A new Social Network Analysis-based approach to extracting knowledge patterns about research activities and hubs 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 International,
[3] DICEAM, University "Mediterranea" of Reggio Calabria
{paolo.lo.giudice@unirc.it, paolo.russo@negg.it, ursino@unirc.it}

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 hubs operating therein. This knowledge allows the understanding of the impact of different socio-economic conditions on research. Furthermore, it may help the design of policies for sustaining the accumulation of scientific and technological capabilities. 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. We also investigate (both scientific and commercial) related approaches and specify the main novelties of our approach with respect to them.

Keywords: Bibliometrics, Social Network Analysis, Hubs, Hub Neighborhoods, Innovation Support.

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. Indeed, research collaborations across institutions, firms and countries have been largely investigated in strategy and management literature [57, 48, 19, 49, 28, 52, 29, 56, 54, 24]. Moreover, different studies have been performed to understand whether international flows from developed countries to developing and less-developed ones have some positive effects in these last ones [31]. Furthermore, many studies investigate the impact and the effects of international knowledge flows by focusing on R&D collaborations and inventions and on their impact on innovation [41, 46, 27, 30, 58, 18, 10, 8, 39, 35, 42, 45].

Currently, data available for scientometrics and bibliometrics investigations are growing at a very rapid rate. As a matter of fact, the problem of extracting useful knowledge from these data can be

seen as a Data Mining problem, and in the very next future, big data approaches for solving it will be unavoidable. The obvious consequence of this reasoning is that more and more innovative approaches to addressing this issue are necessary.

Social Network Analysis [61, 15, 14, 9, 22, 23, 40, 50] and, more in general, graph theory, have been a prominent family of approaches adopted in the past in this context (see, for istance [43, 11, 13, 17, 60, 53, 5, 7, 44, 36, 21, 16, 6]). Furthermore, it is possible to foresee that they will be even more employed in the future, due to the more and more increasing number of proposals someway involving them.

All these studies have certainly contributed to a development of the research in innovation dynamics. However, there are still several aspects that need to be deepened. For instance:

- Most of these approaches focus on authors, whereas investigations on institutions would be extremely interesting. This fact is also valid for the paper that, to the best of our knowledge, is the only one analyzing hubs in the past [11]. In fact, in this paper, the definition of hub is centered on authors.

- Most of the previous approaches employed only centrality measures in their analysis, whereas, in Social Network Analysis, there are several other parameters (e.g., the connection level of a network), which are at least as important as centrality.

- The past approaches did not investigate the neighbors of authors or institutions, whereas we know that, owing to the concept of homophily (that is a key concept in Social Network Analysis), the neighbor of a node can strongly influence the behavior of the corresponding author or institution.

This paper aims at providing a contribution in this setting. Indeed, it proposes a new Social Network Analysis-based (hereafter, SNA-based) approach to extracting knowledge patterns about research activities and hubs in a set of countries of interest. As for this paper, a hub is a research institution that operates as a guide or stimulus to the research in its country and, at the same time, is capable of stimulating cooperations with institutions of other countries. Our hub definition is strongly fitted to our scenario of interest. It strongly benefits from the observations, suggestions and experience of innovation management researchers, who guided 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 used 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 [3].

The most important support data structure (already introduced in the past literature) is a social network with nodes that represent institutions and with edges that 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.

Thanks to our approach, it is possible to reconstruct a very detailed and multi-dimensional picture of the research scenarios of a set of countries, as well as to determine analogies and differences among them. In this way, innovation managers have at their disposal some empirical instruments helping their

decisions. Beside providing several knowledge patterns about institutions and their collaboration, not known in the past, this paper provides several other contributions and, in our opinion, some of them are even more important than the extracted knowledge patterns. Indeed:

- it presents a general SNA-based approach that can be applied to extract knowledge about research scenarios and the corresponding institutions for a set of countries;
- it redefines some SNA metrics in such a way as to make them suitable for this application scenario;
- it defines new metrics about institutions and their cooperations not presented in the past;
- it introduces the concept of hub and provides a method to determine the hubs of each country, as well as to investigate their main features;
- it defines new data structures (such as the *clique social network* and the *nbh social network*) allowing the extraction of interesting knowledge about hubs and their neighbors;
- it provides both a visual and a quantitative method to determine the core hubs (if they exist) of a given country.

For an expert, the extracted knowledge patterns are important for at least two reasons. First of all, they may improve her understanding of the impact of different socio-economic conditions on the structure and evolution of scientific collaborations. In this sense, the four countries, which our approach was applied on, present a great heterogeneity along several socio-economic dimensions, such as type and degree of economic specializations, language and culture. Secondly, this analysis may help the design of policy interventions aimed to sustain the accumulation of scientific and technological capabilities in the countries on which our approach is applied. For instance, the identification and analysis of hubs and their interactions with local research communities may lead to the design of policies that explicitly target hubs as key vectors to access and disseminate knowledge from advanced countries.

The algorithms implementing our approach are in Python [2] and the underlying DBMS is MongoDB [1]. As a consequence, our approach is already compliant with big data technology and, therefore, can help very large investigations (for instance, a large set of countries, or countries having a very high number of research institutions and publications, like United States and European countries).

This paper is organized as follows. In Section 2, we present related literature. In Section 3, we describe available data and illustrate the pre-processing activities performed on them. In Section 4, we present our approach. In Section 5, we apply it to the four North African countries mentioned above. In Section 6, we illustrate the main novelties of our approach w.r.t. the related ones and we compare it with three commercial systems, i.e., Elsevier Pure, Scopus and Fingerprint Engine. Finally, in Section 7, we draw our conclusions and overview some possible future developments.

2 Related Literature

Research collaborations across firms and countries have been largely investigated in strategy and management literature. In this field, authors showed that these collaborations play a key role in the acquisition of external knowledge [57, 63] and in the creation of new knowledge [48, 55].

Specifically, in [48], the authors show that cross-regional networking positively influences innovation, at least in Europe. However, they also show that regional labor mobility plays an even more relevant role. In line with the approach adopted in [48], [19] investigates patent application in biotechnology, organic chemistry and pharmaceutical, and shows that network activity across firms and location is extremely important in the localization of knowledge flows. In [63], the authors employ data mining techniques to show that local research groups, characterized by a very high internal cohesion, hinder knowledge transmission. At the same time, they show that scientists with a centralized position in a network have a positive effect on knowledge flow. [49] analyzes the interactions among researchers coming from developing and advanced countries and finds that innovation in Latin American countries was largely influenced by R&D activities carried out on some OECD countries. [28] investigates cross-border inventions between BRICS firms and European Union actors and finds that these inventions are growing more valuable than the domestic ones. [52] studies and analyzes some survey interviews about Nigerian firms and employs them to determine which factors guide these firms to successfully or unsuccessfully adopt industrial innovations. [29] investigates the learning processes and the linkage behavior of small and large, local and foreign firms in Tanzania. [56] analyzes the efficiency of South Africa's innovation system. [54] investigates innovation in Ghana through a multilevel theoretical framework. In [24], the author proposes a framework aimed to evaluate the optimal conditions for innovation in emerging economies, with a special focus on Kenya and Uganda. The paper shows that, in both countries, the human capital and the firm's internal infrastructure play a significant role in innovation.

Another important investigation in innovation is aimed to understand whether international flows from developed countries to developing and less-developed ones have some positive effects in these last countries. The role played by knowledge spillovers is well known in the literature (see, for instance, [31]). These spillovers can operate through many channels, ranging from formal communication methods (e.g., scientific publications) to informal ones (e.g., person-to-person contacts).

In [55], the authors analyze the institutions having the highest impact on collaboration networks. Their researches confirm the existence of elite groups that cooperate with other minor institutions. This form of cooperation allows the creation and diffusion of knowledge on management.

