The Collaboration Potential, an index to assess the roles of scientists in their coauthorship networks

Francesco Giuliani[1], Michele Pio De Petris[1] and Giovanni Nico[2]

[1] Innovation and Technological Development, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy,
f.giuliani@operapadrepio.it, m.depetris@operapadrepio.it,
[2] Istituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche (CNR-IAC), Italy, g.nico@ba.iac.cnr.it

Abstract. Over the past decade there have been many investigations aimed at defining the role of scientists in their coauthorship networks. In this work we propose an analytical definition of a collaboration potential between authors of scientific papers based on both coauthorships and content sharing. The collaboration potential can also be considered a tool to investigate the weakness of the network in terms of 'lost collaborations' between authors with same scientific interests. This work is an abbreviated version of the original article from the same authors [1].

Keywords: social network analysis, scientific collaborations, coauthorships, collaboration potential

1 Introduction and methodological approach

In this work we present a method aimed at investigating the informative potential that modern bibliographic databases offer. We study the publication output of researchers and try to find an index describing both collaborations and content sharing in scientific networks.

Considering coauthorship as synonymous of collaboration, we can define a collaboration index between author A and author B as

$$P_{AB} = \frac{dim(P_A \cap P_B)}{dim(P_A)} \ , \tag{1}$$

where P_A and P_B represent the sets of papers authored by A and B respectively, $dim(P_A)$ represents the number of elements (articles) authored by author A and $dim(P_A \cap P_B)$ represents the number of articles shared by authors A and B as coauthors. This index represents for author A the fraction of articles he has written in collaboration with author B. This index, taken alone, does not tell the whole story about collaboration as it is independent from article contents.

One can build an index to express content sharing defining it as the number of keywords author A and author B share divided by the number of keywords of

author A. This index is a measure of the commonality of scientific interests, but does not take into account collaborations between authors. If we want to measure the collaboration potential between two authors we need to build a consistent index taking into account both coauthored papers and contents of such papers, that in our model are represented by article keywords.

We want keywords to come from an unambiguous and limited set of terms, so we chose to study only publications indexed by the PubMed search engine. For such publications, keywords come from MeSH (Medical Subject Headings) database, a controlled vocabulary thesaurus used for indexing articles in PubMed. We simply used a custom query and XML parsing in order to associate keywords to articles of our interest.

2 Measuring the collaboration potential

In our simple model we start taking coauthorships into account. We can define the set of articles author A has not co-authored with author B (and vice versa) as

$$\overline{P}_A = P_A - (P_A \cap P_B) \ , \ \overline{P}_B = P_B - (P_A \cap P_B) \tag{2}$$

The articles belonging to these sets are associated with their respective keywords, i.e. we can define the sets $\overline{K}_A$ and $\overline{K}_B$ containing the keywords of the papers the two authors have not respectively coauthored. The intersection between these two sets

$$\overline{K}_{AB} = \overline{K}_A \cap \overline{K}_B \ , \tag{3}$$

represents the keywords shared by the articles the authors have not co-authored. So we can formulate the collaboration potential based on non-coauthorship for author A towards author B as

$$m_{AB} = \frac{\dim(\overline{K}_A \cap \overline{K}_B)}{\dim(\overline{K}_A)} \ . \tag{4}$$

This index has many interesting characteristics. It is defined in the interval $[0, 1]$ and is 0 in two circumstances:

- First case: the two authors have coauthored all their articles. In this case the sets $\overline{K}_A$ and $\overline{K}_B$ are both void so $\dim(\overline{K}_A \cap \overline{K}_B)$ is 0, the authors having fully exploited their collaboration potential, having co-authored all they could, i. e. all the articles they wrote.
- Second case: for the articles they have not coauthored, they worked on totally different subjects. In this case $\overline{K}_A \cap \overline{K}_B$ is void meaning that the authors, excluding coauthored articles, share no common scientific interests and so, according to our model, no collaboration potential exists between them.

We can discriminate between the two cases in which $m_{AB} = 0$ according to the corresponding value of P_{AB}. In fact a value $P_{AB} = 1$ corresponds to the first case, while a $P_{AB} \neq 1$ to the second case.

In all other cases m_{AB} different from 0 implies the existence of a not fully "exploited" collaboration between authors A and B. The other extreme value of the index is 1. In this case $\overline{K}_A = \overline{K}_{AB} = \overline{K}_B$, i.e. author A and author B share all their keywords for the articles they have not coauthored. It is worth noting that the collaboration potential we've just defined should not be considered a "predictor" of future collaborations but it is intended to investigate the role of scientists in the collaboration network. Other approaches were proposed in the literature [2], [3] and methods were presented in order to predict the evolution of links in a social network based on topology taken alone. Our method is quite different because it relies on intrinsic node properties (identified as keywords), and tries to investigate properties of links in terms of 'lost collaboration' between the authors.

