Community structure in co-authorship networks: the case of Italian statisticians

Domenico De Stefano, Maria Prosperina Vitale, Susanna Zaccarin

Abstract Community detection is a very appealing topic in network analysis. A precise definition of community is still lacking, so the comparison of different methods is not a simple task. The paper shows exploratory results by adopting two well-known community detection methods and a new proposal to discover groups of scientists in the co-authorship network of Italian academic statisticians.

Key words: co-authorship networks, community detection algorithms, modularity, Italian statisticians

1 Introduction

In the last decades social network analysis (SNA) has become a wide spread methodological approach to study scientific collaboration. As stated in several studies [6,8], scientific collaboration is a crucial factor to enhance publication productivity and research quality. The role of scientific collaboration allowing a fertile ground for the development of new ideas is also recognized in research funding European programmes as well as national projects.

Thanks to the availability of international bibliographic archives, co-authorship networks –in which the connection between two researchers are given by the number

Domenico De Stefano
Department of Social and Political Sciences, University of Trieste
e-mail: ddestefano@units.it

Maria Prosperina Vitale
Department of Political and Social Studies, University of Salerno
e-mail: mvitale@unisa.it

Susanna Zaccarin
Department of Business, Economic, Mathematics and Statistics, University of Trieste
e-mail: susanna.zaccarin@deams.units.it

of papers they co-authored– are used as a proxy of scholars' collaborative behavior in science [2]. Usually, binary networks –setting the connections greater than zero to one– are considered in empirical analysis. A common aim in co-authorship studies through SNA perspective is the understanding of network properties since the evolution of topics and methods in scientific fields appears strongly related to the topological structure of the collaboration patterns among scholars. In this stream of research, the recovery of *communities* –the term used to identify groups or clusters of actors in a graph– shaping the network structure sounds very appealing and informative. Unfortunately, a precise definition of what constitutes a community –broadly, part of a network where internal links are denser than external ones– is still lacking [16]. As a consequence of this conceptual vagueness, several community detection algorithms have been proposed in the literature [9].

Starting from previous findings on small-world topology in the co-authorship network of Italian academic statisticians [5, 11], the present contribution intends to deepen the analysis of this case study uncovering a meaningful community structure for Italian scholars. To this aim, results from three community detection methods, the Girvan-Newman algorithm [13], the Louvain algorithm [3] and a new method –Modal clustering algorithm [12]– will be compared. The evaluation of performance measures [16] and the interpretation of main results should benefit of the common clustering perspective shared by the three algorithms.

The paper is organized as follows. Section 2 reviews the main characteristics of the three methods and their performance in identifying communities within an illustrative example. Section 3 discusses the main results obtained by adopting the aforementioned methods on the co-authorship network of Italian statisticians using also available scholar's attributes (i.e., scientific field and university affiliation). Section 4 reports new lines of research for future work.

2 Community detection methods

Similarly to the problem of clustering for attribute data, the lack of a unique definition of community in presence of network data has lead to the proliferation of several methods in different theoretical contexts. Among them, some are explicitly designed to handle these kind of data. For instance Blockmodeling [7, pp. 11-12] is a methodological approach "*to identify, in a given network, clusters of actors that share structural characteristics in terms of some relations*", mainly based on partitioning the relational matrix by the clusters into a set of blocks.

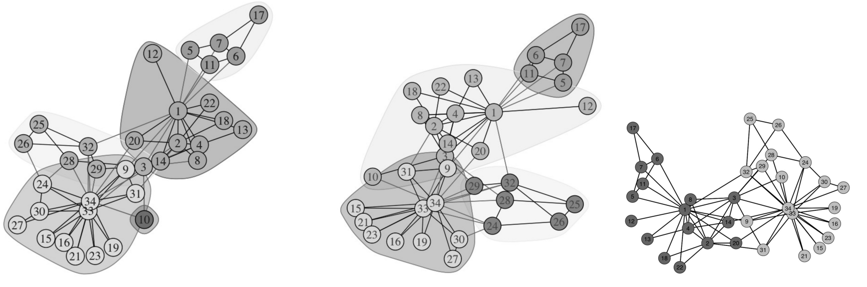
Recently, a huge variety of network-based clustering techniques, the so-called community detection methods, have been developed based on hierarchical clustering techniques [13], locating network communities by statistical analysis of the raw data [14] or optimizing different quality functions [9]. These general methods have been also used in the literature for analyzing co-authorship networks.