Many studies investigate the impact and the effects of international knowledge flows by focusing on international R&D collaborations and inventions and on their impact on innovation. This investigation was performed at two different levels (i.e., theoretical and empirical ones). From a theoretical point of view, some authors argued that these collaborations can lead to higher-quality innovations, thanks to the contamination of different skills and pieces of knowledge [41, 46]. Other authors hypothesized that international collaborations are not efficient owing to high coordination costs and difficulties to integrate knowledge of different research teams [27, 30, 58]. Empirical studies produced mixed results. In fact, [18] found that, as far as Indian and Chinese inventors are concerned, cross-border inventions receive more citations than the ones produced by the inventors of only one country. In [10], the

authors show that international collaborations positively influence patent quality; at the same time, they evidence the difficulties of research teams to absorb external knowledge. In [8], the authors show that research collaboration in Africa presents an inhomogeneous structure. They also evidence that these collaborations are strongly constrained by several factors, ranging from geography to history, culture and language. In [39], the authors conduct a study on North African countries. They evidence that, in these countries, research collaborations are rapidly changing although they are still weak. In [35], the authors analyze the international transmission of knowledge in USA. In [42], the authors present a deep research about the geography of innovation, based on patent analysis. They show how localized knowledge flows are largely mediated by labor and technology markets. In [45], the author shows that both the links and the h-indexes of co-inventors and co-authors highly enhanced the flows of academic knowledge into industrial patents in South Africa's firms, as well as knowledge diffusion in large R&D and innovation clusters and hubs. In [12], the authors investigate the diffusion of European knowledge. Specifically, they analyze the diffusion of knowledge between European countries and European Neighboring Countries (ENCs). For this purpose, they use several indicators allowing them to evaluate how European knowledge is employed by ENCs. Obtained results show that ENCs can benefit from the interaction with European countries and can "transform" European knowledge and tools in new knowledge and innovation.

Social Network Analysis and, more in general, graph theory have been largely employed to investigate co-authorship networks and research scenarios in the past. The structure of co-authorship networks in three different fields (i.e., Nanoscience, Pharmacology and Statistics) in Spain in the time interval $[2006, 2008]$ is analyzed in [16]. Here, the authors investigate if there exists a relationship between the research performance of authors and their position in co-authorship networks. A co-authorship network in the interdisciplinary field of "evolution of cooperation" is analyzed in [43]. To carry out their investigation, the authors adopt SNA and a modularity measure. A co-authorship network regarding Digital Libraries is investigated in [44]. For this purpose, some support social networks are constructed and analyzed to determine the impact of authors in the co-authorship network. To evaluate this impact, several SNA measures, such as centrality and PageRank, along with a new metric, called AuthorRank, are employed. In [11], the authors investigate co-authorship networks involving four institutions to understand how information flows therein and to detect the leader authors (called hubs). Hubs are defined as those authors having both a high eigenvector centrality and a high betweenness centrality. Hub detection is performed by means of SNA-based techniques. After the detection of the hubs of the four considered institutions, the authors analyze the relationships among them. In [13], a co-authorship network is analyzed to understand if it is possible to link centrality measures with author performances and if the authors' gender can have an impact on their performance. A co-authorship network, constructed starting from the publications in "information visualization" field in the time interval $[1974, 2004]$, is investigated in [17]. In [60], the authors hypothesize that international cooperation networks are self-organizing. To verify their hypothesis, they employ SNA-based techniques, capable of analyzing the growth of these networks. A method for building link predictors in networks with nodes that represent researchers and with links that denote collaborations is proposed in [53]. SNA-based techniques are employed in [5] for examining the effect of social networks on the performance of scholars in a given discipline. A co-authorship network concerning the "industrial ecology" field is investigated in [36], with the goal of evaluating the corresponding research efforts

and results. Blockmodeling techniques are employed in [21] for analyzing a co-authorship network. In [6], the authors investigate whether preferential attachment in scientific co-autorship networks is different for authors with different forms of centrality. An exploratory analysis of co-authorship in the field of management and organizational studies is presented in [7]. Here, the authors determine the frequency of collaboration in the most prominent journals in the field. In [34], the authors apply classical SNA-based techniques on a complex co-authorship network to find knowledge patterns about paper citation, cooperation trends, the evolution of key components and author ranking. Furthermore, they employ the network diameter, clustering coefficient and degree distribution to find connectivity patterns, small-world network phenomena, and several other properties. In [25], the authors analyze a set of shared papers by constructing a two-mode network with node that represent both authors and papers, and by applying new regression models on it. After this, they employ the obtained knowledge to perform an empirical analysis on a larger co-authorship network. In [26], the authors apply SNA-based techniques on a co-authorship network for estimating cooperation trends and for identifying the most important scientists and institutions. Furthermore, they investigate the possible application of their approach and of the derived knowledge to a medical context. In [20], the authors start from a demographic analysis to provide an overview of the corresponding distribution of both scientific labels and academic titles. In particular, they employ Social Network Analysis to investigate a co-authorship network and several citation metrics.

3 Available data and preprocessing

The dataset we used was derived from Web of Science of Thomson Reuters. It stores all the publications performed by all the research institutes of the four countries into examination from 2003 to 2013. It consisted of four parts concerning:

- *Institutions*. This part contains information about all the research institutions of the four countries into consideration, as well as about the research institutions of the other countries cooperating with them from 2003 to 2013.
- *Authorships*. This part contains information about all the authorships concerning papers involving at least one of the institutions into consideration from 2003 to 2013.
- *Publications*. This part stores information about the publications of the authors affiliated to at least one of the institutions into consideration.
- *Research areas and fields*. This part stores information about the research areas and fields, as classified by Web of Science.

A first analysis of our dataset allowed us to verify that the parts concerning institutions and publications needed some adjustments. In the following subsections we describe these adjustments.

3.1 Choice of similarity metrics

The first task to do for cleaning data was the choice of one or more metrics capable of indicating if two strings are similar or not. In the literature, several string similarity metrics have been already

proposed in the past. When adopted, they are generally coupled with a threshold in such a way that two strings can be considered similar if the value returned by a similarity metric is higher than the threshold. The choice of the threshold is extremely difficult. In fact, if it is excessively low, too much false positives could be obtained; in this case, dissimilar strings would be considered similar. By contrast, an excessively high threshold would lead to too much false negatives.

In the literature the most used similarity metrics are Levenshtein, NeedlemanWunch, Smith-Waterman, Jaro, QGram Distance, Block Distance, and Jaccard Similarity. After a first analysis of the strengths and the weaknesses of the most known metrics, it was necessary to determine the most suited to our scenario. For this purpose, we considered all the institutions of a country (i.e., Afghanistan) and we applied all metrics to them. We chose Afghanistan because it was the first country in our list and because the number of its institutions made it possible a manual (and, therefore, much more precise) check of the results obtained by applying all candidate metrics.

We almost immediately determined that only one metric was not sufficient to obtain accurate results. After several tests, we found that it was sufficient to detect a pair of (at least partially) complementary metrics. Further tests showed that the most promising pair was formed by Jaccard Similarity and QGram Distance. As previously pointed out, the choice of the metrics was strictly connected with computing the most suited thresholds. For this purpose, we conducted an experimental campaign by executing an optimization algorithm, based on a hill climbing methodology. This algorithm aimed at maximizing the number of corrected results on Afghanistan data. It found that the best threshold value was 0.71 for Jaccard Similarity and 0.75 for QGram Distance.

3.2 Description of the algorithm for determining string similarity

After having chosen metrics and thresholds, we had to define a cleaning algorithm to use in the next steps of our ETL activity. This task was difficult. In fact, it was necessary to guarantee a possible "transitive closure" of similarities, assuming that the choice of thresholds in the previous step was capable of avoiding that an excessive usage of this closure would have led to consider as similar some strings that actually were dissimilar. To better explain this problem, consider the following example. We have three strings, namely "Paolo Russo", "Pao Russo" and "Pao Ru", representing the (possibly abbreviated) surname and name of an author. If our algorithm would have determined a similarity between "Paolo Russo" and "Pao Russo" and another similarity between "Pao Russo" and "Pao Ru", it should have been capable of understanding that there exists a similarity between "Paolo Russo" and "Pao Ru".

