the value observed in random networks of equal size – together with a high tendency towards actor clustering. Specifically, in small-world networks, the clustering coefficient, $\Gamma(G)$, is much larger than that measured among nodes in a random network. The simultaneous presence of dense local clustering with short network distances in co-authorship networks indicates a mechanism which can facilitate knowledge flows among actors. In these networks, small-world patterns can also support disciplinary fractionalisation and specialty areas, clustered into distinct groups of scientists (Moody, 2004), mainly due to scientists' research group membership, university affiliations or geographic proximity.

The consistency with a "scale free" topology, instead, implies the existence of a peculiar tie formation mechanism named preferential attachment. In co-authorship networks, this mechanism formally accounts for the tendency to interact with the best connected authors (i.e., actors with the highest degree, usually

[2] For instance see Lazega et al. (2008) for the construction of advice networks at individual and institutional level within the "elite" of French cancer researchers.

[3] In ER random graphs, the degree of any given node follows a binomial distribution, which becomes a Poisson for $n \to \infty$. This feature is quite unrealistic in real networks. A more flexible model for random graphs is the so-called configuration model (CM) (Bender and Canfield, 1978).

called "star" authors). If the actor degree distribution follows a power law, then a scale-free structure emerges. Basically, there are two types of power law distributions. The first is defined by the probability distribution function (Nicholls, 1986):

$$P(x) = Cx^{-\alpha} \quad (1)$$

where $P(x)$ is the degree distribution (i.e., the proportion of nodes in the network with degree x), C is a normalising constant, and α is the power law parameter, ranging in a predetermined interval (typically $2 < \alpha < 3$). Since C is a constant function, Eq. (1) holds for all values of x.

Clauset et al. (2009) affirm that empirical data follow a power law distribution only for values of x above some lower bound x_{min}. Then, provided $\alpha > 1$, it is straightforward to calculate the normalising constant and Eq. (1) becomes:

$$P(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha} \quad (2)$$

In Eq. (2), x it is assumed to be continuous (for discrete variable x, see Clauset et al., 2009, p. 3).

In the literature, clear evidence of small-world properties have been observed in Economics (Goyal et al., 2006) and Physics (Newman, 2004). Physics, Mathematics and Neurosciences (Albert and Barabási, 2002), and Economics (Goyal et al., 2006) also show statistical properties consistent with a preferential attachment mechanism (although not all have a strictly power-law distribution). Sociology is the one exception, because it is better represented by an integrated (cohesive) collaboration network structure resembling a random network (Moody, 2004).

The findings for these disciplines could reflect the differences in the way research is done and internal organisation of disciplines. Natural sciences are mainly characterised by the use of quantitative methods, while social sciences consider a mix of quantitative and qualitative methods requiring different level of collaboration. Specifically, in Sociology "quantitative work is more likely to be coauthored than non-quantitative work" and "the coauthorship pattern shows a steadily growing cohesive core, suggesting that while authors might specialise their skills marry well with others creating an integrated collaboration network" (Moody, 2004, p. 235). Instead, for example in Biology (Newman, 2004, p. 5201) states: "biological research consisting often of work by large groups of laboratory scientists".

Studies focusing on specific scientific communities at country level, such as Italian academic economists (Maggioni and Uberti, 2011) and Slovenian scientists belonging to Physics, Biotechnology, Mathematics and Sociology (Kronegger et al., 2012), show evidence of small-world structure. In addition, for Slovenian scientists, some features of preferential attachment principle have only been confirmed for Mathematics and Sociology.

3. Data sources on co-authorship for Italian academic statisticians

Seminal studies in scientific collaboration are based on international databases containing mainly high-impact publications (for instance, *Sociological Abstracts* in Moody, 2004, MEDLINE in Newman, 2004, and Econlit in Goyal et al., 2006). These bibliographic databases allow exploration of the collaboration patterns among scientists working on topics covered by the editorial policies on which the archives are based. The advantages of using such data sources are that they are relatively inexpensive, do not impose a burden on informant time and effort, and may be less prone to missing data and inaccuracy problems.

If the interest is to describe collaboration in a target population involved in a scientific field and/or affiliated to a specific institution, the main problem in using international databases is the partial coverage of scientists' production. Writing articles or books and publishing in international or national journals may depend on discipline specialty (Hicks, 1999) and community traditions. In this regard, thematic and local research archives may be more complete because they allow to consider the entire scientists' output (books, articles in local journals, technical reports, book chapters).

Our target population is composed of the 792 academic statisticians who have permanent positions in Italian universities, as recorded in the MIUR database in March 2010[4], belonging to the five subfields (Table 1): Statistics (Stat), Statistics for Experimental and Technological Research (Stat for E&T), Economic Statistics (Economic Stat), Demography (Demo), and Social Statistics (Social Stat).

Similar co-authorship studies focusing on specific scientific communities can be found in the recent literature. Among others, we mention studies on: co-authorship networks of Italian academic economists recognised by the Econlit database (Maggioni and Uberti, 2011); the effect of co-authors past productivity on scientific productivity of Italian and French academic physicists, considering high impact-factor journals from WoS (Lissoni et al., 2011); co-authorship of the Slovenian scientific community (Kronegger et al., 2011) with data from the Co-operative On-Line Bibliographic System & Services (COBISS).

To the best of our knowledge, only few studies have been specifically devoted to the Statistics field. Baccini et al. (2009) 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 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. Lastly, the contribution of De Battisti and Salini (2010) investigates the publication style of Italian academic statisticians from several data sources (WoS, Scopus, CIS and Google Scholar) according to standard multivariate techniques. The authors recognize that the use of a single data source can led to biased and partial results.

In this study we aim to compare network results on collaboration of Italian academic statisticians using three bibliographic archives. In particular, we refer to co-authorship data collected by two international archives – one general (WoS) and one thematic (CIS) – and one national. WoS covers over 10,000 high-impact journals and over 110,000 conference proceedings in all disciplines and it consists of several databases for all sciences. For the analysis of a specific scientific discipline, however, one can consider the use of thematic databases. For statisticians CIS represents the principal available data source because it contains publications in Statistics and related fields. Since 1975, it covers over 160 core statistical journals, around 1200 additional journals with statistical oriented articles and 10,000 books in Statistics. Finally, if the interest is to take into account all kind of formal collaboration among scientists in a national community, other data sources can be explored. In Italy, bibliographic information is available from publications forms filled in individual scholars' web pages ("sito docente"), managed by the MIUR and the Cineca consortium. Unfortunately, access to this database is denied, due to the privacy policy. The only bibliographic information provided by the Cineca consortium regards selected publications of statisticians involved as national managers or members in PRIN projects from 2000 to 2008[5]. The three data sources differ in terms of coverage and information overlap, which

[4] For further details, see http://cercauniversita.cineca.it/php5/docenti/cerca.php.

[5] A network analysis was carried out for Italian physicists on data from this database (Bellotti, 2012).

Table 1
Italian academic statisticians by Statistics *subfields*, *academic ranking* and *university geographic location* (%). Source: MIUR 2010.

	All	Stat	Stat for E&T	Economic Stat	Demo	Social Stat
Academic ranking						
Researcher	38.0	38.6	46.7	41.9	31.8	29.7
Associate professor	28.3	27.8	33.3	25.0	28.2	36.5
Full professor	33.7	31.8	20.0	33.1	40.0	33.8
University geographic location						
North	39.1	44.0	13.3	34.4	32.9	37.8
Center	26.9	26.0	16.7	30.0	32.9	23.0
South	34.0	30.0	70.0	35.6	34.1	39.2
Total	792	443	30	160	85	74

may greatly affect the resulting co-authorship patterns between authors.