In the following, we focus on two well-known community detection algorithms, and a new proposed method based on an adaptation to network data of modal clustering procedure (for an overview with standard data, see [1]):

1. the Girvan-Newman algorithm [13], one of the most popular community detection approach. It is based on a hierarchical divisive procedure in which links are iteratively removed based on the value of the edge's betweenness. The procedure of link removal ends when the value of the modularity index Q is maximized. This index [4, 13] measures the fraction of the edges in the network that connect nodes within-community minus its expected value in the case of a network with edges placed at random. It assumes a minimum value of 0, when the number of within-community edges is no better than the randomized network, and a maximum value of 1 in presence of strong community structure. The index usually falls in the range 0.3 to 0.7, and a value of around 0.3 is a good indicator of significant community structure in the network;
2. the Louvain algorithm [3], also based on the modularity index and on a hierarchical approach. Initially, each node is assigned to a community on its own. In every step, nodes are re-assigned to communities in a local, greedy way: each node is moved to the community in which it achieves the highest contribution to the modularity;
3. the Modal clustering algorithm [12], which starts from the idea that highly connected sets of nodes can be detected around the modes of a "density" function f reflecting the cohesiveness between nodes –e.g. centrality measures [10] like the node degree (i.e., the number of links a node has with the other nodes in the network) or the actor betweenness (i.e., the number of those shortest paths passing through a specific node connecting two other nodes). The modes of f are seen as the archetypes of the clusters, which are in turn represented by their surrounding regions. Any section of f, at a level λ, identifies a level set, namely the region with f value above λ. The key idea is that when f is unimodal, there is no clustering structure, and the level set is connected for any choice of λ. Conversely, when f is multimodal, the identified level set may be connected or not, depending on λ value. In particular, nodes are clustered together when they have a value of f above the examined threshold λ and they are connected in the underlying network. Clustering is performed around the modal actors, namely actors showing the largest value of the chosen function. Furthermore, by varying the level set the method gives rise to a tree diagram, called cluster tree (which is graphically similar to a dendrogram), where each leaf corresponds to a mode of the function.

The first two algorithms are particularly suited for undirected and unweighted relational data (likewise the most usual case of co-authorship data obtained disregarding the number of papers co-authored by pairs of scholars), while the third one is more flexible since different concepts of cohesiveness among actors can be used.

To compare the three approaches in discovering communities, we consider the Zachary's karate club network data [17] describing the friendship relationship among 34 members of a karate club at an US university in the Seventies. A useful feature of this dataset is that, during the period of observation, the club split into two factions, due to a dispute between the administrator and the karate instructor. Thus, a true cluster membership of the actors in the network is known and can be used as a benchmark to evaluate the performance of different methods. Figure 1 shows the communities identified by using the three algorithms. It is possible to appreciate that the Modal clustering method, using node degree as density function to reflect actors' cohesiveness, allows to detect the two factions underlying the networks. In particular, the method works by clustering every actors around the modal actors –that are the two most central ones in terms of their degree in Figure 1c– that, incidentally, are the members around which the Karate club splits into two distinct factions. The other approaches are able to detect different partitions, in particular consisting of four groups.



a) b) c)

Fig. 1 Comparison of the three community detection methods for Zachary's karate club network data: a) Girvan-Newman algorithm; b) Louvain algorithm; c) Modal clustering algorithm.

3 Community detection results for Italian statisticians

The three aforementioned community detection methods are used to analyze the co-authorship network defined for the population of the 792 Italian academic Statisticians belonging to five scientific subfields[1], as recorded in the Italian Ministry of University and Research (MIUR) database at March 2010. To collect publications three bibliographic archives –two international (Web of Science and Current Index to Statistics)– and one national based on publications attached to the nationally funded grants (PRIN projects)– are considered [5]. Hence the co-authorship network under analysis is the result of combining multiple data sources [11].

The general aim of the community detection procedures here adopted is to discover if the co-authorship network of Italian statisticians can be clustered into com-

[1] The five subfields established by the Italian governmental official classification are: Methodological Statistics, Statistics for Experimental and Technological research, Economic Statistics, Demography, and Social Statistics.

munities. In order to let the results comparable, the three community detection methods are performed on the largest connected component of the given graph (i.e., giant component). This approach, recognized in the related literature in order to isolate disjoint components [15], is useful in our case given that only the Modal clustering algorithm is able to handle disconnected graphs. In the observed co-authorship network, the giant component consists of 660 authors, representing the 82% of statisticians. Therefore the analysis can be restricted to this set of authors without loss of generality. In performing the Modal clustering method, two different density functions (degree and betweenness) are chosen.

