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).