

The analysis of collaboration networks: methodological and applied issues

Susanna Zaccarin

Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Trieste
susanna.zaccarin@econ.units.it

Abstract: Collaboration networks in which the actors are scientists and the ties between them are scientific collaborations, as documented in the papers they write or in the common research projects and teamwork they are involved in, provide an example of a genuine network of human acquaintances for which a precise definition of acquaintance is given. Derived from archival data, these networks can be of very large size achieving the required degree of complexity to be considered in the realm of “complex systems”. In this work specificity and statistical properties shared by this kind of large socially-generated network will be reviewed.

Keywords: co-authorship, electronic database, complex networks, network properties, network modelling

1. Networks and scientific interaction

A social network can be viewed as a relational system characterized by a set of actors (or groups) – possibly with their attribute variables – and the connections each of them has with the others (Wasserman and Faust, 1994, p. 89). A network can be represented as a set of points (also called vertices or nodes) to denote actors, joined in pairs by lines (also called edges or ties).

Social networks have been extensively studied in the social sciences for at least the last fifty years because of inherent interest in the patterns of human interaction, but also because the implications of the structures of network have on the spread of information and knowledge in general. 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 (Knoke and Yang, 2008).

An alternative source can be provided from archival documents. Although the information has not been collected for network studies, compared with surveys, archival data are relatively inexpensive, do not impose a burden on informant time and efforts and may be less prone to inaccuracy problems and small sample size. Archival data are of various forms, including Internet Web pages, personal letters, patent citations, patent inventors, joint research venture projects among firms, references from books and articles, co-author relationships among academics.

This work focuses on collaboration networks in which the actors are scientists and the ties between them are scientific collaborations, as documented in the papers they write or in the common research projects and teamwork they are involved. Collaborations networks provide an example of a genuine network of human acquaintances for which a precise definition of acquaintance is given. These networks can be of very large size achieving the required degree of complexity to be considered in the realm of "complex systems", where complexity may have to do with the intricate dynamics of the

interacting components with the non-trivial properties of the underlying network topology, or with the size of the system itself.

Issues on the specificity of collaboration networks and the main available archival data sources are discussed in the next section. The final section 3 reviews statistical properties that have been found to be common to large socially-generated networks and the main approaches proposed to account for these characteristics in formal models for network data.

2. Collaboration networks

Collaboration networks attract a lot of attention from the scientific community for several reasons. Compared with other kinds of relational data derived from archival documents, they can be better interpreted as an example of socially generated networks with the meaning usually provided in the social sciences studies. It is probably fair to think that most pairs of people who have written a paper together or have shared a common teamwork are genuinely acquainted with one another (Newman, 2004) or, at least, have exchanged crucial scientific or technical information. On the contrary, for example, in networks of citations, the actors are papers and the directed ties between them are citations of one paper by another. However, citation networks are not true social networks since the authors of two papers need not be acquainted for one of them to cite the other's work.

The idea of analyzing co-authorship is not new. Mathematicians are well familiar with the concept of the Erdos number, that is the shortest distance (geodesic distance in the jargon of social network analysis) between a scientist and the Hungarian mathematician Paul Erdos in the co-authorship network from that field (Batagely and Mrvar, 2000). A substantial body of work in the analysis of collaboration patterns can be found in various fields: from physics and biomedical research (Barabasi, 2002; Newman, 2004) to research in economics (Goyal *et al.*, 2006) and sociology (Moody, 2004). For an application on population studies network in Italy see also Rivellini *et al.* (2006). The common aims of these studies are the understanding of the properties of networks and the implications 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.

Further, recent papers in economics of science and innovation stress the interactive and collective properties of innovation. Interactions among actors within scientific and technological networks of cooperation have been growing stronger over the past two or three decades – particularly in the fields of high technology. The linkages among these actors may take the form of joint research, personnel exchange, cross-patenting, and a variety of other channels. Many recent studies highlight that knowledge production involves an increasing number of actors who are more interconnected to each other (see, among others, Balconi *et al.*, 2004; Roijakkers and Hagedoorn, 2006).

Data on collaboration are usually reconstructed from the affiliation network that links scientists (actors) to the papers (patents or projects) they co-authored together. An affiliation network is a network in which actors are joined together by common membership to groups of some kind (*e.g.*, papers or patents). Affiliation networks can be shown as a graph consisting of two kinds of vertices, one representing the actors and the other the groups (*e.g.*, papers). In order to analyse the patterns of relations among

actors, however, affiliation networks are often represented simply as unipartite (or one-mode) graphs of actors joined by undirected edges (Wasserman and Faust, 2004). Very large networks can be assembled in this way as well, since in many cases group membership can be ascertained from membership lists. Several types of data source can be considered: 1) electronic bibliographies that index general and thematic literature which includes journal articles, books, book reviews, collective volume articles, working papers and dissertations (see for instance, ISI Web of Science, Google Scholar, Scopus or Econlit and Medline); 2) thematic references reviews; 3) electronic journal catalogs; 4) books of abstracts distributed during formal meetings; 5) applications for patents (see for instance, the European EPO-INV database). Although sharing some common characteristics, the different sources provide specific aspects and specific coverage of the collaboration among authors.