The main results of the three procedures are reported in Table 1. In general, the methods are quite comparable in terms of number of detected communities and of their sizes. The Girvan-Newman algorithm produces the larger number of communities (#. 22). Also the quality of the partitions, measured by the modularity index Q, is quite similar across methods. The lower value is associated with Modal clustering with the betweenness as density function, that is the method that also gives raise to communities of relative larger sizes with respect to the other two methods.

Table 1 Performance measures of giant component of the Italian statisticians co-authorship network by methods. **C**= #. of detected communities, **Average**= Average number of authors in communities (St.Dev.), **Q**= modularity index

Method	**C**	**Average (St.Dev.)**	**Q**
Girvan-Newman	22	30.000 (15.754)	0.752
Louvain	18	36.667 (17.283)	0.762
Modal clustering (betweenness)	13	50.769 (30.444)	0.702
Modal clustering (degree)	18	36.667 (23.118)	0.761

The Modal clustering (with degree as density function) and the Louvain algorithm show the highest –and similar– values of the modularity index as well as the same total number of detected communities (#. 18). In the following, the composition of the first 9 larger communities identified by these two approaches is analyzed. These larger communities are quite representative since for both methods they comprise about the 70% of the 660 statisticians in the giant component.

Table 2 reports some descriptive measures of the 9 communities listed in descending order by size. In both algorithms, the detected communities share quite similar structural characteristics. By way of example, the largest community (C1) comprises 91 and 69 statisticians, for the Modal clustering and the Louvain algorithm, respectively.

The author average degree –computed within the community– is usually comparable across methods, ranging from a minimum of 1.75 (community C4 by Modal clustering) to a maximum of 4.04 (community C3 for Louvain algorithm). The ratio between within-community links (edges representing the relationship in the same community) and the external links (edges activated with non members of the community) are quite small for both methods. Looking at the internal composition by scientific subfield and university affiliation, in the Louvain method, the largest community includes several authors in the Statistics subfield.

In the Modal clustering, the emerging largest community is composed mostly of authors in Statistics subfield and some authors in Economic statistics subfield, mainly clustered according to the geographic proximity of their universities. In particular, the majority of authors in this cluster is affiliated to the universities located in the North and in the Centre of Italy (e.g., Florence, Padua, Rome and Milan). The same differences arise looking at the composition of the other larger detected communities. Both methods find clusters that are homogeneous by scientific sectors (demographers and social statisticians, on the one hand, and methodological statisticians, on the other hand, tend to create strong communities), although it seems that Modal clustering groups together authors on the basis of links mainly driven by the geographic proximity of the universities in which they are affiliated, while Louvain algorithm aggregates authors on the basis of network characteristics.

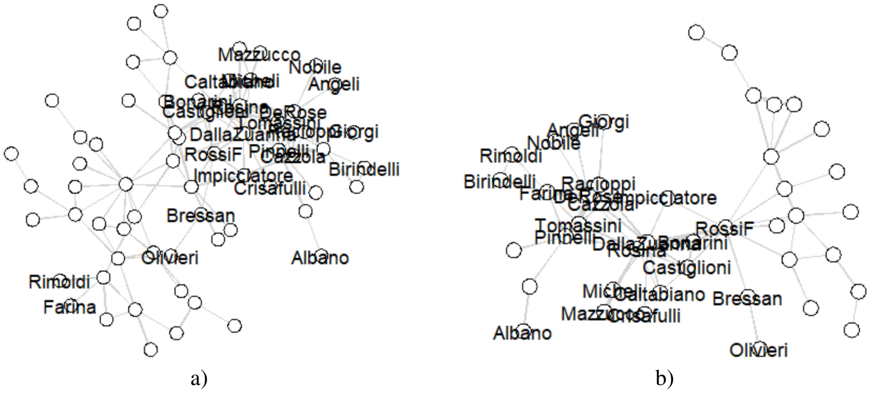
Generally speaking, comparing all possible couples of communities, the overlapping among the detected communities is low. The average Jaccard index is indeed equal to 0.02. Only some communities present a sort of overlapping with about 30% of common members, as showed in the example in Figure 2 for community 1 (C1) in the Louvain algorithm and community 4 (C4) in the Modal clustering algorithm. These methods are therefore able to capture common relational aspects of the observed co-authorship network enriching the interpretation of the findings related to the authors' attributes.

Table 2 Descriptive measures of the first 9 detected communities obtained by the Modal clustering (MC) and the Louvain algorithms for the giant component of the Italian statisticians co-authorship network.