In order to handle transitive closure to the best, we adopted the support data structure that appeared the most adequate to conceptually representing and explaining this phenomenon, i.e., a graph. This graph G_{Sym} consists of a set N_{Sym} of nodes and a set E_{Sym} of edges. There is a node n_i for each string to evaluate. There is an edge (n_i, n_j) if the strings associated with n_i and n_j have been found to be similar by applying the Jaccard Similarity with a threshold of 0.71 and/or the QGram Distance with a threshold of 0.75.

Once G_{Sym} has been created, finding all the possible similarities among sets of strings can be carried out by finding all possible connected components in G_{Sym}. After the sets of similar strings have been found, ETL represents all the strings associated with the same university by means of a

unique string in the underlying database.

3.3 Application of our ETL algorithm on available data

Our ETL activities concern the fields `City` and `Inst_name` of the part *Institutions*. As for `City`, our approach clusters the values of this field on the basis of the corresponding country. In this way, it avoids homonymies concerning cities having the same name but belonging to different countries. For the same reason, the values of `Inst_name` are clustered on the basis of the country, the city and the category.

For each cluster of `City`, we constructed a graph G_{Sym} and applied the algorithm described in Section 3.2. This algorithm computed the connected components and, for each of them, selected a city name to represent it and stored this name in `City` accordingly. For each cluster of `Inst_name`, we proceeded analogously to `City`, but the strings representing connected components were suitably stored in the field `Institution_Name1`. Furthermore, for each connected component, we registered an auto-increment number in the field `Inst_ID`.

4 Description of our approach

As pointed out in the Introduction, our approach is aimed to extract knowledge patterns about research activities and hubs in a set of countries of interest starting from the publications of their research institutions, as stored in the Web of Science repository. Before starting its description, we must define some sets that formalize available data and, therefore, will be extensively used below.

The first set regards the set RA of research areas. It consists of the following elements:

$$RA = \{\text{'NS', 'AS', 'MH', 'SS', 'HU', 'ET'}\}$$

where 'NS' (resp., 'AS', 'MH', 'SS', 'HU', 'ET') stands for 'Natural Science' (resp., 'Agricultural Science', 'Medical and Health Science', 'Social Science', 'Humanities', 'Engineering and Technology').

The second set concerns the overall set Pub of publications at our disposal. Given a publication $p \in Pub$, we indicate by $Authors_p$ the set of its authors and by $Areas_p$ the set of the research areas it belongs to.

The third basic set regards the set C of the countries to investigate.

4.1 Hub characterization and detection

In this section, we define 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.

Starting from this support structure, we can now define some sets regarding the neighborhoods of a node in G. Specifically, we define the neighborhood nbh_i of a node $n_i \in N$ as the set of the nodes of G directly connected with n_i:

$$nbh_i = \{n_j | (n_i, n_j, w_{ij}) \in E, n_j \neq n_i\}$$

Then, we can define the sets nbh_i^I (resp., nbh_i^F) of the neighbors of n_i belonging to the same country as (resp., to different countries from) the one of n_i:

$$nbh_i^I = \{n_j | n_j \in nbh_i, l_i = l_j\} \qquad nbh_i^F = \{n_j | n_j \in nbh_i, l_i \neq l_j\}$$

Now, we introduce the set N_k of the nodes (i.e., institutions) of a given country k:

$$N_k = \{n_i | n_i \in N, l_i = k\}$$

Another group of sets can be defined for representing several features about publications:

$$Pub_k = \{p \in Pub | \text{ at least one element of } Authors_p \text{ operates at an institution of } N_k\}$$
$$Pub_k^I = \{p \in Pub | \text{ all the elements of } Authors_p \text{ operate at institutions of } N_k\}$$

After this, we introduce Pub_{ij} as the set of publications simultaneously having researchers of both n_i and n_j as their authors. We also introduce the set $JPub$ (resp., $CPub$) as the set of papers published in a journal (resp., proceedings of a conference). We define $JCPub$ as $JCPub = JPub \cup CPub$. We also define the set Pub^q of the publications belonging to the research area q:

$$Pub^q = \{p \in Pub | q \in Areas_p\}$$

Finally, we define Pub_k^q as the subset of Pub^q having at least one author operating at an institution of the country k.

Now, we are able to introduce the concept of hub. For this purpose we need to introduce three metrics.

The first metric, M_1, is defined in such a way that, given a node n_i, M_{1_i} is equal to the sum of the weights of the edges linking n_i. Observe that this metric coincides with the classical weighted degree centrality [32, 62, 38, 37, 4]. Formally speaking:

$$M_{1_i} = \sum_{j \in nbh_i} w_{ij}$$

The second metric, M_2, is defined in such a way that, given a node n_i, M_{2_i} is the ratio of the sum of the weights of the edges linking n_i to nodes associated with foreign institutions to the average number of publications relative to the country of n_i. Observe that this metric coincides with the normalized weighted degree centrality [32, 62, 38, 37, 4]. Formally speaking:

$$M_{2_i} = \frac{\sum_{j \in nbh_i^F} w_{ij}}{AvgPub_k}$$

where $AvgPub_k = \frac{\sum_{n_i \in N_k, n_j \in N, n_i \neq n_j} w_{ij}}{|N_k|}$.

The third metric, M_3, is analogous to M_2 except that, in the numerator, the sum of the weights of the edges linking n_i to nodes of the same country is considered, since this metric takes publications with internal institutions into account. Interestingly, this metric is analogous to the E-I index [32, 47, 33, 59]. Formally speaking:

$$M_{3_i} = \frac{\sum_{j \in nbh_i^I} w_{ij}}{AvgPub_k}$$

According to both the theoretical and the experimental results described in [32, 62, 38, 37, 4, 47, 33, 59], and as verified in our case study (see Section 5.1), 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 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).

The set $\mathcal{H}^X$ of hubs is defined as:

$$\mathcal{H}^X = \{n_i \in N | n_i \in (I_1^X \cap I_2^X \cap I_3^X)\}$$

where:

$I_1^X = \{n_i \in N | M_{1_i}$ belongs to the top $X\%$ of the values of M_1, when applied to the nodes of $N\}$
$I_2^X = \{n_i \in N | M_{2_i}$ belongs to the top $X\%$ of the values of M_2, when applied to the nodes of $N\}$
$I_3^X = \{n_i \in N | M_{3_i}$ belongs to the top $X\%$ of the values of M_3, when applied to the nodes of $N\}$

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 (see Section 5.1), 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.

The rationale underlying this definition is that a hub is an institution that simultaneously belongs:

- to the top $X\%$ of the institutions publishing more papers (we call this condition C_1; it is handled by metrics M_1);
- to the top $X\%$ of the institutions publishing more papers with institutions of a country different from their own (we call this condition C_2; it is handled by metrics M_2);
- to the top $X\%$ of the institutions publishing more papers with institutions of their own country (we call this condition C_3; it is handled by metrics M_3).

It is worth pointing out that our hub definition could be seen as an attempt to introduce a new form of node centrality (specific to the context of interest), which takes into account both the number of edges relative to a node and their weights. In this sense, our hub definition follows the same general philosophy proposed in [51], where the authors present new versions of degree, closeness and betweenness centrality that take both incoming edges and their weights into consideration.

In the following, we use the symbol $\mathcal{H}_k^X$ to indicate the hubs of a given country k. The application of the parameters introduced in this section to our case study can be found in Section 5.1.

4.2 Investigation of the research scenarios for the countries of interest

In this section, we aim at analyzing the research scenarios of the countries of C in such a way as to detect their most important features and to highlight similarities and differences among them. Initially, we define I_1' as the set of the institutions of I_1 belonging to a country of C.

Now, we can introduce three indicators that could give us some knowledge about the research scenarios of the countries of C.

