

The scientific collaboration network of Italian academic statisticians

Domenico De Stefano, Maria Prosperina Vitale and Susanna Zaccarin

Abstract The aim of this paper is to describe the collaboration network among Italian academic statisticians and to compare the resulting network characteristics with those of other disciplines in social and natural sciences. Usually, data on collaboration are derived looking at the co-authorship among scientists, mainly because of the increasing availability of archival database on publications. Several data sources can be considered but each of them presents pros and cons related to both the coverage of the target population and the inclusion of different kind of publications. In this paper we exploit the potential of integrated information to study the patterns of interactions in the given scientific community.

Key words: Scientific Collaboration, Archival data, Co-authorship network, Social networks analysis

1 Introduction

An increased attention on scientific collaboration can be noted in the last years. Collaboration is considered to be a key element in the advancement of knowledge because scientists embedded in collaboration networks share ideas, use similar techniques and influence each other's work. Several studies have also shown that scien-

Domenico De Stefano
Dept. of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste,
e-mail: domenico.destefano@econ.units.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@econ.units.it

tific productivity depends, among other things, on scientists attitude toward collaboration in research (e.g. [6]). By the means of collaboration, scientists can benefit of methodological and technological complementarities (different expertises) as well synergy improving quality and quantity of their research output. As an empirical evidence collaboration among scientists is increasing in all sciences.

It is straightforward to think about collaboration among scientists as a social networks where the actors ties could be represented by the co-production of a given scientific output or the co-participation to a common research project. Longterm trends in scientific work might depend on the broader pattern of disciplinary networks. Empirical evidence of disciplinary fractionalization and specialty areas, clustered into distinct groups of scientists, can support 'small world' pattern [14]; while a small number of high connected scientists (stars scientists) can be better described by a scale-free network with preferential attachment [1]. On the other hand, broad connectivity among a large portion of scientists (theoretical integration) implies structural cohesion whose behavior can be simulated (at least of the large component) by a random network [9].

A substantial body of works in the analysis of collaboration patterns can be found in various fields at international level: from Physics and Biomedical research ([1], [10]) to research in Economics [4] and Sociology [9]. The common aims of these studies are the understanding of the properties of networks and the implications that collaboration patterns have on the evolution over time of topics and methods, as well as the evaluation of scientific productivity and quality among scientists involved.

Traditionally, social relational data are collected through field studies, mainly by direct interviews or by filling in questionnaires according to some suitable name generator procedure. An alternative source can be provided from archival documents. Since collaboration is a mix of *informal* (advice, face to face contacts, personal exchange) and *formal* activities (writing a paper, participating in a project) among scientists, direct interviews can be very useful to gain insights on informal collaboration, while archive data can provide good information on several kinds of formal collaboration. Although the information retrieved in archival data has not been collected for network studies, these data sources are relatively inexpensive, do not impose a burden on informant time and efforts and may be less prone to missing data and inaccuracy problems.

In the field under study, several types of archive data are available: 1) electronic bibliographies that index general and thematic literature which includes journal articles, books, book reviews, collective volume articles, working papers and dissertations (see ISI Web of Science - WoS, Scopus or Current Index to Statistics - CIS, Econlit and domestic research archives tenured by universities and research institutions); 2) thematic references reviews; 3) electronic journal catalogs (Caspur, Jstor); 4) books of abstracts distributed during formal meetings; 5) research projects (see the Italian database of national project PRIN provided by Minister of University and Research- MIUR); 6) applications for patents (see the European EPO-INV database). Although sharing some common characteristics, the different sources provide specific aspects and partial coverage of the collaboration among scientists. The main problems in the use of international databases regard which kind of sci-

entific production is included and its implications on the coverage of the target population. As a matter of fact, writing articles or books, publishing in international or national journals, can be dependent by disciplines or specialty within disciplines. An additional issue, common to these data sources, is the misreporting in author identification.

This paper focuses on scientific collaboration among academic statisticians in Italy - specifically scientists classified as belonging to SECS-S/01, -S/02,-S/03,-S/04 and -S/05 sectors in the MIUR database - with the aim to explore the style of collaboration in this national scientific community by means of Social Network Analysis (SNA) techniques. The purpose of the study is twofold. First, we are interested in investigating if the emerging pattern of diffusion of statistical knowledge among Italian scholars resembles the typical one observed in the literature in natural or in social sciences. Second, we want to discuss coverage aspects connected to available data sources. To this aim, international general (Scopus or WoS) and thematic (CIS) databases will be considered along with collaboration information retrieved by MIUR database of granted national projects (PRIN).