3.1. Data retrieval and cleansing

Publications by Italian statisticians were separately extracted from the three data sources. Data retrieval and data cleansing must be carefully carried out in view of the well-known disambiguation problem, which consists of dividing namesakes appearing in publication records into their real individuals (Kang et al., 2009). The main issues related to disambiguation are *homonymy* and *synonymity* (e.g., Calero et al., 2006). Homonymy occurs when different people have the same name, either due to coincidence or abbreviations of names (e.g., using initials for names instead of full names); whereas synonymity occurs when one person appears with different names. The main source of homonymy is incomplete author data in the publication records of bibliographic archives. This affects both the way of querying data sources (e.g., in WoS and CIS interfaces, only the initial of the first name and not the full name can be specified) and the attribution of a publication to the correct author once the record has been retrieved. The main source of synonymity is often the use of different names by authors who have more than one first name (there are 89 of them out of the 792 in our population) or surname (13 out of 792). Other sources are possible misspellings or nicknames.

Only two cases of homonymy are found among statisticians; whereas the number of cases of homonymy between statisticians and other academic researchers show a high probability (around 50%) of obtaining publications attributable to other authors. For this reason the queries in the retrieval step must be carefully composed. In addition, a data cleansing phase is necessary to eliminate possible errors.

3.1.1. Data retrieval

For the WoS and CIS international databases, data were retrieved through a Web-based interface, queried by filling in a Web form[6]. The interfaces of both sources are rather similar and allow users to compose queries as logical expressions by specifying one or more parameters, chosen through a combo box. Common information to both interfaces is: topic (keyword), author name, publication title, and journal title. CIS reports more information and enables further parameters to be specified, including time interval, publication type, file format. Instead, WoS includes a rich toolbox for data refinement in the result list interface (e.g., *subject categories*, in which each publication is classified).

Since CIS is a thematic data source, it presents a lower risk of ambiguity in the results. Hence, a simple query with author data (surname and initial of first name) for each of the 792 statisticians was used to retrieve data from this archive. However, WoS is a general and multidisciplinary data source, and queries giving only author data may produce many undesired results mainly due to homonomy cases (for details, see data cleansing, point 2). In summary, for WoS both author and affiliation (address) data – available in the MIUR register since 2000 – were used respecting the following rules:

- For the parameter *Author*, the value was obtained through the concatenation of the whole surname and the initial of the first name. For authors with one or more middle names, the "*" wildcard was attached to the initial of the first name. Surnames with an accent or apostrophe were listed verbatim in queries and multiple surnames were considered without abbreviations. It should be noted that this method of proceeding may have caused the loss of publication records in which multiple surnames are abbreviated (but, as noted above, this situation regards only 13 statisticians out of 792).
- For the parameter *Address*, the value was a logical expression, including terms referring to all the universities with which authors were affiliated during their career. A single affiliation may produce several terms, including part of the proper name of a university (if it exists), the name of its hosting city, and its English translation (when available).

For the Italian data source PRIN, the information provided by the Cineca consortium gives both selected publications by research project managers (maximum 5 publications until 2004, and maximum 30 publications since 2005) and selected scientific publications of the other research project members (maximum of 30 publications since 2007). Although these publications only represent a partial list, at the present time they are the only official data available from this national archive.

3.1.2. Data cleansing

Before obtaining co-authorship information, cleansing was carried out:

1. *Removal of duplicated publication records*. For all databases, duplicated records were due to the retrieval of the same publications through queries associated with the various co-authors of the document belonging to the target population. In addition, for PRIN, they are due to the presence of the same publication in different projects and years. As a result, a total number of 973, 175 and 1458 publications was deleted in WoS, CIS and PRIN databases, respectively;
2. *Removal of publication records erroneously attributed to authors*. Due to the homonymy problem, the data retrieved through queries may contain publications not authored by statisticians in our target population. This problem is especially apparent in data retrieved from WoS. Automated filtering was therefore carried out on these data: first, records presenting a mismatch on the full first author name reported in the WoS pages (associated to

[6] For further details, see websites of WoS http://apps.isiknowledge.com/ and CIS http://www.statindex.org/.

Table 2
Authors (All) and full professor (FP) coverage rates by Statistics *subfields* in the three data sources.

Subfields	**WoS**		**CIS**		**PRIN**		Never found	
	All	FP	All	FP	All	FP	All	FP
Stat	71.3	77.9	85.1	97.3	72.7	83.9	7.9	1.3
Stat for E&T	86.7	83.3	60.0	100.0	73.3	100.0	13.3	0.0
Economic Stat	42.5	34.0	65.0	90.6	59.4	71.7	20.0	3.8
Demo	40.0	47.1	48.2	67.6	67.1	85.3	27.1	8.8
Social Stat	50.0	46.4	55.4	72.6	81.1	96.0	12.2	4.0
Total	60.7	65.2	73.4	89.9	70.2	83.1	13.0	3.0

the publication title) were removed. Unfortunately this information is available only in a few cases, therefore the records which had *subject categories* not relevant to Statistics were then marked for further checks. In particular, the *marked* records were manually checked against other data sources (e.g., author's webpage if available, journal website). If the marked publication was not present in other sources, the record was deleted. As a result, a total number of 4948 publications was deleted in WoS publications in this phase;

3. *Correction of misspellings of authors' names*. Only for retrieved data, misspellings of authors' names could be corrected. Obviously, a publication containing a misspelling of the queried author's name is not shown in the query results and is completely lost. In order to treat misspellings, we performed a pairwise comparison of the names of all authors: those with an edit distance lower than three characters were manually inspected and, if possible, corrected.

After the data cleansing step, the highest number of publications was collected through the PRIN database (5608), followed by CIS (3518) and WoS (2289). We expected this result, due to the different kinds of publications collected in the three databases.

3.2. Author coverage rates and publication characteristics

A different coverage rate was obtained from the three data sources for all statisticians and the five subfields (Table 2). The lowest authors coverage rate is observed in WoS database (60.7%), with substantial subfield differences. Statistics for E&T research is quite well represented (86.7%), whereas only 40.0% of scientists is found in Demography. Statistics and Economic Statistics are well covered within CIS (85.1% and 65.0%, respectively), while authors in Demography and Social Statistics appear more frequently in PRIN (81.1% and 67.1%, respectively). In international databases, Demography, Economic Statistics and Social Statistics show low author coverage rates. This result may be the consequence of two combined aspects: partial inclusion of publications focusing on the specific research topics of these fields (e.g., the Econlit database would be more appropriate for Economic Statistics) and a higher tendency to produce publications at national level.

Considering *academic ranking*, the full professor coverage rate was lower in the WoS database (65.2%) with respect to the other sources (89.9% for CIS and 83.1% for PRIN). As before, Economic Statistics, Demography and Social Statistics show the lowest full professor coverage rates in WoS. The good coverage of full professors in CIS may be explained both by the inclusion in the past of the Conference Proceedings of the Italian Statistical Society and by the irregular updating, which does not include the publications of the youngest scientists. The total percentage of missing authors never found in the three databases was lower for Statistics (7.9%) with respect to Demography (27.1%) and Economic Statistics (20.0%).

The highest percentage of co-authored publications was found in WoS (about 85% on average) and the lowest value was in CIS (55.3%). PRIN reported an intermediate value equal to 71.2%. Statisticians belonging to the Statistics for E&T research showed the highest propensity to collaborate in all data sources (99.2% in WoS, 79.7% in CIS and 83.5% in PRIN), probably due to their attitude towards working with external co-authors involved in other disciplines (e.g., Medicine, Physics, Chemistry, etc.), in which the practice of collaboration is well established. The average number of authors per publication is around 3 for all statisticians in the CIS and PRIN databases (see Table 3). This value increases in WoS (12.6), due to the high number of authors per publication in Statistics for E&T research (49.2) and, to a small extent, in Social Statistics (7.2). From scientists' complete bibliographies (COBISS database), Kronegger et al. (2011) report comparable values to CIS and PRIN for Slovenian mathematicians (2.8) and sociologists (3.7), whereas physicists and biotechnologists show an higher value (both 4.6). Our findings are higher with respect to the results given by Newman (2004), referring to publications in Natural Sciences databases (MEDLINE, SPIRES, NCSTRL). Lastly, the average number of publications per author (Table 3) is around 6 in WoS and CIS, but much higher in PRIN (14.8 publications). The highest value was found in the Statistics for E&T research, and was observed in both WoS (about 15.7 publications) and PRIN (27.8 publications). Newman (2004) reports values of 11.6 for the SPIRES database and around 6 for MEDLINE, whereas the values in COBISS database (Kronegger et al., 2011) are higher: 52.5 (Physics), 29.9 (Sociology), 23.9 (Mathematics), and 21.4 (Biotechnology).