- The first one, RQ, is an indicator of the overall research quality in the countries of C:

$$RQ = \frac{|I_1'|}{|I_1|}$$

- The second one, FC, indicates how many institutions, among the top ones of the countries of C, publish many papers with foreign institutions:

$$FC = \frac{|I_1' \cap I_2|}{|I_1'|}$$

- The third one, TP, indicates how many institutions that publish very much with foreign institutions belong to the top institutions of the countries of C:

$$TP = \frac{|I_1' \cap I_2|}{|I_2|}$$

In the investigation of the research scenario of a country k and of the role of its hubs, it appears very interesting to study its paper distribution. For this purpose, we introduce the average number $AvgPub_k^{\mathcal{H}}$ of the publications of its hubs:

$$AvgPub_k^{\mathcal{H}} = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N, n_i \neq n_j} w_{ij}}{|\mathcal{H}_k|}$$

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:

- the average number $AvgHubPub_k^I$ of publications performed by the hubs of $\mathcal{H}_k$ with other institutions of the same country (here, the apex "I" stands for "Internal"):

$$AvgHubPub_k^I = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N_k, n_i \neq n_j} w_{ij}}{|\mathcal{H}_k|}$$

- the average number $AvgHubPub_k^F$ of publications performed by the hubs of $\mathcal{H}_k$ with other institutions of a country different from k (here, the apex "F" stands for "Foreign"):

$$AvgHubPub_k^F = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N - N_k} w_{ij}}{|\mathcal{H}_k|}$$

- the average number $AvgHubPub_k^A$ of publications performed alone by the hubs of $\mathcal{H}_k$, i.e., with authors that belong all to the same institution of the country k (here, the apex "A" stands for "Alone"):

$$AvgHubPub_k^A = \frac{\sum_{n_i \in \mathcal{H}_k w_{ij}}}{|\mathcal{H}_k|}$$

A further interesting analysis is devoted to understand if, in their cooperation with foreign institutions, the hubs of $\mathcal{H}_k$ privilege one or few countries. For this purpose, we specialize the Herfindahl Index to our research context. Specifically, in our case, we define the Herfindahl Index HI_k associated with the papers published by the hubs of $\mathcal{H}_k$ 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.

In order to apply the Herfindahl index to our context, we must introduce the following support parameters:

- number of publications that the hubs of $\mathcal{H}_k$ performed with foreign institutions: $PubH_k^F = \sum_{n_i \in \mathcal{H}_k, n_j \in N - N_k} w_{ij}$
- fraction of the external publications that the hubs of $\mathcal{H}_k$ performed with the institutions of a country q: $PubFr_{kq}^F = \frac{\sum_{n_i \in \mathcal{H}_k, n_j \in N_q} w_{ij}}{PubH_k^F}$
- set of countries having at least one paper with the institutions of a country k: $Cntr_k^F = \{q | \exists (n_i, n_j, w_{ij}) \in E, n_i \in \mathcal{H}_k, n_j \in N_q\}$

We can, now, define the Herfindahl Index associated with the papers published by the hubs of $\mathcal{H}_k$ as follows:

$$HI_k = \sum_{q=1..|Cntr_k^F|} \left(PubFr_{kq}^F \right)^2$$

The possible values of the Herfindahl Index range in the real interval $\left[\frac{1}{|Cntr_k^F|}, 1\right]$, where $\frac{1}{|Cntr_k^F|}$ is obtained when each paper is published with an institution of a different country, and 1 in the opposite case.

The application of the parameters introduced in this section to our case study can be found in Section 5.2.

4.2.1 Cooperation among hubs of the same country

In this section, we aim at investigating the cooperation levels of the hubs $\mathcal{H}_k$ 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 4.1 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 define a normalized version $\widehat{CG_k}$ of CG_k.

$\widehat{CG_k}$ is obtained by performing the same steps carried out for constructing CG_k but on a graph $\widehat{G_k}$, instead of on G_k. $\widehat{G_k}$ has the same nodes as G_k. There is an edge (n_i, n_j) in $\widehat{G_k}$ only if $\frac{Pub_{ij}}{Pub_k}$ is higher than a threshold th. We have experimentally verified that, generally, $\frac{Pub_{ij}}{Pub_k}$ follows a power law distribution. As a consequence, we have chosen to set th in such a way as to discard the 20% of the lowest values of $\frac{Pub_{ij}}{Pub_k}$.

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[1].

The application of these parameters to our case study is reported in Section 5.2.1.

4.3 Investigation about research areas

All reasonings and computations performed above for countries can be repeated for research areas. For this purpose, we define a support social network, called *RA social network*. In particular, the *RA social network* S_q, associated with the research area q, is defined as:

$$S_q = \langle N_q, E_q \rangle$$

Here, N_q is the set of the nodes of G having at least one publication belonging to Pub^q. There exists an edge $e_{ij} = (n_i, n_j, w_{ij}) \in E_q$ if there exists at least one publication of Pub^q involving one author of n_i and one author of n_j. w_{ij} indicates the number of publications of Pub^q performed by at least one author of n_i and an author of n_j.

Another important parameter, very useful in this context, is the set $\mathcal{H}_q$ of the hubs related to the research area q. Its detailed definition and the way to compute it are analogous to the corresponding ones we have described for $\mathcal{H}$ in Section 4.1.

[1]As a matter of facts, this last measure can be derived from the two other ones, but it is very expressing and, consequently, we decided to explicitly consider it.

The application of these data structures and concepts to our case study can be found in Section 5.3.

4.4 Investigation about the quality of publications

All indicators introduced above are based only on the number of publications. However, it would be important to take also their quality into account. One way to do this consists in taking 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 employ 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:

$$G' = \langle N', E' \rangle$$

There is a node $n_i \in N'$ for each institution having at least one author that 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 simultaneously performed by n_i and n_j 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 employed 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.

The application of these data structures and concepts to our case study is reported in Section 5.4.

4.5 Characterization of hub neighborhoods

A first parameter useful to characterize hub neighbors is the average number $AvgPub$ of publications of the hub neighborhoods. It is defined as:

$$AvgPub = \frac{\sum_{i \in \mathcal{H}} AvgNbhPub_i}{|\mathcal{H}|}$$

where $AvgNbhPub_i = \frac{\sum_{n_j \in nbh_i^I} \sum_{n_k \in nbh_j^I} w_{jk}}{|nbh_i^I|}$.

Since, in the hub neighborhoods, there could be other hubs, which clearly can strongly influence the neighborhood behavior, we define an additional version of hub neighborhoods $\widehat{nbh_i}$, $\widehat{nbh_i^I}$ and $\widehat{nbh_i^F}$, obtained by filtering out hubs from nbh_i, nbh_i^I and nbh_i^F, respectively. Then, we define $\widehat{AvgPub}$ by simply substituting nbh_i with $\widehat{nbh_i}$.

We call $AvgPub_k$ (resp., $\widehat{AvgPub}_k$) the "projection" of $AvgPub$ (resp., $\widehat{AvgPub}$) on the country k: $AvgPub_k = \frac{\sum_{i \in \mathcal{H}_k} AvgNbhPub_i}{|\mathcal{H}_k|}$.

A second parameter for evaluating hub neighborhoods regards their average dimension $AvgDim$:

$$AvgDim = \frac{\sum_{i \in \mathcal{H}} |nbh_i|}{|\mathcal{H}|}$$

Also 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 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 the possible number of them. It is an indicator of the cooperation level among hubs. It is defined as:

$$AvgCFrac = \frac{\sum_{i \in \mathcal{H}} NbhCFrac_i}{|\mathcal{H}|}$$

Here, $NbhCFrac_i = \frac{\widetilde{C_i}}{2^{|nbh_i|} - |nbh_i| - \frac{|nbh_i|(|nbh_i|-1)}{2}}$, where $\widetilde{C_i}$ represents the number of cliques in $NbhG_i$, whereas the denominator of $NbhCFrac_i$ indicates the possible number of cliques in nbh_i. 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:

$$AvgCNbh = \frac{\sum_{i \in \mathcal{H}} NbhCNum_i}{|\mathcal{H}|}$$

Here, $NbhCNum_i = \frac{\widetilde{C_i}}{|nbh_i|}$. Again, we call $AvgCNbh_k$ the "projection" of $AvgCNbh$ on the country k.

