A Social Network Analysis based approach to deriving knowledge about research scenarios in a set of countries*

Paolo Lo Giudice[1], Paolo Russo[2], and Domenico Ursino[3]

[1] DIIES, University “Mediterranea” of Reggio Calabria
[2] NEGG
[3] DICEAM, University “Mediterranea” of Reggio Calabria

(DISCUSSION PAPER)

Abstract. In this paper, we propose a new Social Network Analysis based approach to providing a multi-dimensional picture of the research scenarios of a set of countries of interest and to detecting possible “research hubs” operating therein. This knowledge allows innovation managers to understand the impact of different socio-economic conditions on the research level of a country. Furthermore, it may help the design of policies for supporting the accumulation of scientific and technological capabilities. In the last part of this paper, we apply our approach to four North African countries (i.e., Algeria, Egypt, Morocco and Tunisia) in such a way as to show its potential.

1 Introduction

In the last years, scientometrics and bibliometrics received a growing interest both in research literature and as objective ways for evaluating the performances of researchers, universities, institutions, etc. [12, 10, 9, 7, 3]. The obvious consequence of this fact is that more and more innovative approaches to supporting these activities have been recently proposed. Social Network Analysis [13] and, more in general, graph theory, represent a prominent family of approaches adopted in the past in this context (see, for instance, [4, 5, 8, 6, 2]). Furthermore, it is possible to foresee that they will be even more adopted in the future.

This paper aims at providing a contribution in this setting. Indeed, it proposes a new Social Network Analysis based approach to extracting knowledge about research scenarios and spillovers [11] or, better, hubs in a set of countries. As for this paper, a *hub* is a research institution that operates as a guide or stimulus to research in its country and, at the same time, is capable of stimulating cooperations with institutions of other countries. Our hub definition is fitted to the scenario of our interest and strongly benefits from the observations, suggestions and experience of innovation management researchers, who guided

* The authors thank Proff. Fabio Landini and Roberto Mavilia of the Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship of University “Bocconi”, who supplied data for experiments and cooperated in the analysis of results.

us in its formulation. Our approach is general and can be directly applied to any set of countries. The only requirement is to have at disposal the set of the publications of all the research institutions of the countries to investigate. In this paper, we applied it to four North African countries (e.g., Algeria, Egypt, Morocco and Tunisia), and we exploited all the publications of all the research institutions of the four countries of interest in the time interval $[2003, 2013]$, as stored in the Web of Science repository [1]. The most important support data structure is a social network whose nodes represent institutions and whose edges denote collaborations among institutions. Starting from it, other important support data structures and accompanying parameters (some of which were never defined in the literature) are introduced.

This paper is organized as follows. In Section 2, we illustrate our approach. In Section 3, we apply it to the four North African countries mentioned above. Finally, in Section 4, we draw our conclusions and overview some possible future developments.

2 Description of our approach

Our approach operates on a set Pub of publications at our disposal and on a set C of countries to investigate. It will be described in the next subsections.

2.1 Hub characterization and detection

In this section, we aim at defining a method for detecting both hubs and their features in a set of countries. For this purpose, we preliminarily introduce a first support data structure. It is a social network: $G = \langle N, E \rangle$. N is the set of the nodes of G. A node $n_i \in N$ corresponds to exactly one institution registered in our database. Since there is a biunivocal correspondence between a node of N and the corresponding institution, in the following, we will use the symbol n_i to indicate both of them. Each node of N is labeled with an element of C depending on the country of the corresponding institution. We indicate by l_i the label of n_i. E is the set of the edges of G. There exists an edge $e_{ij} = (n_i, n_j, w_{ij}) \in E$ if there exists at least one publication involving one author of n_i and one author of n_j. w_{ij} is the weight of e_{ij}; it denotes the number of publications having at least one researcher of n_i and one researcher of n_j among their authors. If a paper has more than one author of the same institution n_i, then there is a self-edge linking n_i with itself.