4. Co-authorship patterns in Statistics: research hypotheses

Starting from the co-authorship networks derived from the three data sources, we provide evidence on several research hypotheses on scientific collaboration patterns among Italian academic statisticians:

- H1: *The number of co-authored publications by Italian academic statisticians is growing faster than the number of single-authored publications, as observed in other scientific disciplines.*

  The probability of co-authoring differs across disciplines and over time but, in the last few decades, it has been increasing steadily across all fields (Moody, 2004, p. 217). We thus also expect to observe growth in scientific collaboration for Italian statisticians, as reported in the literature on other disciplines. This increasing co-authorship behaviour is supported by the three data sources.
- H2: *The collaboration style of the overall Italian statistician community – disregarding the five subfields– resembles the typical style observed in the literature for social sciences (in particular, according to the topological network structures found in Economics in international and national studies).*

  The small-world configuration appears as the most appropriate underlying mechanism to explain cooperative behaviour, mainly

Table 3
Main characteristics and network statistics for *Overall* and Statistics *subfields* in the three data sources.[a]

	Overall	Stat	Stat for E&T	Economic Stat	Demo	Social Stat
WoS [b]						
#. of authors	5291	2501	2152	337	187	791
#. of authors per pub (St. Dev.)	12.6 (61.5)	4.3 (12.5)	49.2 (136.4)	3.2 (2.1)	3.6 (2.3)	7.2 (5.1)
#. of pub per author (St. Dev.)	6.1 (8.8)	6.0 (5.9)	15.7 (27.7)	3.9 (3.9)	4.5 (5.8)	5.3 (6.7)
#. of statisticians	481	317	25	68	34	37
#. of isolated	26	15	0	2	5	4
#. of edges	427,238	81,500	400,829	863	597	5151
#. of internal edges	403	251	15	29	22	4
Density	0.031	0.026	0.173	0.015	0.034	0.016
Average degree	161.5	65.2	372.5	5.1	6.4	13.0
Largest distance	16	16	10	6	7	13
Average path length (ℓ)	5.47	6.70	3.08	2.23	3.07	4.85
Clustering coefficient (Γ)	0.91	0.94	0.91	0.76	0.58	0.59
# of components $\geq$1	77	54	6	41	10	20
Giant component (%)	91.7	80.6	93.9	14.2	76.5	64.6
E–I index	0.76	0.68	0.97	0.67	0.64	0.98
CIS						
#. of authors	1525	1188	100	276	106	126
#. of authors per pub (St. Dev.)	2.4 (0.7)	2.4 (0.9)	2.9 (1.3)	2.3 (0.8)	2.7 (1.1)	2.6 (1.0)
#. of pub per author (St. Dev.)	7.9 (8.6)	9.7 (9.6)	8.3 (8.4)	5.0 (5.1)	2.5 (2.1)	3.6 (3.5)
#. of statisticians	581	377	18	104	41	41
#. of isolated	60	28	0	19	5	8
#. of edges	2534	2012	227	332	136	153
#. of internal edges	631	387	12	63	19	9
Density	0.002	0.003	0.045	0.010	0.024	0.019
Average degree	3.3	3.4	4.5	2.4	2.6	2.4
Largest distance	19	19	5	11	4	8
Average path length (ℓ)	7.15	7.00	2.06	5.56	2.25	3.19
Clustering coefficient (Γ)	0.30	0.29	0.42	0.30	0.57	0.47
# of components $\geq$1	54	30	9	24	19	20
Giant component (%)	87.7	88.7	42.0	56.5	23.6	30.2
E–I index	0.03	0.19	0.63	0.24	0.37	0.68
PRIN						
#. of authors	2839	1669	469	401	292	603
#. of authors per pub (St. Dev.)	2.8 (1.6)	2.7 (1.5)	4.1 (3.3)	2.6 (1.1)	2.6 (1.2)	2.9 (1.6)
#. of pub per author (St. Dev.)	14.8 (12.3)	14.9 (12.2)	27.8 (17.0)	11.6 (9.5)	14.4 (13.5)	17.1 (12.0)
#. of statisticians	556	322	22	95	57	60
#. of isolated	7	4	0	1	0	1
#. of edges	9379	5071	2584	853	724	1686
#. of internal edges	999	458	21	88	96	65
Density	0.002	0.004	0.023	0.010	0.017	0.009
Average degree	6.6	6.1	11.0	4.2	4.9	6.0
Largest distance	17	16	8	12	8	9
Average path length (ℓ)	6.52	6.61	2.39	5.47	4.28	5.32
Clustering coefficient (Γ)	0.54	0.62	0.56	0.53	0.49	0.51
# of components $\geq$1	20	15	8	21	7	11
Giant component (%)	94.9	92.2	46.5	54.4	92.5	76.8
E–I index	0.24	0.33	0.86	0.43	0.29	0.67

[a] In each subfield, external authors include both authors outside Italian statistical community and authors affiliated to other Statistics subfields, except that under analysis.
[b] In this data source there are nine statisticians (two in Stat, five in Stat E&T and two in Social Stat) with very high degree (greater than 100).

due to the proximity of statisticians with other social scientists in the Italian academic context[7]. We also expect some evidence of different network structures related to data sources. A network pattern resembling a random configuration is expected in WoS, due to the main kinds of publications (high-impact journals) collected in this archive and the interdisciplinary openness of statisticians in collaborating with colleagues in other disciplines (e.g., Medicine, Physics, Chemistry, etc.). Journals publishing such interdisciplinary articles have a high probability of being included in international bibliographic archives. A clustered configuration very close to a small-world structure is expected in CIS and, to some extent, in PRIN. CIS is strongly oriented towards statistical journals, so it determines a selection of publications and co-authors only inside the Statistics discipline and its subfields. In PRIN, this network structure may be a direct consequence of the database definition, which focuses mainly on project managers' publications.

- H3: *The subfields of Statistics have different collaboration styles.*

We expect different mechanisms to characterise the subfields. Each subfield focuses on rather different research topics[8] that mainly refer to a more methodological or a more applied research interest in the development of statistical methods. These different focuses can lead to a lower or higher authors' propensity towards interdisciplinary collaboration (e.g., usually very high for Statistics for E&T research and Social Statistics). So we expect that network structure in these subfields could be consistent with a random network configuration. A further reason for subfield structural difference can be due to the presence of well-known scientists, which especially in the smallest groups (Statistics for E&T research, Social Statistics, and Demography) can act as "stars"

[7] Statistics does belong to scientific Area 13, called "Economics and Statistics", that is the institutional group defined by MIUR, comprising the following fields: Business, Economics, Mathematics for Economics, Finance and Insurance, and Statistics.

[8] Detailed descriptions are reported in official documents "declaratorie" published in Miur website http://hubmiur.pubblica.istruzione.it/web/universita/docenti-e-ricercatori.