A final parameter measuring the cooperation level between hub neighbors is the average density $AvgDens$ of the nbh social network:

$$AvgDens = \frac{\sum_{i \in \mathcal{H}} NbhSNDens_i}{|\mathcal{H}|}$$

Here, $NbhSNDens_i = \frac{|nbhE_i|}{\frac{|nbh_i|(|nbh_i|-1)}{2}}$. As usual, we call $AvgDens_k$ the "projection" of $AvgDens$ on the country k.

The application of these parameters to our case study is reported in Section 5.5.

5 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 introduced in Section 4, consisted of the following elements:

$$C = \{\text{'A'}, \text{'E'}, \text{'M'}, \text{'T'}, \text{'O'}\}$$

where 'A' (resp., 'E', 'M', 'T', 'O') stands for 'Algeria' (resp., 'Egypt', 'Morocco', 'Tunisia', 'Others'). Clearly, 'O' does not represent a specific country, but it indicates all the ones different from the four into examination. The reasons for adding 'O' will be clear below.

Our dataset was stored in a MongoDB database [1]. 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.

5.1 Hub characterization and detection

We computed the distribution of M_1 for the publications of $JCPub$ for each year. For instance, in Figure 1, we show the distribution of M_1 for $JCPub$ in the year 2013. It is a very steep power law distribution; in other cases, the trends are less steep, but, anyhow, they follow power law distributions. We do not report the other trends for space reasons; in any case, all of them are similar to the one of Figure 1. Obtained results confirm that, in our case study, the theoretical conjecture about the trend of weighted degree centrality [32] is valid.

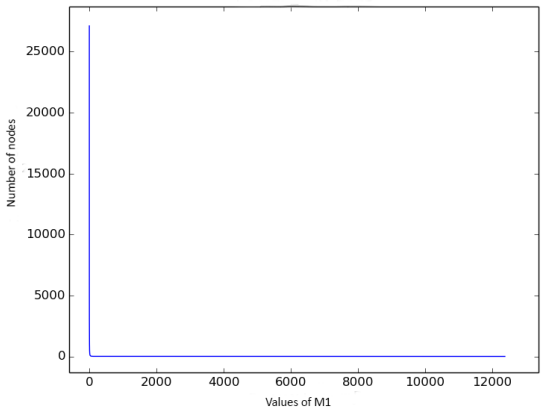

Figure 1: Distribution of M_1 for the publications of $JCPub$ in the year 2013

Then, we computed the distribution of M_2 and M_3 for the publications of $JCPub$ for each year. Analogously to what happened for M_1, all the trends were the same and followed power law distributions, thus confirming what theoretically said in [32].

In order to understand the filtering level of hubs against the increase of X, and in order to choose a default value for this parameter, we computed the number of hubs belonging to the four countries into consideration over time for the three values of X chosen in Section 4.1. We report obtained results in Figure 2. From the analysis of this figure, we can see that the trend of selected hubs is always increasing over time and very similar for the three values of X. This implies that all the three values of X would lead to the same behavior of our approach. The only difference regards the desired tradeoff between the number and the strength of the identified hubs. The higher X, the stronger (but, the less

numerous) the identified hubs. We have preferred to let our approach to be more "permissive", i.e., to let it privilege hub number on hub strength. As a consequence, we set X to a default value of 20. However, in case of hub strength needs be privileged on hub number, it would be sufficient to set X to a low value, for instance to set it to 10.

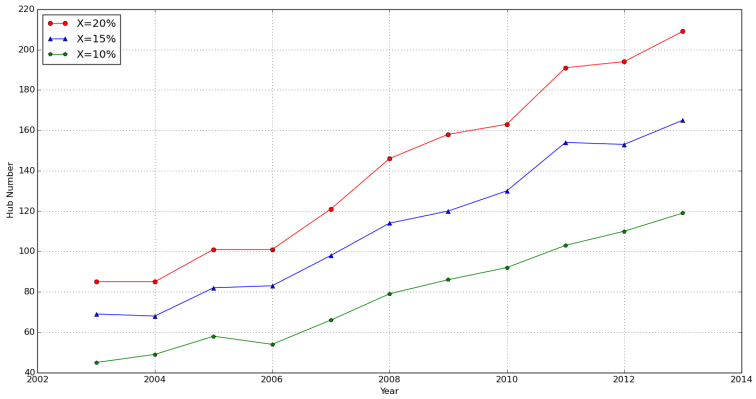

Figure 2: Hub number over time for several values of X

In many research fields, conferences are not considered in the computation of bibliometric indices. As a consequence, we judged interesting to remake all the previous investigations considering journals only. This corresponded to analyze publications belonging to $JPub$, instead of to $JCPub$. All the analyses performed for $JPub$ confirmed the general trends and the results found for $JCPub$. For instance, also in this case, M_1, M_2 and M_3 presented a power law distribution for all the four countries into consideration. Interestingly, in case of $JPub$, power law distributions are generally steeper than the ones of $JCPub$.

Another very interesting, and quite unexpected, result regards the number of hubs when only journals are considered. In fact, although the number of involved institutions decreases, the number of hubs generally does not decrease and, in several cases, increases. This quite surprisingly result can be explained by considering that the publication of a paper on conference proceedings has quite high costs (think, for instance, of costs for conference registration, travel, stay, etc.). These can disadvantage the institutions of the four countries into consideration, since all of them are characterized by a low average income per capita. If we consider $X = 10$ or $X = 15$, we obtain the same results.

An important characterization of hubs regards their capability of cooperating each other. In other words, it is interesting to verify if there exists a sort of backbone comprising hubs of different countries. To perform this investigation, we considered the concept of clique. Recall that a clique of dimension η is simply a complete subgraph consisting of η nodes. To conduct our analysis we carried out the following steps:

- We considered two time intervals. The former is $[2003, 2009]$, the latter is $[2007, 2013]$. We considered them expressly overlapped to avoid the risk of discontinuity.
- We "projected" the social network G in two social networks G' and G'' in such a way as to

Years	RQ	FC	TP
2003	0.0593	0.667	0.785
2004	0.0572	0.731	0.819
2005	0.0577	0.748	0.865
2006	0.0598	0.616	0.850
2007	0.0574	0.638	0.860
2008	0.0555	0.629	0.830
2009	0.0612	0.602	0.852
2010	0.0555	0.621	0.891
2011	0.0516	0.658	0.892
2012	0.0503	0.660	0.888
2013	0.0471	0.701	0.894

Table 1: Values of RQ, FC, and TP in the year interval [2003,2013] when both conferences and journals are considered

consider only hubs and only publications of the period $[2003, 2009]$ in the former, and of the period $[2007, 2013]$ in the latter.

- We computed all the cliques of G' and G''.

After this, we analyzed the number and the dimension of obtained cliques, as well as the institutions belonging to them. As a general trend, we found that there are many cliques and most of them are very small. This indicates that there are some contacts among hubs but there is not a strict cooperation among many of them in such a way as to have "research backbones".

Furthermore, the largest clique of the period $[2003, 2009]$ consisted of 13 hubs, whereas the largest one of the period $[2007, 2013]$ was formed by 17 hubs. In both cases all hubs forming these cliques are only Egyptians. From this analysis, we can draw the following knowledge patterns:

- Cliques tend to enlarge over time, although slowly. For instance, the largest clique of the period $[2007, 2013]$ is obtained by aggregating four further hubs to those belonging to the largest clique of the period $[2003, 2009]$.
- The largest cliques are formed by hubs of the same country; for instance, the top 5 cliques in the two periods are all formed by Egyptian hubs only. This last result has a further important consequence in that it shows that hubs of different countries tend to not cooperate each other.