Now, we are able to introduce the concept of hub. With regard to this fact, we point out that we do not aim at proposing a new concept, characterized by a mathematical foundation supporting it. Instead, we would like to introduce an informal and empirical, yet reasonable, concept, which can support innovation managers to make their decisions. In carrying out this activity, we strongly benefited from the observations, suggestions and experience of innovation management researchers, who guided us in its formulation. Taking this purpose into account, we can say that, in order to be a hub, an institution must satisfy the following conditions: *(i)* C_1: it should have published *a very high number* of papers; *(ii)* C_2: it should have published *a high number* of papers in cooperation with institutions different from the ones of its country; *(iii)* C_3: it should have published *many papers* in cooperation with institutions of its country.

To "quantify" conditions C_1, C_2 and C_3, we use three metrics, namely M_1, M_2 and M_3, respectively. M_1 coincides with the classical weighted degree centrality, M_2 coincides with the normalized weighted degree centrality and M_3 is analogous to the E-I index [13].

As theoretically conjectured in the past literature, and as verified for the countries composing our case study, M_1, M_2 and M_3 follow a power law distribution. Taking all these considerations into account, the set $\mathcal{H}^X$ of hubs for the countries into consideration can be defined as the set of those institutions simultaneously belonging to the top $X\%$ of the institutions with the highest values of M_1, M_2 and M_3 (we call I_1^X, I_2^X and I_3^X these three sets, when considered separately). In this definition, X is a threshold allowing the selection of the institutions having the highest values of M_1, M_2 and M_3. The choice to use X as a threshold parameter derives from the power law distributions characterizing all the three metrics. Reasonable values of X could be 10, 15 and 20. After several experiments, we decided to consider a default value of X equal to 20. As a consequence, in the following, when X is not specified, we intend that it is equal to 20. Below, we use the symbol $\mathcal{H}_k^X$ to indicate the hubs of a given country k.

2.2 Investigation of the research scenarios

In this section, we aim at analyzing the research scenarios of the countries into examination. Initially, we can introduce three indicators that could give us some knowledge about the research scenarios of the countries into consideration. The first one, RQ, is an indicator of the overall research quality in the countries of interest. In fact, it measures how many institutions of I_1 belong to these countries. The second one, FC, indicates how many institutions, among the top ones of the countries of interest, publish many papers with foreign institutions. The third one, TP, indicates how many institutions that publish very much with foreign institutions belong to the top institutions of the countries of interest.

In the investigation of the research scenario of a country k and of the role of its hubs, it appears very interesting to analyze its paper distribution. For this purpose, we introduce the average number $AvgPub_k^{\mathcal{H}}$ of the publications of its hubs. Another interesting issue to investigate is to verify if a hub of k publishes more with institutions of k (we call "internal" the corresponding publications) than with foreign ones (we call "external" the corresponding publications) or alone. To carry out this investigation, we introduce: *(i)* the average number $AvgHubPub_k^I$ of publications performed by the hubs of k with institutions of k; *(ii)* the average number $AvgHubPub_k^F$ of publications performed by the hubs of k with foreign institutions; *(iii)* the average number $AvgHubPub_k^A$ of publications performed alone by the hubs of k (we call them "alone publications" in the following).

A further interesting analysis is devoted to understand if, in their cooperation with foreign institutions, the hubs of a given country k privilege one or few countries. For this purpose, we specialize to our research context the Herfindahl Index. This index is very used in economics. It is defined as the sum of the squares of the market shares of the firms within the industry, where market shares are expressed as fractions. It can range from 0.0 to 1.0, moving from a huge number of very small firms to a single monopolistic producer. In our case, we extend the Herfindahl index to our context and define the Herfindahl Index HI_k associated with the papers published by the hubs of k to verify if these hubs published in cooperation with institutions of few (implying high values of HI_k) or many (implying low values of HI_k) countries.

2.3 Cooperation among hubs of the same country

In this section, we aim at investigating the cooperation levels of the hubs of a given country k. For this purpose, we preliminarily define a support data structure called *clique social network*. In particular, let G be the social network defined in Section 2.1 and let G_k be its "projection" on the country k. Let $\mathcal{C}_k$ be the set of cliques of G_k and let $\mathcal{H}_k$ be the set of the hubs of k. A *clique social network* CG_k has a node for each hub of $\mathcal{H}_k$ belonging to at least one clique of $\mathcal{C}_k$. Each node n_i of CG_k has associated a weight w_i denoting the number of cliques of $\mathcal{C}_k$ which it belongs to. An edge (n_i, n_j) of CG_k denotes that n_i and n_j together belong to at least one clique of $\mathcal{C}_k$.