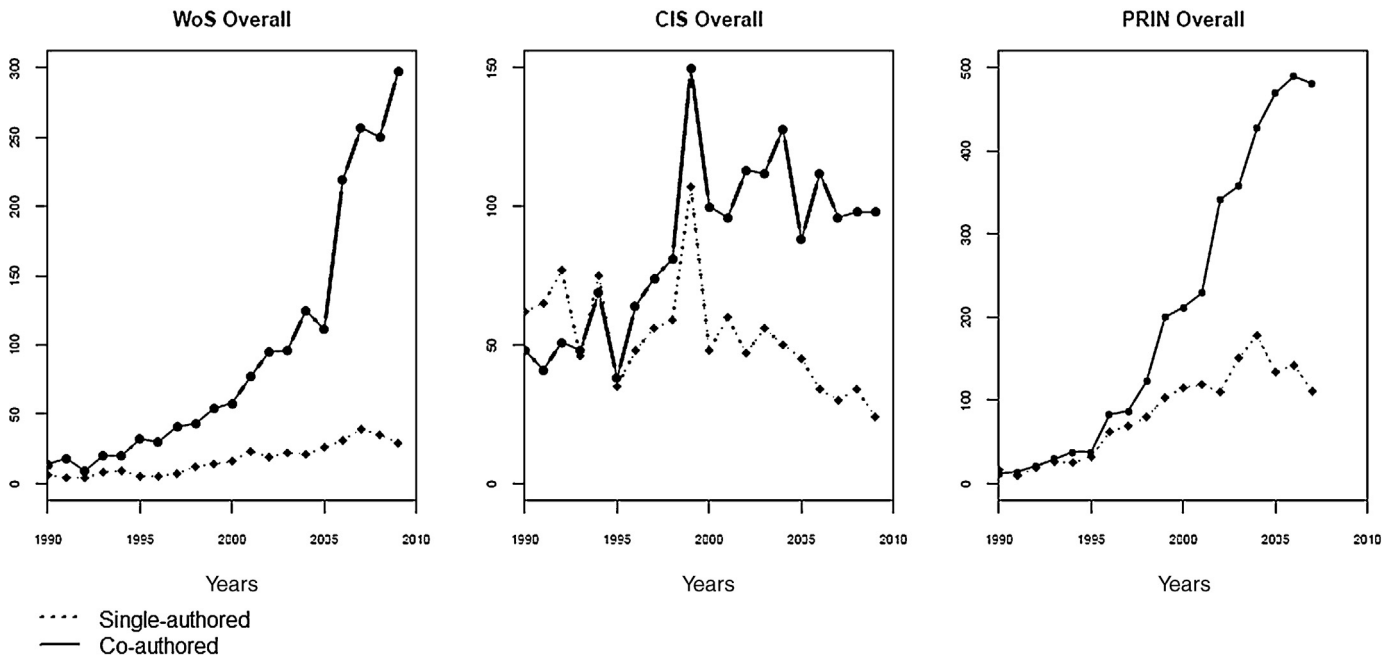

Fig. 1. Trend of co-authored (solid line) and single-authored (dashed line) publications for Overall statisticians in the three data sources. years 1990–2009.

authors. This behaviour can lead to a scale-free configuration within these groups. Instead, we expect that the two largest subfields, Statistics and Economic Statistics, resemble the collaboration style (small-world) we hypothesised for the overall Italian statisticians community, as stated in H2. These characteristics may emerge in different ways in the three data sources (H2).

- H4: *The scientific performance of Italian statisticians is related to authors' collaboration style in co-authorship networks.*

Several studies have shown that scientific productivity depends, among other things, on scientists' attitudes towards collaboration in research (e.g., Lee and Bozeman, 2005; Wuchty et al., 2007; Abbasi et al., 2011). By collaborating, scientists can benefit by both methodological and technological complementarities and synergy, improving the quality and quantity of their research output. Hence, once academic ranking and propensity to collaborate with subjects external to the statistical discipline is controlled for, positive associations between scientific performance and actors' central positions in the network are to be expected. Nonetheless, the strength of the association may differ in the three data sources, as stated in H2.

5. Analysis of co-authorship of Italian academic statisticians

In the following we present both collaboration trend and network analysis results for Italian academic statisticians related to our research hypotheses.

5.1. Scientific collaboration trend in the field of Statistics

In order to set up a common time-frame, mainly to make WoS data consistent with those in CIS and PRIN, we consider the number of co-authored and single-authored publications in the period 1990–2009[9].

We observe a significant increase in the proportion of co-authored publications in almost all the Statistics subfields since the end of 1990, as stated in H1 (results for overall statisticians are provided in Fig. 1). This finding confirms the tendency shown in the literature in the global increase of collaboration as from the early 1990s (Kronegger et al., 2011). Specifically, in our population, the mid-2000s for all subfields were crucial years for scientific collaboration within the WoS and PRIN databases. The proportion of co-authored publications began to grow very fast, and almost no slowdown can be observed. Instead, in CIS, there is a more variable trend in co-authorship in this period, mainly due to archive maintenance policies, often based on voluntary updates by country managers.

Several explanations may be given to the increasing of co-authorship over time. The growing scientific complexity and high degree of specialisation both appear to contribute to collaborative research (Babchuk et al., 1999) and then require interactions by scholars with different scientific skills. But funding requirements may also induce collaboration (Laband and Tollison, 2000) and the development of the Internet may facilitate it. For statisticians, this trend may partly be due to the central role played by Statistics in all sciences, in view of the importance of statistical methods in everyday applications. It seems that several problems in different disciplines may be addressed by Statistics. Cox (1997) reported the mid-1990s as a period of rapid development of the statistical sciences in many directions. This general tendency is shown in our data with a 10-year delay.

5.2. Co-authorship networks: main characteristics

Taking into account both overall statisticians and the five Statistics subfields, 18 adjacency data matrices are defined from the affiliation matrices retrieved from the three data sources. We consider a binary version of these matrices, setting all entries in the original valued matrices greater than zero to 1. Our choice is based

[9] Papers published before 1989 in WoS have not been considered during the query process due to license restriction in 2010 at our universities. The percentage of dropped publications was 0.1% for WoS, 21.1% CIS and 8.1% for PRIN before the 1990 and 3.8% for WoS, 0.3% for CIS and 0% for PRIN after the 2009, respectively.

on two main reasons: (i) we aim to make a comparison with results reported in seminal papers on co-authorship (in particular, Moody, 2004; Newman, 2004; Goyal et al., 2006), adopting the same dichotomisation; (ii) we observe that link values greater than 1 are quite rare in our networks (the percentage of links greater than 1 on total links among statisticians and their co-authors is around 43% in the worst case, observed in PRIN database), therefore disregarding link values, setting a threshold equal to 1, does not produce substantial information loss.

Density is very low for all networks (Table 3), and the average degree is particularly high for WoS (161.5) with respect to the other two databases (3.3 for CIS and 6.6 for PRIN), due to some publications with more than 100 authors[10]. The average degree computed without these outliers becomes 10.6 in the overall network, 9.2 in Stat, 13.3 in Stat for E&T, and 12.7 in Social Stat. In any case, these values are higher than those reported for Italian economists (on average about 2 in the period 1986–2006, Maggioni and Uberti, 2011) and are in line with values found for scientists in the Slovenian study in the period 1991–2005, apart from Mathematics which has a degree around 2 (Kronegger et al., 2012).

The three overall networks (see Table 3) show an important largest component (which accounts for 94.9% in PRIN, 91.7% in Wos, and 87.7% in CIS); some isolated scholars (i.e., authors having only single-authored publications), especially for CIS; and a large number of small components with a minimum number of 2 authors, especially in WoS.

The extent of collaboration closure of all statisticians and the five subfields was evaluated through the E–I index (Krackhardt and Stern, 1988) based on the comparison of the number of internal links among statisticians and external links between statisticians and outsider authors. The group level E–I index[11] in some cases shows very high positive values (near to 1), indicating that scientists start up collaboration mainly with external authors. As expected, some differences may be noted by data sources and subfields. In particular, WoS reports higher scores for overall statisticians and the five subfields; whereas CIS generally shows lower values. Statistics for E&T research and Social Statistics have higher scores in the three data sources, especially WoS and PRIN. In general, the E–I index values indicate high interdisciplinarity behaviour, with different levels by subfields and data sources. These results show some evidence in favour of our hypotheses H2 and H3.