5.2 Investigation of the research scenarios for the countries of interest

First of all, we computed the three indicators RQ, FC and TP, whose formalization has been provided in Section 4.2.1, for the four North African countries of interest. This computation (for both $JCPub$ and $JPub$) returned very interesting knowledge patterns about the research scenarios in the four countries (see Table 1). In particular, the first indicator shows that the research institutions in the four countries do not present excellent performances. Furthermore, this indicator does not show a significant increase over time. The second and the third indicators highlight that an institution of one of the four countries benefits very much from the cooperation with foreign institutions for reaching and maintaining a high performance in its own country.

After these analyses, we started to investigate the similarities and the differences for hubs in the four countries. First, we computed the values of M_1, M_2 and M_3 in the four countries for all the

years into consideration. We obtained that both M_1 an M_3 present a power law distribution and, therefore, confirm what we have seen for the general case. An interesting trend is shown by M_2 for these countries (Figure 3). Indeed, this measure presents a distribution characterized by a broken line with quite a rapid decrease and a possible starting peak. This suggests a very interesting scenario for the hubs in each of the four countries. This scenario is the typical one of an oligarchy of hubs for each country and is very different from the two ones we had initially hypothesized (i.e., a lot of quite weak hubs, corresponding to a smoothly decreasing distribution for M_2, or a very few number of very strong hubs, corresponding to a power law distribution for M_2).

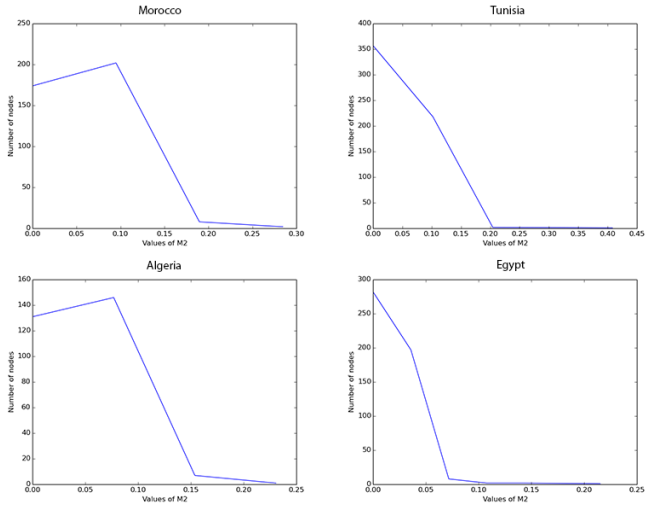

Figure 3: Trend of M_2 for the four countries in the year 2013

In Figure 4, 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 5, we report the values of the average number $AvgPub_k^{\mathcal{H}}$ of hub publications over time (see Section 4.2) for the four countries of interest. From the analysis of this figure we can see that Egyptian hubs generally publish much more papers than the hubs of the other countries. This result, along with the ones of Figure 4, suggests that research in Egypt is much more concentrated than in the other three countries.

A final report about this issue regards the total number of publications $|Pub_k|$ (see Section 4.1) over time for each country. Obtained results evidence that that Egypt has a number of publications much higher than the other three countries. This result, along with the ones reported in Figure 4, is a further confirmation that research in Egypt is much more concentrated than in the other three countries.

After this, we computed the average number of internal, external and alone publications for the four countries of interest. Obtained results evidence that the hubs of all countries always publish more with foreign institutions than with internal ones. Interestingly, Egyptian hubs have a significant fraction of alone publications.

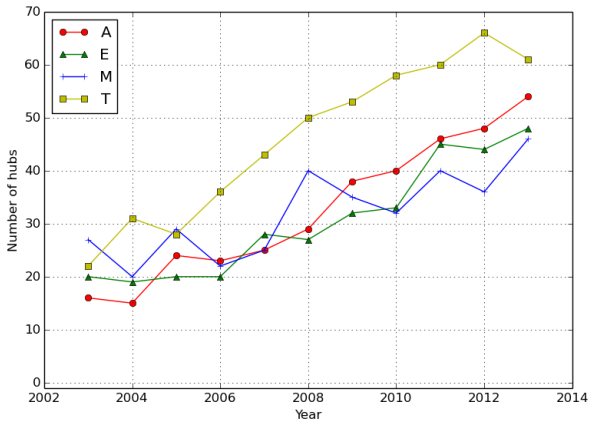

Figure 4: Number of hubs for each country in the year interval [2003,2013]

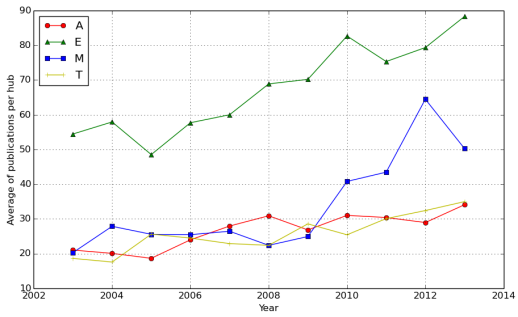

Figure 5: Average number of publications per hub over time for the four countries

In Figure 6, we report the Herfindahl index HI_k for the four countries (see Section 4.2). 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. 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.

A possible objection to the previous way of proceeding could be that the computation of the Herfindahl index of a country k (e.g., Egypt) could be "biased" by the presence of many institutions of different countries each having only one publication with a hub of k. To overcome this objection, for each country k, we considered the top 5 countries T_k sharing publications with its hubs. Then, we recomputed the Herfindahl index considering, for each k, only the institutions belonging to the countries of T_k. Obtained results show that all the main conclusions we have drawn from Figure 6 are still valid. This not only overcomes the previous objection, but it is also a further confirmation of the power law distribution of hubs' publications, which we have detected by studying the trend of M_1.

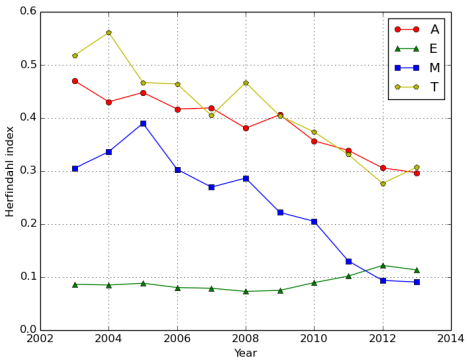

Figure 6: Herfindahl index over time for the four countries

5.2.1 Cooperation among hubs of the same country

To determine the cooperation levels among hubs for the four North African countries into consideration, for each country k, we performed the following tasks:

- We considered the two time intervals $[2003, 2009]$ and $[2007, 2013]$.
- We computed the clique social networks (see Section 4.2.1) $CG1_k$ (resp., $CG2_k$), corresponding to the first (resp., the second) time interval.
- We measured the four parameters introduced in Section 4.2.1 for quantitatively evaluating clique social networks.

Obtained results are reported in Table 2. From the analysis of these tables we can draw the following conclusions:

- Egypt has the largest clique in both periods; the clique is much larger than the maximum cliques of the other countries;
- 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 7, 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 little nodes, allows us to derive another important knowledge pattern, i.e., that research cooperation in Egypt is more advanced than in the other countries.

In Figures 8 and 9, we report the graphs $\widehat{CG1_k}$ and $\widehat{CG2_k}$ (corresponding to $\widehat{CG_k}$ for the first and the second time interval) for the four countries. From the analysis of these figures we can observe that

Country	$\|\mathcal{C}1_k\|$	$d1_{\mathcal{C}_k}$	$\frac{d1_{\mathcal{C}_k}}{\|\mathcal{H}_k\|}$	$f1^{\mathcal{H}}_{\mathcal{C}_k}$
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

Country	$\|\mathcal{C}2_k\|$	$d2_{\mathcal{C}_k}$	$\frac{d2_{\mathcal{C}_k}}{\|\mathcal{H}_k\|}$	$f2^{\mathcal{H}}_{\mathcal{C}_k}$
Algeria	234	8	0.121	0.939
Egypt	94	17	0.27	1.0
Tunisia	304	9	0.081	0.847
Morocco	106	9	0.134	0.821

Table 2: Quantitative differences characterizing the cooperation behaviors of hubs in the four countries (first time interval on the top and second time interval on the bottom)

Country	*number of nodes*	*number of edges*	*density*
Algeria	41	166	0.200
Egypt	34	196	0.349
Tunisia	66	272	0.127
Morocco	45	174	0.176

Country	*number of nodes*	*number of edges*	*density*
Algeria	62	249	0.132
Egypt	60	450	0.254
Tunisia	103	416	0.079
Morocco	62	221	0.117

Table 3: Number of nodes, number of edges and density of $CG1_k$ (on the top) and of $CG2_k$ (on the bottom) for all countries

the different behavior of Egyptian hubs with respect to the ones of the other countries is confirmed, although slightly attenuated.