Some measures capable of quantitatively representing the differences that characterize the cooperation among hubs are the following: *(i)* the number of cliques $|\mathcal{C}_k|$; *(ii)* the absolute dimension $d_{\mathcal{C}_k}$ of the largest clique in $\mathcal{C}_k$; *(iii)* the relative dimension $\frac{d_{\mathcal{C}_k}}{|\mathcal{H}_k|}$ of the largest clique in $\mathcal{C}_k$; *(iv)* the fraction $f^{\mathcal{H}}_{\mathcal{C}_k}$ of hubs belonging to at least one clique of $\mathcal{C}_k$. In order to avoid that results are biased by the number of publications (which can be very different in the different countries of interest), we defined a normalized version $\widehat{CG_k}$ of CG_k. Finally, we searched for some measures to compare clique social networks. After several experiments, we found that the most significant ones were: *(i)* the number of nodes; *(ii)* the number of edges; *(iii)* density[4].

2.4 Investigation about the quality of publications

All indicators introduced above are based only on the number of publications. Actually, it would be important to take also their quality into account. One way to do this consists in taking their impact factor into consideration; another way consists in considering the number of citations received by papers. Impact factors are measured only for journal papers. As a consequence, if we want to exploit this measure, we must define a new support data structure. This structure, that we indicate by G', is, once again, a social network. It is defined as $G' = \langle N', E' \rangle$.

[4] Actually, this last measure can be derived from the two other ones. However, it is very expressing and, consequently, we decided to explicitly consider it.

There is a node $n_i \in N'$ for each institution having at least one author, who published at least one journal paper. An edge $e'_{ij} = (n'_i, n'_j, w'_{ij})$ has a semantics similar to the one of e_{ij} except that the weight $w'_{ij} = \sum_{p \in (Pub_{ij} \cap JPub)} IF_p$ considers both the number of publications performed by n_i and n_j simultaneously and the corresponding impact factors. Paper citations are valid both for conference proceedings and for journal papers. However, in order to make our analyses about the quality of publications homogeneous, we chose to investigate only journal papers. In this case, we used the same support social network as the one exploited for impact factors, but the edge weights w'_{ij} was computed as: $w'_{ij} = \sum_{p \in (Pub_{ij} \cap JPub)} CitN_p$, where $CitN_p$ is the number of citations of p.

2.5 Characterization of hub neighborhoods

A first parameter useful to characterize hub neighbors is the average number *AvgPub* of publications of the hub neighborhoods. A second parameter regards their average dimension $AvgDim$. Even in this case, we disaggregate data per country, and we call $AvgDim_k$ the corresponding parameter for the country k.

A next analysis regards the cooperation level among the institutions belonging to hub neighborhoods. To perform this task, we define a new support social network. We call it *nbh social network* and we represent it by means of the symbol $NbhG_i$. Given a neighborhood nbh_i, the corresponding nbh social network is defined as follows: $nbhG_i = \langle nbh_i, nbhE_i \rangle$. There is a node in $NbhG_i$ for each node of nbh_i; there is an edge $(n_i, n_j) \in nbhE_i$ if there exists at least one publication between an author of n_i and an author of n_j.

After having introduced this social network, we define a first parameter on it. This parameter is called $AvgCFrac$ and corresponds to the average fraction of the real number of cliques existing in hub neighborhoods against their possible number. It is an indicator of the cooperation level among hubs. As usual, we call $AvgCFrac_k$ the "projection" of $AvgCFrac$ on the country k. A second parameter about intra-neighborhood cooperation regards the average fraction $AvgCNbh$ of the number of cliques existing in hub neighborhoods against the number of neighborhood nodes. Again, we call $AvgCNbh_k$ the "projection" of $AvgCNbh$ on the country k. A final parameter measuring the cooperation level among hub neighbors is the average density $AvgDens$ of the nbh social network. As usual, we call $AvgDens_k$ the "projection" of $AvgDens$ on the country k.