5.3. Assessment of structural hypotheses in networks topology

In this section we test the consistency of the observed networks with topological structures emerging in co-authorship settings (i.e., small-world and scale-free networks), described in Section 2.

5.3.1. Assessment of small-world properties

Small-worldliness is characterised by small dense network regions – revealed by high clustering coefficient $\Gamma(G)$ – and by short paths connecting any two actors – revealed by low average path length $\ell(G)$, typically bounded by $O(\log n)$. Specifically, it is required that $Q_\ell^{\mathcal{R}} = \ell(G)/\ell(\mathcal{R}) \approx 1$ and $Q_\Gamma^{\mathcal{R}} = \Gamma(G)/\Gamma(\mathcal{R}) \gg 1$, where $\Gamma(\mathcal{R})$ and $\ell(\mathcal{R})$ are the values of clustering coefficients and average path lenght averaged over K graphs, generated from a random model $\mathcal{R}$.

Typically, assessment is made by assuming $\mathcal{R} = ER$ as baseline model. However, since ER models are limited in the types of degree distributions they may account for, we also carried out the assessment by simulating random graphs from the more general configuration model (CM) which allows for more complex degree sequences, which would be extremely rare under the ER model assumption.

We simulate $K = 1000$ random graphs from both the ER model, fixing $\pi = \Delta(G)$, that is the observed network density, and the CM model, fixing the degree sequence $(d_1, \ldots, d_n)$ on the observed ones.

As expected, in the three data sources and for both overall and subfield networks, the ratio between observed $\Gamma(G)$ and theoretical ones[12] computed from both random models – $\Gamma(ER)$ and $\Gamma(CM)$ – is always very large (see Table 4). It should be noted that these values are higher than the values computed in the above-mentioned studies, especially in PRIN, probably due to the inherent clusterisation induced by project participation. This finding indicates that observing such $\Gamma(G)$ values by chance alone (according to different random network models) is very unlikely. Each of the observed co-authorship networks, irrespective of both scientific subfield and data source, are characterised by a significantly large number of small subgroups, which can potentially determine the emergence of a small-world structure. However, the other required property – i.e., the shortness of $\ell(G)$ compared with $\ell(ER)$ and $\ell(CM)$ – is not met, either for all data sources or for each network in the three data sources.

The small-world structure clearly characterises collaboration only within the CIS database for all networks although with border values for average path length, according to the CM comparison for overall network ($Q_\ell^{CM} \approx 1.382$) and Statistics subfields ($Q_\ell^{CM} \approx 1.393$). Evidence of small-worldliness also arises, although to a lesser extent, in a few networks in the other data sources. Co-authorship networks of Economic Statistics and Demography in WoS may be regarded as small-world structures, as well as Statistics for E&T research in PRIN and with border values in Demography in the CM model ($Q_\ell^{CM} = 1.212$).

5.3.2. Assessment of scale-free networks

In order to evaluate whether the observed co-authorship networks may be viewed as structures forming from a preferential attachment process, a power law distribution is fitted to the observed degree distributions by the maximum likelihood estimation (Nicholls, 1986). A power law distribution including only scale parameter α as well as an alternative formulation with the additional parameter for the lower-bound on scaling region x_{min} (as proposed by Clauset et al., 2009), are considered (see Section 2).

The Kolmogorov–Smirnov (KS) test (Table 4) shows that the hypothesis of the presence of a scale-free configuration must be rejected for all analysed networks at 1% significance level. When the fit is made with lower bound distribution, we obtain the same results for the overall networks and for most of the subfield networks in the three data sources, with the exception of CIS. In particular, the degree distributions of 4 out of 5 subfields from CIS and of the Demography subfield from PRIN are clearly described by a power law from a given lower bound.

The absence of a power law degree distribution in the three complete overall networks implies that this scientific community is not affected by prominent researcher effects.

[10] The value of average degree is affected by the presence of a few authors in Statistics for E&T research and to a lesser extent in Statistics and in Social Statistics, in which co-authored publications on natural science topics show a large number of co-authors.

[11] The E–I index may be applied at three levels: whole network, group level, and individual level. The whole network E–I index was not considered, because its computation is affected by the presence of external authors who present homophily behaviour by network construction, due to the lack of co-authorship data for them, apart from the links they have with Italian statisticians. The E–I index at individual level, which accounts for the embeddedness of each scientist in the group, is considered as covariate in the model specified in Section 5.4.

[12] Both observed and simulated results are reported considering the whole network. Results based only on the giant component show slight variations, given its size in all networks (see Table 3).

Table 4
Small-world and scale-free topology assessment for *Overall* and by Statistics *subfields* in the three data sources.[a]

	Overall	Stat	Stat for E&T	Economic Stat	Demo	Social Stat
WoS						
Small world						
$\ell(G)/\ell(ER)$	2.769	3.113	1.684	0.597	1.018	1.697
$\Gamma(G)/\Gamma(ER)$	29.663	36.002	5.253	49.699	17.037	35.718
$\ell(G)/\ell(CM)$	2.135	2.535	1.497	0.619	1.028	1.709
$\Gamma(G)/\Gamma(CM)$	2.510	2.359	2.176	21.337	6.783	13.150
Scale free						
Power law						
C	0.240	0.281	–	0.419	0.383	0.296
$\hat{\alpha}$	1.281	1.339	–	1.565	1.450	1.360
Power law l-b						
$\hat{x}_{min}$	3	3	348	2	5	13
$\hat{\alpha}$	1.500	1.520	2.850	1.900	2.380	3.120
CIS						
Small world						
$\ell(G)/\ell(ER)$	1.166	1.198	0.650	0.923	0.480	0.626
$\Gamma(G)/\Gamma(ER)$	138.195	98.901	9.189	33.765	24.533	24.965
$\ell(G)/\ell(CM)$	1.382	1.393	0.656	1.046	0.485	0.660
$\Gamma(G)/\Gamma(CM)$	45.903	35.142	5.011	19.403	19.749	17.128
Scale free						
Power law						
C	0.494	0.494	0.333	0.558	0.531	0.546
$\hat{\alpha}$	1.716	1.715	1.417	1.866	1.799	1.836
Power law l-b						
$\hat{x}_{min}$	3	3	4	3	3	3
$\hat{\alpha}$	2.610	2.630	2.810***	3.140***	3.500***	3.280***
PRIN						
Small world						
$\ell(G)/\ell(ER)$	1.473	1.531	0.850	1.280	1.153	1.361
$\Gamma(G)/\Gamma(ER)$	231.842	170.212	23.911	49.878	28.409	55.343
$\ell(G)/\ell(CM)$	1.632	1.676	0.846	1.363	1.212	1.442
$\Gamma(G)/\Gamma(CM)$	50.188	43.431	6.888	20.983	8.743	18.612
Scale free						
Power law						
C	0.391	0.402	0.266	0.450	0.440	0.407
$\hat{\alpha}$	1.515	1.534	1.316	1.625	1.605	1.544
Power law l-b						
$\hat{x}_{min}$	11	6	7	2	2	17
$\hat{\alpha}$	3.100	2.480	2.340	2.100	2.020***	3.330

[a] Significant parameter at: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Our results indicate that the emergence of small-worldliness and scale-free topologies depends on data sources as well as on Statistics subfields. Our H2 hypothesis is therefore confirmed for CIS and WoS overall networks and the effect of data sources on subfields stated in H3 is completely confirmed. Subfields in CIS also reveal a topology with interconnected stars, consistent with small-world and scale-free behaviour, as reported for economists in Goyal et al. (2006) and Slovenian mathematicians and sociologists in Kronegger et al. (2012).