Finally, we computed the number of nodes, the number of edges and the density of $CG1_k$ and $CG2_k$ for all countries. Obtained results are reported in Table 3. From the analysis of this table we can observe that the three measures quantitatively depict very well what we have expressed previously. Furthermore, if we compare their values in the two periods, we can draw some interesting knowledge patterns. In fact, we can observe that the number of nodes always increases, which implies an increase of the hub capability of cooperating each other. This increase is quite high (i.e., about 38%) for Morocco, high (i.e., about 50%) for Algeria and Tunisia, and very high (i.e., about 76%) for Egypt. The same trends can be observed for the increase of the number of edges (i.e., about 27% for Morocco, about 50% for Algeria and Tunisia, and about 129% for Egypt). By contrast density always decreases. These last results represent a further confirmation about the fact that hubs continue to cooperate a little each other.

Finally, if we consider the ratio of the increase of the number of edges to the increase of the number of nodes for the four countries when passing from the first to the second time interval, we can observe that this ratio is about 1 for Algeria and Tunisia, about 1.70 for Egypt and about 0.73 for Morocco. This indicates that, in the second time interval, Egypt had a spectacular increase of the hub cooperation. This also reflects in the density decrease, which is much more reduced in Egypt than in

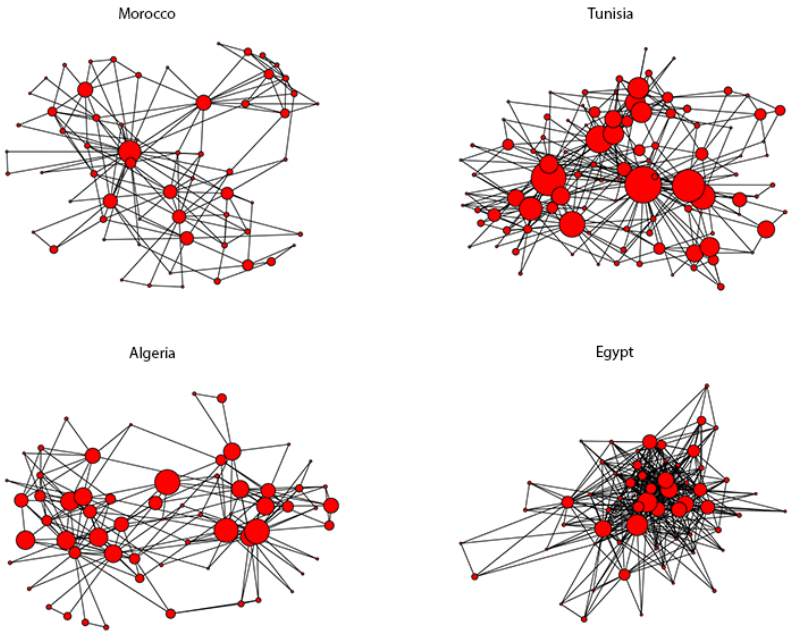

Figure 7: Graphs $CG2_k$ for the four countries

the other countries (in fact, it is about -27% for Egypt, -34% for Algeria, -38% for Tunisia and -59% for Morocco). As for density, its decrease must not be misleading since, to avoid it, the number of edges should have increased against the square of the number of nodes, which is almost impossible. As a matter of fact, the number of edges always increases in all the four countries, but slightly.

5.3 Investigation about research areas

For each research area of RA (see Section 4), we computed the corresponding RA network and, then, we repeated all the tasks described in the previous sections in such a way as to disaggregate the corresponding results per research area.

A first analysis regarded the distribution of M_1, M_2 and M_3 over time for each research area. In this case, we obtained that these distributions are analogous to the ones obtained for aggregated data.

For each research area, we computed the number of publications of hubs over time. Obtained results show that the research areas having the highest number of hubs are 'NS', 'ET' and 'MH'. This result confirms the ones reported in [39] concerning the diffusion of research areas in the same countries.

Then, we computed the number of publications per hub over time for each research area. Obtained results are in line with the ones shown in the previous figures.

A further, very interesting, disaggregation of results is obtained by separating data for pairs (country, research area). In fact, in this way, we can verify if the four countries into consideration present similar or dissimilar features and behaviors in the different research areas.

A first analysis of this disaggregation level regarded the distribution of M_1, M_2 and M_3. Obtained