3 Application of our approach to four North African countries

As pointed out in the introduction, we applied our approach to four North African countries, namely Algeria, Egypt, Morocco and Tunisia. As a consequence, in our case study, the set C of the countries to investigate (see Section 2) consisted of the following elements: $C = \{$'A', 'E', 'M', 'T', 'O'$\}$, where 'A' (resp., 'E', 'M', 'T', 'O') stands for 'Algeria' (resp., 'Egypt', 'Morocco', 'Tunisia', 'Others'). Clearly, 'O' indicates all the countries different from the four into

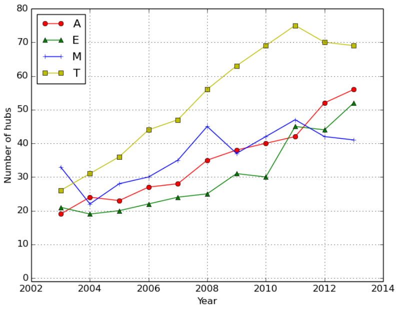

Fig. 1. Number of hubs for each country in the year interval [2003,2013]

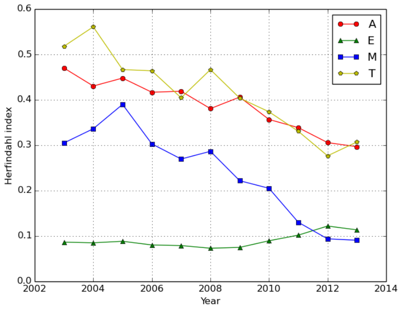

Fig. 2. Herfindahl index over time for the four countries

examination. We chose these four countries because we had the possibility to access the corresponding data thanks to an academic partner that was carrying out a project aiming at investigating their research and innovation scenarios. Our dataset was stored in a MongoDB database. To give an idea of it, we report some of its features: *(i)* dimension = 10.27 GB; *(ii)* number of institutions = 278,696 ; *(iii)* number of authorships = 89,008,846; *(iv)* number of publications = 6,599,104; *(v)* number of research areas = 6; *(vi)* number of research fields = 251. Due to space limitations we can present only a very limited number of the results that we obtained.

In Figure 1, we report the variation of the number of hubs for each country. From the analysis of this figure, we can see that the country with the highest number of hubs is Tunisia. This result was unexpected also because both the extension and the number of citizens of Tunisia were smaller than the ones of the other three countries.

In Figure 2, we report the Herfindahl index HI_k for the four countries. From the analysis of this figure we can observe that Tunisia and Algeria have a high Herfindahl index, which implies that their hubs cooperate mostly with one or few countries. In particular, after some investigations, we have seen that Tunisia and Algeria cooperate mostly with France, which is reasonable if we consider that they are French past colonies. By contrast, Egypt has a very low Herfindahl index, i.e., its hubs cooperate with many countries. An interesting trend is the one of Morocco; in fact, it initially has a behavior like the ones of Tunisia and Algeria, whereas, in the last years, it shows a behavior like the one of Egypt.

To determine the cooperation levels among hubs for the four North African countries into consideration, for each country k, we performed the following tasks: *(i)* we considered the two time intervals $[2003, 2009]$ and $[2007, 2013]$; *(ii)* we computed the clique social networks $CG1_k$ (resp., $CG2_k$), corresponding to the first and the second time intervals, respectively; *(iii)* we measured the four parameters introduced in Section 2.2 for quantitatively evaluating clique social networks. Obtained results for the first time interval are reported in Table 1.

Country	$\lvert \mathcal{C}1_k \rvert$	$d1_{\mathcal{C}_k}$	$\frac{d1_{\mathcal{C}_k}}{\lvert \mathcal{H}_k \rvert}$	$f1^{\mathcal{H}}_{\mathcal{C}_k}$
Algeria	292	7	0.152	0.913
Egypt	38	13	0.351	0.973
Tunisia	130	8	0.116	0.942
Morocco	82	7	0.127	0.818

Table 1. Quantitative differences characterizing the cooperation behaviors of hubs in the four countries of interest in the time interval $[2003, 2009]$