The remainder of the paper is organized as follows: in Section 2 the current framework linking network structure to the knowledge diffusion is presented. Main empirical results from different disciplines are also reported. In Section 3 ins and outs related to both the coverage of the target population and the inclusion of different kind of publications in some databases are discussed. Some preliminary results are presented in section 4.

2 Pattern of collaboration in scientific communities

Scientific collaboration, and specifically co-authorship, are becoming more common irrespective of discipline and several studies have shown how cooperation manifests itself in many different ways across scientific domains as well as over space and time (e.g. [9]).

However, remarkable differences there exist between natural sciences and social sciences and between natural sciences and humanities. The main reason of such diversity can be principally due to the use of quantitative methods [9] and to the disciplines internal organization. Researchers in natural sciences are more likely to work in large lab settings [10] giving rise to team-based research output, whereas in social sciences this does not happen very often.

These aspects determine the emergence of peculiar structures governing the collaboration behavior within scientific fields. A frequent finding in collaboration networks, for instance measured through co-authorship, is that they assume the aspect of some theoretical structure with well-defined topological and relational properties. A number of recent papers highlight how SNA provides useful insights to discover the prominent structure of the collaboration network in different disciplines by detecting peculiar statistical properties that explain network ties configuration (see for instance [10], [9]). Network models start from the idea that the connections between

actors occur at random. Random networks are represented by the *Erdos-Renyi random graph* where the probability of having a tie between actors' pairs is equal to p . Usually, random structures are not observed in real phenomena but they represent the baseline model to assess non-random behavior in observed networks. Empirical evidence shows that co-authorship networks are usually non-random because they tend to exhibit distinctive statistical properties deriving from peculiar ties generating mechanisms.

One of the most diffused topology emerging within several scientific communities/disciplines is the so-called “small world” network configuration [14]. The relevant feature in networks having this property is that, despite the growing of its size and heterogeneity in actors' characteristics, their connectedness is very high and the distance among regions of the structure is shorter than the one observed in other kind of networks. In particular, average actors pairs distance (i.e. *average path length*) in small worlds is shorter than the distance observed in random networks. The shorter distance reveals actors tendency to be optimal connected with respect to random behavior. In any case, random networks with $p \rightarrow 1$ are also well-connected although small world networks are connected in an “optimal” sense because the links among actors are more parsimonious and functional to the average path length reduction (see Fig.1).

Another characteristic that distinguishes small world networks is the high tendency toward actors clustering. This tendency is measured by the *clustering coefficient* (CC), a network level index ([10], [14]). The CC value observed in a small world network is much larger than the one measured among nodes in a random network.

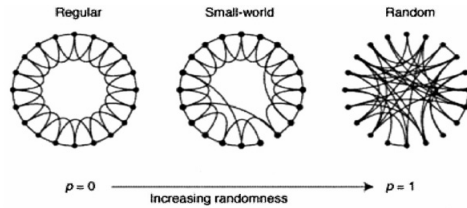


Fig. 1 Small world network compared with a regular lattice (on the left) and with a random graph with wiring probability equal to p . Source: Watts, Strogatz, 1998.

Small world is not the only well-known topology that can be found in collaboration networks. Another theoretical configuration, often observed in empirical studies, is the so-called “scale-free” topology [1]. The scale-free networks are related to the distributional properties of a specific network feature, namely the actors degree distribution. If it follows a “power law” distribution, with parameter α ranging in a predetermined interval, a scale-free structure is observed. This implies the existence of a peculiar tie formation mechanism named “preferential attachment”, which formally accounts for the tendency to interact with the best connected actors (i.e. actors with highest degree) (see Fig.2). Hence, there exist very few actors holding the role of stars and acting as attractors of collaborators, students and resources. Stars can also influence funding politics. Scientific collaboration in presence of one

(or more) star, tends to have a more centralized structure that is rather independent on the discipline.

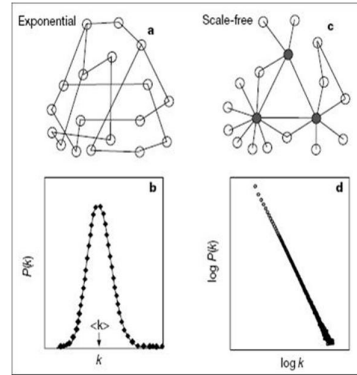


Fig. 2 Exponential degree distribution (a) which is typical of random networks (b) compared to scale free network (c) with Power Law degree distribution (d). Source: Albert, Barabási, 2002.

Clear evidence of small world properties have been observed in Economics [4] and Physics [10]. Physics, Maths and Neurosciences [1], and Economics also present statistical properties consistent with preferential attachment mechanism (although not all having strictly power-law distribution). Sociology is the one exception because it is better represented by an integrated (cohesive) collaboration network structure [9].