The absence of authors acting as stars in the overall community of Italian academic statisticians does not mean that prominent statisticians are not important within their respective subfields and also within the whole structure. In particular, as suggested by Goyal et al. (2006), the arrangement of links in the networks must be explored for deeper insights of processes responsible for network aggregate features. Actor-level network statistics (degree, closeness, betweenness, and the clustering coefficient) for the most highly linked statisticians[13] and their individual characteristics (subfield, affiliation, number of publications, and h-index) are listed in Table 5 for the overall community in the three data sources.

Besides their role as connectors, on average the five most prominent statisticians also show very high closeness and betweenness centrality with respect to the whole population in the three data sources. The average degree of the top 100 statisticians also has a high value with respect to the average degree computed for all authors. Instead, their clustering coefficient is smaller than the overall average.

In order to examine the role of "star" authors in network connectivity and clustering, we compared the effects of randomly removing 5% (see Goyal et al., 2006) of all authors and only statisticians, with the effect of deleting the same percentage of star actors and star statisticians from the overall networks. The random removal of 5% of authors has negligible effects on both network connectivity and clustering for the three data sources. When 5% of star actors are removed, there is a great reduction in the largest components in both CIS (42.6% of authors) and PRIN (63.5% of nodes). Likewise random removal, also in this case the effect on the clustering coefficient is negligible. When the removal concerns the 5% of star actors among statisticians, we note a remarkable reduction of the largest component in CIS (20.0% of authors) and PRIN (41.9% of authors) and an increase in the clustering coefficient. Similarly to findings for Economics (Goyal et al., 2006), also for Statistics in Italy, the most highly linked authors act like interconnected stars, and their removal greatly increases the distance between different groups of statisticians. Again, the strength of their role differs in the data sources (as stated by our research hypothesis H2).

5.4. Network position and scientific performance

In the following, examining the three overall co-authorship networks, we analyse the relationship between scientific performance, measured by the h-index and central positions in the

[13] The ranking is first obtained by the degree value and then by closeness, betweenness, and the clustering coefficient.

Table 5
Network statistics[a] for most highly linked statisticians in the three data sources.

Rank	Subfield	Univ. geograph. location	# Publ.	Deg.	Clos. (rank)	Bet. (rank)	Γ (rank)	*h*-index(rank)
WoS								
1	Stat E&T	South	35	890	0.213 (19)	0.001 (191)	0.841 (79)	21 (2)
2	Stat	North	23	446	0.235 (9)	0.061 (6)	0.721 (92)	14 (5)
3[b]	Stat E&T	Center	136	392	0.281 (1)	0.341 (1)	0.037 (363)	28 (1)
4	Stat E&T	South	8	358	0.225 (14)	0.017 (25)	0.945 (113)	3 (162)
5	Stat E&T	North	57	190	0.204 (22)	0.07 (3)	0.060 (15)	20 (3)
Avg. top 100	–	–	14.01	51.750	0.17	0.02	0.34	5.82
Avg. all	–	–	6.12	14.26	0.11	0.00	0.46	3.14
CIS								
1[b]	Stat E&T	Center	35	38	0.168 (10)	0.066 (7)	0.130 (266)	28 (1)
2	Stat	South	36	34	0.148 (53)	0.061 (10)	0.055 (317)	3 (145)
3	Stat	North	41	33	0.123 (211)	0.018 (50)	0.110 (276)	7 (32)
4	Stat	North	51	29	0.141 (92)	0.084 (4)	0.090 (301)	3 (154)
5	Stat	North	37	28	0.172 (7)	0.087 (3)	0.040 (324)	9 (14)
Avg. top 100	–	–	19.42	12.41	0.14	0.02	0.16	5.16
Avg. all	–	–	8.97	5.13	0.10	0.01	0.33	3.01
PRIN								
1	Stat E&T	Center	28	122	0.195 (13)	0.039 (21)	0.152 (297)	7 (30)
2[b]	Stat E&T	Center	69	116	0.220 (1)	0.166 (1)	0.787 (375)	28 (1)
3	Stat E&T	South	31	69	0.167 (397)	0.063 (6)	0.115 (326)	6 (51)
4	Social Stat	North	25	64	0.169 (352)	0.046 (13)	0.101 (278)	8 (20)
5	Stat	North	17	54	0.198 (10)	0.026 (41)	0.340 (137)	8 (24)
Avg. top 100	–	–	30.05	25.90	0.17	0.02	0.22	4.56
Avg. all	–	–	16.11	10.47	0.15	0.01	0.36	3.10

[a] Network statistics are computed only on statisticians, disregarding outsider authors.
[b] Same author ranked in the three data sources.

co-authorship networks – measured by degree (d_i), closeness (c_i), betweenness (b_i) – and the local clustering coefficient (Γ_i). We also include the individual E–I index (EI_i) to account for the propensity to collaborate inside or outside the field of Statistics and a dummy variable for the academic ranking "Full Professor" (FP_i) as a proxy for academic seniority as well as anagraphic age.

It should be noted that using the *h*-index as a measure of scientific performance has some limitations. As reported by Costas (2007, p. 194) these drawbacks are mainly related to: (i) the different productivity and citation practices of fields; (ii) the duration of each scientist's career; (iii) the artificial increase in the number of self-citations. Nevertheless, this index combines a measure of quantity (publications) and impact (citations) in order to characterise the scientific productivity of a researcher performing better than other single indicator. We mainly consider *h*-index thanks to its availability for all authors in our target population, as retrieved from Scopus.

It is interesting to note that the correlation of clustering coefficient is always negative with respect to both *h*-index and centrality measures, which indicates that in general collaboration within closed groups has a negative influence on scientific performance and actors' network position; whereas the centrality measures have a positive relation with the *h*-index. The correlation between actor relational variables is not very high, except for degree and betweenness in CIS and PRIN (in both data sources $r = 0.76$).

For evidence regarding the influence of actor relational covariates on scientific performance, a generalised extreme value distribution (GEV) is fitted[14]. The choice of GEV is due to the particular nature of the *h*-index distribution, which is generally highly skewed and heavy tailed[15]. GEV is a family of distributions combining the Gumbel, Fréchet and Weibull families also known as type I, II and III extreme value distributions (Coles, 2001), having a cumulative distribution function of the following form:

$$F(z;\mu,\sigma,\xi) = exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \quad (3)$$

where $\mu \in \Re$ is the location parameter, $\sigma > 0$ the scale parameter, and $\xi \in \Re$ the shape parameter. Shape parameter ξ governs the tail of the distribution: the higher ξ, the heavier the tail. In particular, in Eq. (3), a value of $\xi > 0$ corresponds to the Fréchet type, which has a heavy-tailed distribution, power law-like; $\xi < 0$ corresponds to the Weibull distribution, which is light-tailed; $\xi \to 0$ corresponds to the Gumbel type, which is a distribution with an exponential tail.

In detail, we model *h*-index distribution (h) as:

$$h \sim GEV(\mu_i, \sigma_i, \xi_i) \quad (4)$$

where

$$\mu_i = const_1 + \alpha_1 d_i + \alpha_2 c_i + \alpha_3 b_i + \alpha_4 \Gamma_i + \alpha_5 EI_i + \alpha_6 FP_i \quad (5)$$

$$\sigma_i = \sigma \quad (6)$$

$$\xi_i = const_2 + \beta_1 d_i + \beta_2 c_i + \beta_3 b_i + \beta_4 \Gamma_i + \beta_5 EI_i + \beta_6 FP_i \quad (7)$$

In order to obtain the simplest model which explains as much of the variation in the data as possible, we first include all actor covariates in both location and shape parameters in Eq. (4) (model 1). Then, by means of the likelihood ratio test, we compare model 1 with a simpler model including covariates only in the location parameter (model 2). After selecting one of these two models, we estimate the final one, omitting non-significant terms. Results for the three data sources are shown in Table 6.