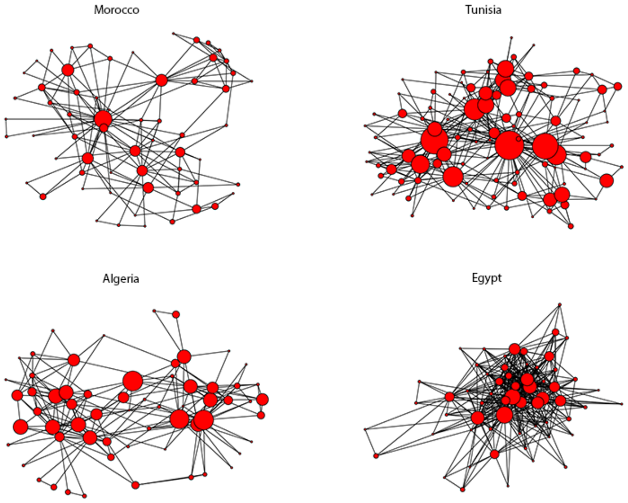

Fig. 3. Graphs $CG2_k$ for all the four countries

From the analysis of this table we can draw the following conclusions: *(i)* Egypt has the largest clique; this clique is much larger than the maximum cliques of the other countries; *(ii)* in Egypt almost all hubs belong to at least one clique. These results indicate that Egyptian hubs are more prone to cooperation than the hubs of the other countries.

In Figure 3, we report the graphs $CG2_k$ for all the four countries; in these graphs the dimension of nodes is proportional to the corresponding weight, i.e., to the number of cliques they belong to. The analysis of this figure confirms the previous conjecture; in fact, the number of edges in the Egyptian graph is much higher than in the other graphs. This fact, along with the presence of many not very large nodes, allows us to derive another important knowledge pattern, i.e., that research activities in Egypt are more "distributed" among hubs.

4 Conclusion

In this paper, we have proposed a new Social Network Analysis based approach to investigating the research scenarios of a set of countries of interest and to detecting possible hubs operating in these countries. Extracted knowledge allows the evaluation of the impact of different socio-economic conditions on research and favors the design of policies for supporting innovation in the countries of interest. We applied our approach to four North African countries. In the future, we plan to exploit techniques for analyzing information diffusion in social networks to understand how the possible mobility of top researchers from one institution to another can impact on the quality of both of them. Moreover, we plan to investigate the possible application of classification techniques to derive hub profiles in different countries. Furthermore, it is necessary to consider that a network changes over time. As a consequence, hubs could also change over time. The evolution of an institution's capability of becoming hub, remaining hub or no longer being a hub is a challenging task that we plan to investigate. Finally, we plan to analyze the possible application of prediction techniques to understand what kind of financial investment must be performed for maximizing the increase of both the number and the quality of hubs and publications in the countries of interest.

References

1. Web Of Science. `http://wokinfo.com/`, 2017.
2. A. Abbasi, L. Hossain, and L. Leydesdorff. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6 (3):403–412, 2012.
3. J. Adams, K. Gurney, D. Hook, and L. Leydesdorff. International collaboration clusters in Africa. *Scientometrics*, 98(1):547–556, 2014. Springer.
4. T. Arif, R. Ali, and M. Asger. Scientific co-authorship social networks: A case study of computer science scenario in India. *Science*, 52 (12):38–45, 2012.
5. K. Badar, J.M. Hite, and Y.F. Badir. Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in Pakistan. *Scientometrics*, 94 (2):755–775, 2013. Elsevier.
6. M. Bordons, J. Aparicio, B. González-Albo, and A.A. Díaz-Faes. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, 9 (1):135–144, 2015.
7. K. Börner, L. dell'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10 (4):57–67, 2005.
8. J. Kim and C. Perez. Co-authorship network analysis in industrial ecology research community. *Journal of Industrial Ecology*, 19 (2):222–235, 2015.
9. F. Landini, F. Malerba, and R. Mavilia. The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics*, 105(3):1787–1807, 2015. Elsevier.
10. P. Liu and H. Xia. Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, 103 (1):101–134, 2015.
11. F. Montobbio and V. Sterzi. Inventing together: exploring the nature of international knowledge spillovers in Latin America. *Journal of Evolutionary Economics*, 21(1):53–89, 2011. Springer.
12. J. Singh. Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96, 2008. Elsevier.
13. S. Wasserman and K. Faust. *Social network analysis: Methods and applications.* 1994. Cambridge University Press.