At the best of our knowledge, there are few works that focus on the analysis of collaboration structures within statistical communities. The Statistics field, indeed, presents some characteristics common to natural sciences (e.g. quantitative methods) and social sciences (usually scientific outputs are not realized in lab settings although they are produced from “on-field work”). Recent results on statisticians affiliated to the prominent US universities from CIS database suggest consistence with scale-free configuration [12].

3 Data sources for scientific collaboration

In the analysis of scientific collaboration a crucial facet is the collection of data to obtain information according to the kind of relationship among scientists we are interested in. In the following we discuss the use of some different databases in terms of coverage and overlap information to study the formal relationship expressed by co-authorship among Italian academic statisticians.

WoS and Scopus represent the most used general and multidisciplinary data sources that contain abstracts and citation databases of peer-reviewed literature (journals, conference proceedings, book series). Each of these general archives has a different collection policy which affects database coverage both in number of journals as well as in publications and references (see for instance [13]).

WoS covers over 10000 high-impact journals and over 110000 conference proceedings in all disciplines; whereas Scopus covers around 27000 titles (18.167 active). With regards to scientific areas, titles indexed in Scopus present a partial coverage for social sciences (20%) and it seems to be more oriented towards natural sciences (80% for Physical Sciences, Health Sciences and Life Sciences together). Instead, WoS presents seven databases for all sciences. One of them is completely dedicated to social sciences and it contains around 2500 journals across 50 specific disciplines. Scopus website reports information about overlapping titles covered by both Scopus and WoS (about 60% of them). Whereas, about 40% are unique titles contained only in Scopus. Due to this large amount of overlapping, they could be considered as an interchangeable data sources.

For the analysis of a specific scientific discipline, however, one can consider the use of thematic databases. In particular, for the statisticians CIS represents the principal available data source because it contains publications in statistics and related fields. Indeed, it covers over 160 core statistical journals, around 1200 additional journals with statistical oriented articles and 10000 books in statistics published since 1975.

In this framework, there exist several works that focus on Italian scientific community. In particular, some domestic studies describe scientific collaboration patterns by using general and thematic archives. See for example: 1) Italian scholars working on population studies through reference data [11]; Italian academic economists using Econlit database [8], academic Italian and French physicists considering WoS journals [7].

Collaboration patterns derived from international bibliographic databases are constraint by the included publications. Instead, 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. To this regard, domestic research archives could be more complete than bibliographic international databases because they allow to consider all research outputs (books, articles in local journals, technical reports, chapters in books). Indeed, there are some studies that describe collaboration structure through domestic archives containing all scientific publications of active national researchers (see for instance the Slovenian database [3], the Turkish database [5] and the local archive of the University of Salerno [2]).

For Italian scientific community, co-authorship information could be retrieved from the list of research outputs available in the scholars MIUR web page (“sito docente”), managed by Cineca consortium, but the access to this database is restricted due to privacy policy.

4 The scientific community of Italian academic statisticians: some preliminary results

For our study, we consider the 792 Italian academic statisticians, belonging to SECS-S/01 (443), -S/02 (30), -S/03 (160), -S/04 (85) and -S/05 (74) reported in

the MIUR database at the end of March 2010. We derive collaboration from publications retrieved in Scopus and CIS and from bibliographic information, provided by MIUR-Cineca, regarding the publications selected by statisticians involved in granted research projects (PRIN) from 2000 to 2008. The collaboration pattern among statisticians is defined by taking into account the co-authorship relationships, i.e. authors are linked according to the papers they co-authored.

The dataset furnished by MIUR presents a total of 7058 documents (likely with duplicated records) collected from both selected publications of research project coordinators (maximum 5 publications until 2004 and maximum 30 publications since 2005) and selected scientific publications of the other research project members (a maximum of 30 publications since 2007). Even though these publications represent a partial list, at this moment they are the only official data available from MIUR.

In the following some preliminary results for the 30 researchers and professors belong to the SECS-S/02 are presented. For what concerns the coverage of this small group, we found that 22 out of 30 members are included in MIUR database, whereas 26 out of 30 are in Scopus. We first computed two affiliation matrices (authors $\times$ papers) $\mathbf{F}_S$ and $\mathbf{F}_M$ for Scopus and MIUR data, of size respectively (1832×415) and (428×506) , where the authors size represent both the statisticians under analysis and their co-authors. Hence, we take into account complete networks but without collecting the publications of such co-authors and their ties with scientists outside our population.