3. Statistical properties and models for socially-generated networks

From the accumulation of empirical results on real-world networks highly distinctive statistical signatures have been observed to be common to socially-generated networks of a wide variety of types. The most important observation to come out of studies such as these is that networks are generally very far from random (Newman, 2003). Main properties such networks tend to exhibit can be summarize as follows:

1. small diameter and small average path length. That is known as the “small world” effect. “Small” is on the order of the log of the number of nodes or less and the diameter is the maximum distance between any two nodes of the network, where distance between nodes is defined as the shortest path between them measured in the number of links or edges;
2. high clustering coefficients relative to what would emerge if the links were simply determined by an independent random process. Clustering refers to the extent to which the existence of ties between actors A and B and between actors B and C implies a tie between A and C;
3. degree distributions (the degree of a node is the number of links that involve that node as one of the endpoints) present “fat tails” with many more nodes with very small and very large degrees than one would see if the links were formed completely independently and in some cases approximate a “scale-free” or “power-law” distribution, at least in the upper tail.

In Table 1 basic statistics on collaboration networks taken from the literature are reported. While these characteristics are far from enough to completely characterize a network, together they give a great deal of information about network structure.

An important issue in the analysis of these structures concerns the formulation of models able to incorporate observed properties. Many researchers have proposed models of networks that typically seek to explain either how networks come to have the observed structure, or what the expected effects of that structure will be.

Different formulations range from random graph model by Erdos and Renyi and its subsequent generalization, Markov graphs, the small-world model to models of network growth, particularly the preferential attachment models by Barabasi and Albert (1999). A complete review is discussed in Newman (2003).

Table 1: Basic statistics for a number of coauthorship networks in the late 1990s

	Nodes (actors)	Average number of collaborators (degree)	Average degree of separation ^a (distance)	Clustering coefficient
Maths	253,339	3.92	7.57	0.34
Physics	52,912	9.27	6.19	0.56
Biology	1,520,251	15.53	4.92	0.60
Economics	81,217	1.67	9.47	0.16
Sociology ^b	87,731	2.70	11.53	0.27

Source: Newman (2004) for Maths, Physics and Biology; Goyal *et al.* (2006) for Economics and Moody (2004) Sociology.

a. Applies only within the largest connected component (giant component).

b. Only scientists with collaborators.

References

- Balconi M., Breschi S., Lissoni F. (2004) Networks of inventors and the role of academia: an exploration of Italian patent data, *Research Policy*, 33, 127-145.
- Barabasi A.-L., Albert R. (1999) Emergence of scaling in random networks, *Science*, 286, 509-512.
- Barabasi A.L. Jeong H., Neda Z., Ravasz E., Schubert A., Vicsek T. (2002) Evolution of the social network of scientific collaborations, *Physica A*, 311, 590-614.
- Batagelj V., Mrvar A. (2000) Some analyses of Erdos collaboration graph, *Social Networks*, 22, 173-186.
- Goyal, S., van der Leij M.J., Moraga-Gonzalez J.L. (2006) Economics: An Emerging Small World, *Journal of Political Economy*, 114 (2), 403-411.
- Knoke D., Yang S. (2008) *Social Network Analysis*, 2nd ed., Series: Quantitative Applications in the Social Sciences, n. 154, Sage.
- Moody J. (2004) The Structure of a Social Science Collaboration Network: Disciplinary cohesion from 1963 to 1999, *American Sociological Review*, 69, 213-238.
- Newman M.E.J. (2003) The structure and function of complex network, *SIAM Review*, 45, 167-256.
- Newman M.E.J. (2004) Who is the best connected scientist? A study of scientific coauthorship networks, in: *Complex Networks*, Ben-Naim E., Frauenfelder H., Toroczkai Z. (Eds.), Springer, 337-370.
- Rivellini G., Rizzi E., Zaccarin S. (2006) Science network in Italian population research: an analysis according to the social network perspective, *Scientometrics*, 67 (3), 407-418.
- Robins G., Snijders T.A.B., Wang P., Handcock M., Pattison P. (2007) Recent developments in Exponential random Graph (p^*) Models for Social Networks, *Social Networks*, 29 (2), 192-215.
- Roijakkers W., Hagedoorn J. (2006) Inter-firms R&D partnering in pharmaceutical biotechnology since 1975: Trend, patterns, and networks, *Research Policy*, 35, 431-446.
- Wasserman S., Faust K. (1994) *Social Network Analysis. Methods and applications*, Cambridge, Cambridge University Press.