Community	Size		Average degree author		Intra-Extra links ratio	
	MC (degree)	Louvain	MC (degree)	Louvain	MC (degree)	Louvain
C1	91	69	3.52	2.92	0.120	0.056
C2	67	65	2.48	3.75	0.063	0.092
C3	57	53	2.60	4.04	0.057	0.022
C4	49	49	1.75	4.00	0.011	0.021
C5	49	49	2.00	2.98	0.016	0.021
C6	48	48	2.00	3.29	0.039	0.026
C7	48	44	2.17	3.36	0.018	0.076
C8	47	41	3.23	3.61	0.038	0.056
C9	44	40	2.32	3.35	0.036	0.050

4 Conclusions

The general aim of the community detection procedures here adopted was to discover if the co-authorship network of Italian statisticians can be clustered into communities. To this purpose, results from three different community detection meth-



a) b)

Fig. 2 Representation of the communities with the largest overlapping number of actors: a) community 1 (C1) Louvain algorithm; (b) community 4 (C4) Modal clustering algorithm. The names of the statisticians common to both communities are displayed.

ods, the Girvan-Newman algorithm, the Louvain algorithm and the Modal clustering algorithm, have been compared by presenting performance measures and specific internal communities interpretations. The most suitable methods in terms of quality of the partitions discovered are the Modal clustering algorithm and the Louvain algorithm.

As general evidence, it seems that the co-authorship network of the Italian statisticians is clustered in a relatively small number of communities with different internal composition that is mainly determined by authors' scientific field and university affiliation.

In order to find denser communities it would be important to consider in the analysis also the strength of the collaboration relationship by using the number of co-authored papers among couples of authors. As future line of research we will intend to extend the described community detection methods to weighted networks. It also would be interesting to explore the community structures dealing with the presence of multiplex networks, when collaboration is described by measuring also other kinds of relationships among scientists (e.g., co-participation on funded projects).

References

1. Adelchi Azzalini and Nicola Torelli. Clustering via nonparametric density estimation. *Stat Comput*, 17(1):71–80, Mar 2007.
2. Elisa Bellotti, Luka Kronegger, and Luigi Guadalupi. The evolution of research collaboration within and across disciplines in italian academia. *Scientometrics*, 109:783–811, 2016.
3. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J Stat Mech*, 2008(10):P10008, 2008.
4. A Clauset, Mark EJ Newman, and C Moore. Finding community structure in very large networks. *Phys Rev E*, 70:066111, 2004.

5. D. De Stefano, V. Fuccella, M.P. Vitale, and S. Zaccarin. The use of different data sources in the analysis of co-authorship networks and scientific performance. *Soc Networks*, 35:370–381, 2013.
6. D. De Stefano and S. Zaccarin. Co-authorship networks and scientific performance: an empirical analysis using the generalized extreme value distribution. *J Appl Stat*, 43:262–279, 2016.
7. Patrick Doreian, Vladimir Batagelj, and Anuška Ferligoj. *Generalized Blockmodeling. Structural Analysis in the Social Sciences*. Cambridge University Press, New York, NY, USA, 2005.
8. A. Ferligoj, L. Kronegger, F. Mali, T.A.B. Snijders, and P. Doreian. Scientific collaboration dynamics in a national scientific system. *Scientometrics*, 104:985–1012, 2015.
9. Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Phys Rep*, 659:1–44, 2016.
10. Linton C Freeman. Centrality in social networks conceptual clarification. *Soc Networks*, 1:215–239, 1978.
11. V. Fuccella, D. De Stefano, M.P. Vitale, and S. Zaccarin. Improving co-authorship network structures by combining multiple data sources: evidence from italian academic statisticians. *Scientometrics*, 107:167–184, 2016.
12. G. Menardi and D. De Stefano. Modal clustering of social network. In S. Cabras, T. Di Battista, and W. Racugno, editors, *Proceedings of the 47th SIS Scientific Meeting of the Italian Statistical Society*. CUEC Editrice, Cagliari, 2014.
13. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E*, 69:026113, 2004.
14. Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.
15. Miloš Savić, Mirjana Ivanović, Miloš Radovanović, Zoran Ognjanović, Aleksandar Pejović, and Tatjana Jakšić Krüger. Exploratory analysis of communities in co-authorship networks: A case study. In *ICT Innovations 2014*, pages 55–64. Springer, 2015.
16. M.T. Schaub, J.C. Delvenne, M. Rosvall, and Lambiotte R. The many facets of community detection in complex networks. *Appl Netw*, 2: 4:1–13, 2017.
17. W.W. Zachary. An information flow model for conflict and fission in small groups. *J Anthropol Res*, 33(4):452–473, 1977.