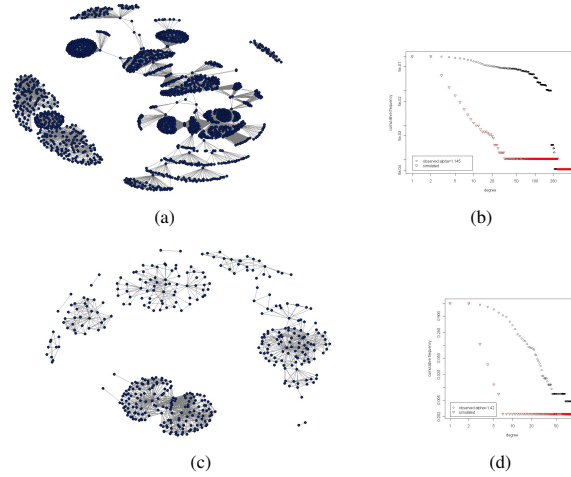


Fig. 3 Co-authorship networks for Italian academic SECS-S/02 statisticians derived by publications contained in Scopus (a) and MIUR (c) databases; Degree distributions for the adjacency matrices $\mathbf{A}_S$ (b) and $\mathbf{A}_M$ (d): red = random degree distribution; black = observed degree distribution.

In order to analyze the patterns of relationships among authors, two undirected binary adjacency matrices $\mathbf{A}_S$ and $\mathbf{A}_M$ are derived from $\mathbf{F}_S$ and $\mathbf{F}_M$. Fig.3 shows the

simple graphs for the two networks embedded in A_S and A_M . Small world configuration is consistent with characteristics observed in the network derived by MIUR data, but not for Scopus where the observed average path length is higher than the corresponding value for random graph. The observed degree distributions of A_S and A_M are fitted to a power law distribution in order to evaluate the presence of scale-free structures. The Kolmogorov-Smirnov goodness of fit statistic shows that the hypothesis of scale-free network have to be rejected for both networks at 1% significance level.

The collaboration network structures, obtained from the two databases for this small group, emphasize how data sources may influence the net configurations describing the collaboration in a scientific community.

References

1. Albert, R., Barabási, A.-L.: Statistical Mechanics of Complex Networks. *Rev. of Mod. Phys.* **74**, 47-97 (2002)
2. De Stefano, D., Vitale, M.P.: Exploring the pattern of scientific collaboration networks. In: Ingrassia, S., Rocci, R. (eds.) *Book of Short Papers Classification and Data Analysis*, pp. 161-164. CLEUP, Padova (2009)
3. Ferligoj, A., Kronegger, L.: Clustering of Attribute and/or Relational Data. *Metodološki zvezki* **6**, 135-153 (2009)
4. Goyal, S., Van der Leij, M.J., Moraga-Gonzalez, J.L.: Economics: An Emerging Small World. *Journal of Political Economy* **114**, 403-412 (2006).
5. Gossart, C., Ozman, M.: Co-authorship networks in social sciences: The case of Turkey. *Scientometrics* **78**, 323-345 (2009)
6. Lee, S., Bozeman, B.: The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science* **35**, 673-702 (2005)
7. Lissoni, F., Mairesse, J., Montobbio, F., Pezzoni, M.: Scientific Productivity and Academic Promotion: A Study on French and Italian Physicists. WP 27, KITEs Knowledge, Internationalization and Technology Studies University Commerciale Luigi Bocconi (2009). Available via DIALOG. http://portale.unibocconi.it/wps/wcm/connect/Centro_KITES/Home
8. Maggioni, M.A., Uberti, T.E.: The economic geography of scientific collaborations. Workshop ARS09 Social Network Analysis: Models and Methods for relational data, University of Salerno, 13-14 July (2009) Available via DIALOG. <http://www.ars2009.unisa.it/>
9. Moody, J.: The Structure of a Social Science: Disciplinary Cohesion from 1963 to 1999. *American Sociological Review* **69**, 213-238 (2004)
10. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci.* (2004)
11. Rivellini, G., Rizzi, E., Zaccarin, S.: The science network in Italian population research: An analysis according to the social network perspective. *Scientometrics* **67**, 407-418 (2006)
12. Said, Y.H., Wegman E.J., Sharabati, W.K.: Author-Coauthor Social Networks and Emerging Scientific Subfield. In: Palumbo, F., Lauro, C., Greenacre, M. (eds.) *Data Analysis and Classification. Series: Studies in Classification, Data Analysis and Knowledge Organization*, pp. 257-268. Springer, Berlin-Heidelberg (2010)
13. Vieira, E.S., Gomesa, J.A.N.F.: Comparison of Scopus and Web of Science for a typical university. *Scientometrics* **81**, 587-600 (2009)
14. Watts, D., Strogatz, S.: Collective dynamics of small world networks. *Nature* **393**, 440-442 (1998)