The GEV estimates show that *h*-index distribution is always heavy-tailed in the three data sources (positive significant values of $const_2$ parameter). The results suggest positive associations between scientific performance and actor's network position, as stated in H4. In particular, for the three data sources we have:

- the model for WoS is the simplest one (model 2), having covariates only on μ. In particular, the *h*-index is positively affected by centrality measures, especially betweenness ($\widehat{\alpha}_3 = 1.25$). In

[14] When dealing with network data, unit independence cannot be assumed (e.g., Doreian et al., 1984). The extension of the GEV model to include autocorrelation parameters requires technically complicated estimation methods not considered here.
[15] We assume *h*-index to be a continuous variable (see Beirlant and Einmahl, 2010).

Table 6
Maximum likelihood estimates of GEV parameters[a]. Standard errors in brackets.

Parameters	WoS	CIS	PRIN
$const_1$	2.14 (0.07)***	1.99 (0.09)***	1.95 (0.08)***
α_1 – Degree (d_i)	0.31 (0.09)***	0.43 (0.07)***	0.61 (0.07)***
α_2 – Clos. (c_i)	0.25 (0.07)***	0.15 (0.08)**	–
α_3 – Bet. (b_i)	1.25 (0.07)***	–	–
α_4 – Γ (Γ_i)	−0.14 (0.06)***	–	−0.14 (0.06)***
α_5 – E–I index (EI_i)	0.18 (0.07)***	0.16 (0.07)***	–
α_6 – Full professor (FP_i)	–	−0.31 (0.14)***	–
σ	1.38 (0.05)***	1.39 (0.06)***	1.54 (0.07)***
$const_2$	0.06 (0.03)**	0.08 (0.05)*	0.16 (0.04)***
β_1 – Degree (d_i)	–	–	0.05 (0.03)**
β_2 – Clos. (c_i)	–	0.07 (0.04)**	–
β_3 – Bet. (b_i)	–	–	–
β_4 – Γ (Γ_i)	–	–	0.07 (0.04)**
β_5 – E–I index (EI_i)	–	–	−0.06 (0.04)*
β_6 – Full professor (FP_i)	–	0.15 (0.09)**	–

[a] Significant parameter at: *$p < .1$, **$p < .05$, ***$p < .01$.

addition, the value of the individual EI_i index is positively related to the location parameter. The only covariate having a negative effect on performance is the clustering coefficient ($\widehat{\alpha}_4 = -0.14$);

- the model for *h*-index distribution in CIS is slightly more complex, with significant covariate effects on both μ and ξ. Only degree ($\widehat{\alpha}_1 = 0.43$), closeness ($\widehat{\alpha}_2 = 0.15$) and EI_i index ($\widehat{\alpha}_5 = 0.16$) have a significant coefficient. The dummy variable coefficient shows a significant negative value for μ, which means that the *h*-index distribution for full professors has a lower median value ($\widehat{\alpha}_6 = -0.31$) whereas the distribution tail is heavier ($\widehat{\beta}_6 = 0.15$) with respect to statisticians not in full professor position. Closeness is the only network centrality measure having a (positive) significant effect ($\widehat{\beta}_2 = 0.07$) on shape parameter ξ;
- the *h*-index distribution observed in PRIN is also described by a model with covariate effects on both parameters (model 1). Among the actor centrality measures, only the degree has a positive effect on μ ($\widehat{\alpha}_1 = 0.61$). As in the WoS model, the clustering coefficient has a significant negative impact on the *h*-index median value ($\widehat{\alpha}_4 = -0.14$). Considering the tail of the *h*-index distribution, the higher number of co-authors belonging to closed groups implies a slightly greater probability of observing very large *h*-indexes ($\widehat{\beta}_1 = 0.05$ for degree, $\widehat{\beta}_4 = 0.07$ for clustering coefficient), whereas the EI_i index is negatively related with ξ. It is worth clarifying the meaning of the opposite sign of coefficients for the same covariate – clustering coefficient – in the μ_i and ξ_i equations. In particular, for large values of clustering coefficients, the probability of observing extreme values of *h*-index increases, especially for values corresponding to the 90th percentile of the estimated GEV. In fact, when the 90th percentile of the estimated GEV is plotted as a function of the standardised clustering coefficient, the relation becomes positive. The change of the clustering coefficient effects on different quantiles (median and 90th percentile) of *h*-index is shown in Fig. 2.

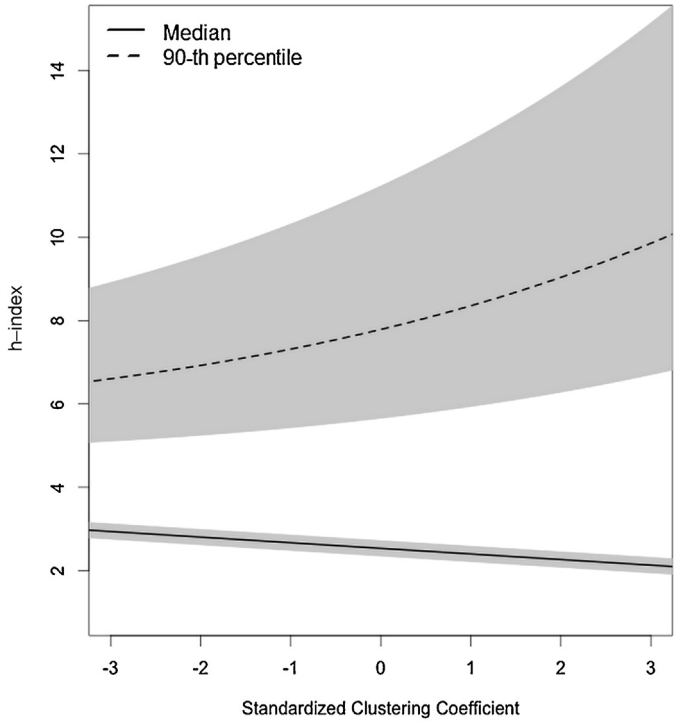

Fig. 2. Median (solid line) and the 90th percentile (dashed line) of the estimated GEV distribution for PRIN database. Both quantiles are represented as function of the standardised clustering coefficient $\Gamma(G)$. Shaded areas represent the 95% confidence interval obtained by the delta method.

6. Discussion and concluding remarks

This study focuses on the co-authorship patterns of the community of Italian academic statisticians as they emerge from three data sources which contain different kinds of scientific publications. A different coverage rate was obtained from the three data sources for all statisticians, and in particular for some subfields. As a general finding, in international databases, Demography, Economic Statistics and Social Statistics have low author coverage rates.

The whole bulk of results on Italian statisticians' co-authorship networks provides strong evidence in favour of our research hypotheses H1–H4.

A general tendency towards increasing co-authorship was observed in Italian statisticians, with a delay of approximately ten years, compared with results at international level. The collaboration style of Italian academic statisticians presents features partly observed in both social and natural sciences. The small-world structure, emerged in Economics at international (Goyal et al., 2006) and national (Maggioni and Uberti, 2011) level, clearly characterises collaboration only within the CIS database for all networks and, to a lesser extent, also in a few networks in the other two data sources. In addition, only within subfields in the CIS network, topologies also appear to be consistent with scale-free behaviour, as reported for other disciplines, in Economics (Goyal et al., 2006) and in Mathematics and Sociology (Kronegger et al., 2012). General evidence of a positive association between performance and actors' central position in the network seems to be confirmed. Occupying a central position in the network – recognised by a brokerage position in WoS and by a high degree in PRIN and CIS – positively affects scientific performance. The attitude towards working with colleagues in closed groups, showing a negative effect on performance in WoS and PRIN, still has the opposite effect in PRIN, favouring the probability of having a very large *h*-index.