Extending 4 we can easily compute the collaboration potential between author A and group G considering the group as a single author, i. e. considering the set of articles written by author A and the set of articles written by all other authors of group G. We thus obtain:

$$m_{AG} = \frac{\dim(\overline{K}_A \cap \overline{K}_G)}{\dim(\overline{K}_A)} \ , \tag{5}$$

where $\overline{K}_G$ represents the set of keywords for the articles author A has not coauthored with the other authors belonging to group G. If author A belongs to group G the index in (5) expresses the collaboration potential the author has with the colleagues of his own group, supposedly studying the same subjects of his researches and publications.

3 Application of the method and discussion of the results

To apply our method we considered the publications and authors of the Casa Sollievo della Sofferenza research hospital in years 2004-2009. For all authors (216) and publications (711 papers, with a mean of 14.42 keywords per article) we computed the collaboration potential according to eqs (4) and (5). We found a mean value for P_{AG} of 90.50%, confirming that scientists coauthor the largest majority of their publications with authors of their own group than with authors belonging to other research groups of the institute.

In order to investigate the role of scientists inside and between research groups, we considered the values of P_{AG} and of m_{AG} for each researcher (see fig.1). The majority of authors concentrate on the bottom-right area of the plot. This result confirms that generally authors have a low collaboration potential with colleagues of their groups. The value of the collaboration potential is exactly zero for 81.48% of authors. This result is simply understandable in terms of coauthorships, in fact we have found that in all cases in which m_{AG} is zero P_{AG} is one, meaning that each of these authors' publication is coauthored by at least one other author of the group the author belongs to. We could define these authors as highly integrated with their research units, writing their papers with

at least one of the colleagues of their groups. Furthermore, we found a small subset of authors having a low value of P_{AG} and an high value of m_{AG} with their group. We can easily define these authors as "independent" as they share no article with the members of their own group, given many subjects on which they "could" have written articles together.

Eventually, generalizing the concept of collaboration to a broader scope, the methods presented herein could easily be used to define a collaboration potential in every case in which one can classify the content of some activities and determine which of them are in common among the actors cooperating to perform such activities.

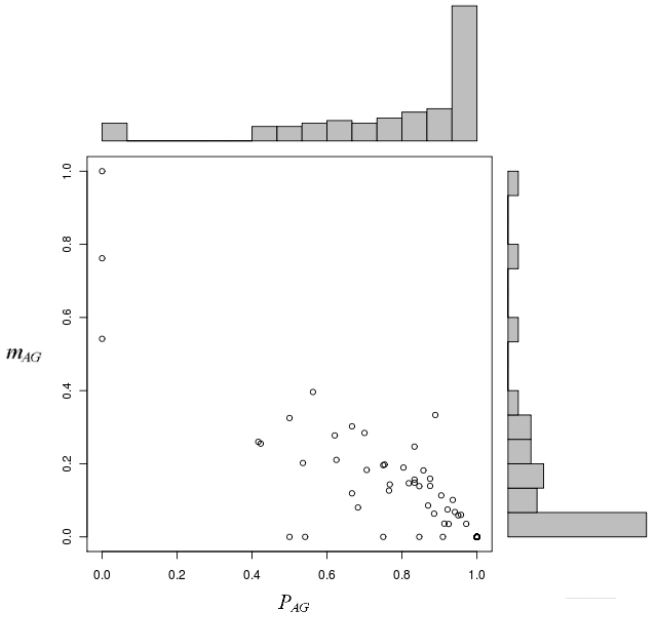

Fig. 1. Distribution of authors according to the collaboration potentials toward their research groups (m_{AG}) and coauthorship sharing (P_{AG}) values. The grey bar graphs on the axes show the frequency distributions of the number of authors for each interval of (m_{AG}) and (P_{AG}).

References

1. *Giuliani F, De Petris MP, Nico G*, Assessing scientific collaboration through coauthorship and content sharing, *Scientometrics*, 85(1), p13–28, October 2010.
2. *David Liben-Nowell and Jon Kleinberg*, The Link-Prediction Problem for Social Networks, *Journal of the American Society for Information Science and Technology*, 58(7), p1019–1031, May 2007.
3. *Leo Katz*, A new status index derived from sociometric analysis, *Psychometrika*, 18(1), p39–43, March 1953.