Network results and their effects on scientific performance appear to be strongly influenced by the features of data sources. On-line bibliographic archives, usually selective on included publications, are not neutral on retrieved results, and the choice of one as opposed to another must be carefully examined according to the aims of the analysis. International databases allow to trace collaborative behaviour of members in a specific target population who usually published in high-impact journals; whereas local research archives (for instance, the Slovenian COBISS database) can be more complete containing both top-international as well as nationally oriented production. In the case of CIS, for example, it represents the principal available on-line international data source for statisticians because it contains publications strongly oriented towards

statistical journals, made by statisticians with co-authors belonging, mainly, to the same field. Then, the more defined patterns (especially small-world configuration, although allowing the presence of some statisticians acting as star authors) we found out in CIS can be reasonably attributed to its specific features, as happens in Econlit database for economists. In summary, collaboration style in CIS database resembles a small-world configuration, with statisticians clustered into distinct groups and connected by few shortcuts. This kind of network structure allows statistical knowledge to flow easily among actors. Compared to the other two data sources, CIS can capture internationalisation openness by research topics and publication style of Italian statisticians rather than their tendency towards an interdisciplinary behavior, the latter being better represented in WoS database. Finally, PRIN mixes up some of CIS and WoS characteristics, although referred only to selected publications (that are limited in number and whose topics are constrained to the project's content).

To conclude, we provide some directions for future work. Co-authorship data retrieval in a target population suffers from several data quality issues, requiring in our case substantial manual checking, usually not possible with large populations. For this purpose, author detection by statistical matching techniques, employing network information as well as actors' attributes, should be considered.

The evidence – for some subfields and data sources – of small-worldliness with relevant star actors roles suggest to move beyond these well established topological structures toward other configurations not yet fully explored in co-authorship (e.g., core-periphery structure), as well as to apply recent methods of community detection (Fortunato, 2010) in order to provide insights on the presence of specific groups acting in the whole network. Co-authorship network analysis could also be improved by enhancing positional analysis through blockmodelling. In order to provide insights on both the determinants of scientific collaboration and network evolution, network statistical modelling (e.g., ERGM) could be considered. Lastly, a deeper investigation of the relationship between scientific performance and network positions is required both as regards suitable indicators to measure performance and statistical modelling to better account for data dependence.

Acknowledgements

The authors would like to thanks Francesco Pauli (University of Trieste) for his useful suggestions in GEV model estimation, the MIUR for PRIN data source availability, the editor and the anonymous reviewers for their helpful comments.

References

Abbasi, A., Altmann, J., Hossain, L., 2011. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. Journal of Informetrics 5, 594–607.
Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Review of Modern Physics 74, 47–97.
Babchuk, N., Keith, B., Peters, G., 1999. Collaboration in sociology and other scientific disciplines: a comparative trend analysis of scholarship in the social, physical and mathematical sciences. The American Sociologist 30, 5–21.
Baccini, A., Barabesi, L., Marcheselli, M., 2009. How are statistical journals linked? A network analysis. Chance 22, 35–45.
Beirlant, J., Einmahl, J.H.J., 2010. Asymptotics for the Hirsch Index. Scandinavian Journal of Statistics 37, 355–364.
Bellotti, E., 2012. Getting funded. Multi-level network of physicists in Italy. Social Networks 34, 215–229.
Bender, E.A., Canfield, E.R., 1978. The asymptotic number of labelled graphs with given degree sequence. Journal of Combinatorial Theory A 24, 296–307.
Burt, R.S., 1978/1979. Stratification and prestige among elite experts in methodological and mathematical sociology circa 1975. Social Networks 1, 105–158.
Calero, C., Buter, R., Valdes, C.C., Noyons, E., 2006. How to identify research groups using publication analysis: an example in the field of nanotechnology. Scientometrics 66, 365–376.
Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data. SIAM Review 51, 661–703.
Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London.
Costas, R., Bordons, M., 2007. The *h*-index: advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics 1, 193–203.
Cox, D.R., 1997. The current position of statistics: a personal view. International Statistical Review 65, 261–276.
De Battisti, F., Salini, S., 2010. Quale profilo per gli statistici italiani? Sis-Magazine, Available at: http://www.sis-statistica.it/magazine/spip.php?article186
Doreian, P., Teuter, K., Wang, C.-H., 1984. Network autocorrelation models: some Monte Carlo results. Sociological Methods and Research 13, 155–200.
Fortunato, S., 2010. Community detection in graphs. Physics Reports 486, 75–174.
Garfield, E., 1979. Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. ISI Press, Philadelphia.
Glanzel, W., Schubert, A., 2004. Analyzing scientific networks through co-authorship. In: Moed, H., Glanzel, W., Schmoch, U. (Eds.), Handbook of Quantitative Science and Technology Research. Kluwer Academic Publishers, Dordrecht, pp. 257–276.
Goyal, S., Van der Leij, M.J., Moraga-Gonzalez, J.L., 2006. Economics: an emerging small world. Journal of Political Economy 114, 403–412.
Hicks, D., 1999. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44, 193–215.
Hummon, N.P., Doreian, P., 1989. Connectivity in a citation network: the development of DNA theory. Social Networks 11, 39–63.
Hummon, N.P., Carley, K., 1993. Social networks as normal science. Social Networks 15, 71–106.
Kagan, J., 2009. The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century. Cambridge University Press, Cambridge.
Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., Sung, W.-K., Lee, J.-H., 2009. On co-authorship for author disambiguation. Information Processing and Management 45, 84–97.
Katz, J.S., Martin, B.R., 1997. What is research collaboration? Research Policy 26, 1–18.
Krackhardt, D., Stern, R.N., 1988. Informal networks and organizational crises: an experimental simulation. Social Psychology Quarterly 51, 123–140.
Kronegger, L., Ferligoj, A., Doreian, P., 2011. On the dynamics of national scientific systems. Quality & Quantity 45, 989–1015.
Kronegger, L., Mali, F., Ferligoj, A., Doreian, P., 2012. Collaboration structures in Slovenian scientific communities. Scientometrics 90, 631–647.
Laband, D.N., Tollison, R.D., 2000. Intellectual collaboration. The Journal of Political Economy 108, 632–662.
Lazega, E., Jourda, M.T., Mounier, L., Stofer, R., 2008. Catching up with big fish in the big pond? Multi-level network analysis through linked design. Social Networks 30, 157–176.
Lee, S., Bozeman, B., 2005. The impact of research collaboration on scientific productivity. Social Studies of Science 35, 673–702.
Leti, G., 2000. The birth of statistics and the origins of the new natural science. Metron 58, 185–211.
Liberman, S., Wolf, K.B., 1997. The flow of knowledge: scientific contacts in formal meetings. Social Networks 19, 271–283.
Liberman, S., Wolf, K.B., 1998. Bonding number in scientific disciplines. Social Networks 20, 239–246.
Lievrouw, L.A., Rogers, E.M., Lowe, C.U., Nadel, E., 1987. Triangulation as a research strategy for identifying invisible colleges among biomedical scientists. Social Networks 9, 217–248.
Lissoni, F., Mairesse, J., Montobbio, F., Pezzoni, M., 2011. Scientific productivity and academic promotion: a study on French and Italian Physicists. Industrial and Corporate Change 20, 253–294.
Maggioni, M.A., Uberti, T.E., 2011. Networks and geography in the economics of knowledge flows. Quality & Quantity 45, 1031–1051.
Melin, G., Persson, O., 1996. Studying research collaboration using co-authorships. Scientometrics 36, 363–377.
Moody, J., 2004. The structure of a social science: disciplinary cohesion from 1963 to 1999. American Sociological Review 69, 213–238.
Newman, M.E.J., 2004. Coauthorship networks and patters of scientific collaboration. Proceedings of National Academy of Sciences of United States of America 101, 5200–5205.
Nicholls, P.T., 1986. Empirical validation of Lotka's law. Information Processing and Management 22, 417–419.
Watts, D., Strogatz, S., 1998. Collective dynamics of small world networks. Nature 393, 440–442.
Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316, 1036–1039.