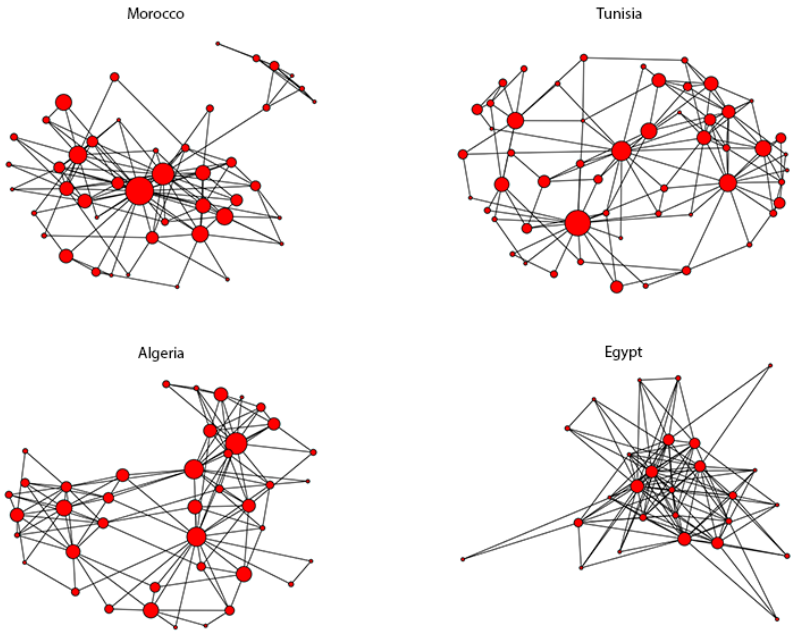

Figure 8: Graphs $\widehat{CG1_k}$ for the four countries

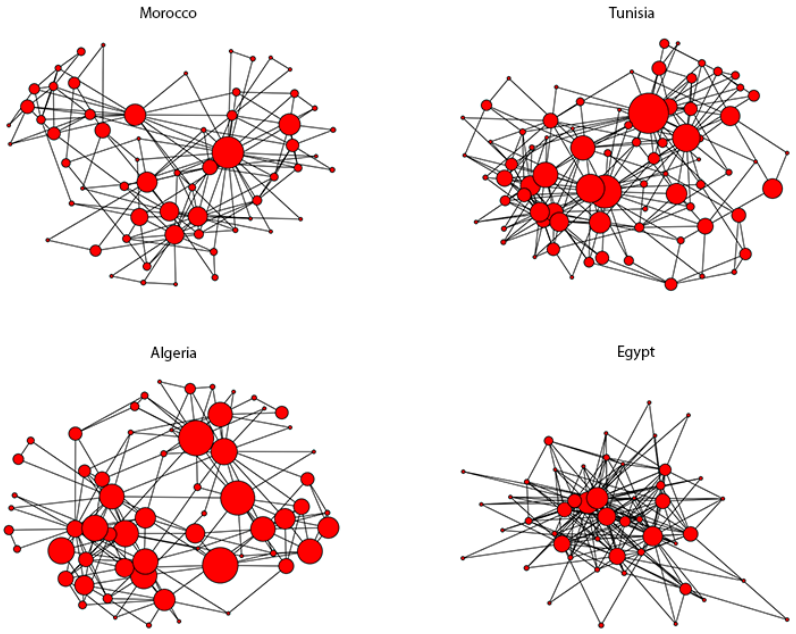

Figure 9: Graphs $\widehat{CG2_k}$ for the four countries

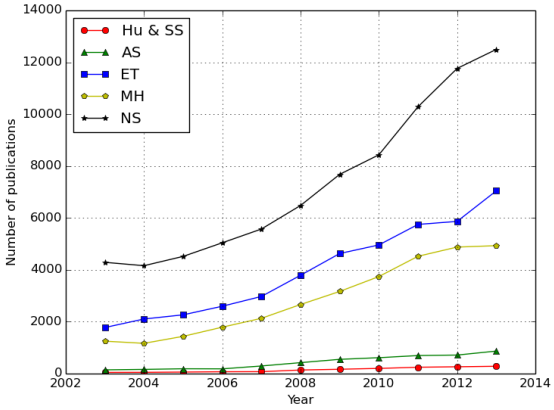

Figure 10: Average number of publications of hubs over time for each research area

results confirm that these metrics follow a power law distribution for 'NS', 'ET' and 'MH'. For the other three research areas, the number of publications performed in the four countries was small and, therefore, we had to discard obtained results, because they were not reliable.

A second investigation at the same disaggregation level concerned the number of hubs over time. Obtained results confirm in principle the ones about the distribution of hubs per country, shown in Figure 2.

We have seen that a particular feature of hubs was the fact that they published more with foreign institutions than with internal ones (see Section 4.2). We repeated this investigation at the new disaggregation level and found that this trend is always valid except for the pair (Tunisia, 'MH'), where it is never valid, and for the pair (Egypt, 'ET'), where it is valid only for the time interval $[2009, 2013]$. Interestingly, for the pair (Egypt, 'ET'), the number of alone publications is higher than the number of internal and external ones in the time interval $[2003, 2010]$.

After this, we investigated the Herfindahl index. Obtained results generally confirm the corresponding aggregated ones (see Figure 6), although with some slight differences. In particular, the trend of 'NS' is identical. As for 'MH', differently from the aggregated case, Morocco always shows a behavior similar to the ones of Algeria and Tunisia. An intermediate behavior w.r.t. the ones of 'NS' and 'MH' is obtained for 'ET'.

The next investigation regarded backbones and cliques, clearly at the new disaggregation level. In this case, after having constructed the corresponding clique social networks, we found that the general trend (i.e., the aggregated results) about country backbones and hub cooperation in the different countries are confirmed in almost all research areas. However, there are a couple of interesting situations and/or exceptions. In fact, we can observe: *(i)* a very few number of hubs and cliques, along with a scarce cooperation, in Algeria for 'MH' and in Morocco for 'ET' in the time interval [2003,2009], and in Morocco for 'MH' in the time interval [2007,2013]; *(ii)* the presence of two disconnected components in Algeria for 'MH' in the time interval [2007,2013] and in Morocco for 'MH' in the time interval [2003,2009]; *(iii)* a much more scarce cooperation for 'MH' than for 'NS' and 'ET' in all the countries and in both time intervals.

Analogously to the aggregated case, also at this disaggregation level we decided to construct the normalized RA social networks. To perform this task, we preliminarily had to verify that weight distribution in the edges followed a power law (see Section 4.3). After this, we constructed the normalized clique social networks and analyzed them. From this analysis we found that the general trends detected for aggregated data are still valid, although the following specificities/exceptions were observed: *(i)* there are few hubs and cliques and a scarce cooperation in Algeria for 'MH' in the time interval [2003,2009]; *(ii)* there is the presence of two disconnected components in Morocco for 'MH' in the time interval [2003,2009], and the presence of three disconnected components in Algeria for 'MH' in the time interval [2007,2013].

5.4 Investigation about the quality of publications

As for this issue, after having constructed G' (see Section 4.4), we computed the distributions of the metrics M_1, M_2 and M_3. We found that these distributions are analogous to the previous ones.

After this, we determined the number of hubs in the four countries of interest. In this case, we observed that, when considering the impact factor, the number of hubs generally decreases w.r.t. the case in which we consider only the number of publications (see Table 4). This is another indicator of the fact that the research performance for the four countries is low. In any case, we found that this number is always increasing over time.

Years	*Number of hubs* (without impact factors and citation numbers)	*Number of hubs* (with impact factors)	*Number of hubs* (with citation numbers)
2003	90	71	58
2004	85	70	65
2005	97	79	69
2006	100	95	82
2007	124	101	103
2008	140	118	111
2009	155	132	120
2010	165	142	123
2011	190	152	140
2012	199	157	127
2013	202	175	147

Table 4: Hub number over time in the three different situations into examination

As a final analysis, we computed the metrics M_1, M_2 and M_3 for each research area. We found that all the previous power law trends, obtained without considering impact factors, are fully confirmed and, even, reinforced, since the new power law distributions are steeper.

After the analyses based on impact factors, we made different analyses taking the citation number into account. In this case, we repeated all the computations already made for impact factors. In particular, we computed: *(i)* the distributions of the metrics M_1, M_2 and M_3; *(ii)* the number of hubs in the four countries (see Table 4); *(iii)* the distributions of the metrics M_1, M_2 and M_3 for each research area. All the results obtained in these computations totally confirm the ones seen for impact factors. Only the distributions of the metrics M_1, M_2 and M_3 present some noise near the elbow of the corresponding curves.

In our opinion, the fact that the previous results are confirmed in this case is extremely important, because this is an indicator of their stability; indeed, although we consider two totally different quality

factors, the obtained results are always the same.

5.5 Characterization of hub neighborhoods

The first task we carried out for characterizing hub neighborhoods was the computation of the average number $AvgPub$ (see Section 4.5) of the publications of hub neighborhoods. As in Section 5.2, we distinguished among internal, external and alone publications. Obtained result is reported in Figure 11.

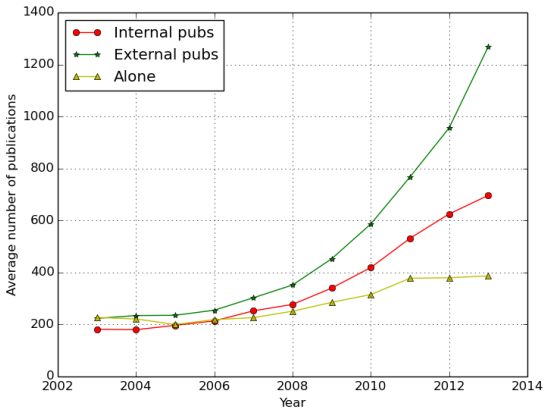

Figure 11: Average number of internal, external and alone publications for hub neighborhoods

This result was quite unexpected. In fact, in Section 5.2, we have seen that hubs tend to publish more with foreign institutions than with internal ones. However, in order to "fulfill its mission" to be a guide for its country, a hub must maintain a strict contact with internal institutions. So, we had hypothesized that this task was performed through its directed neighbors. Nevertheless, this graph seems to contradict this hypothesis.

As a matter of facts, a deeper investigation allows us to better understand this phenomenon. In fact, we must recall that, in the hub neighborhoods, there could be other hubs, which clearly can strongly influence the neighborhood behavior. As a consequence, it appears more correct to consider the trend of $\widehat{AvgPub}$, instead of $AvgPub$ (see Section 4.5), over time.

We carried out this last task by distinguishing among internal, external and alone publications. Obtained results, reported in Figure 12, fully confirm our hypothesis. In fact, in this case, the average number of internal publications is higher than the average number of external ones. Interestingly, the average number of alone publications is significant, at least from 2003 to 2010.

To verify if the four countries showed identical or different behaviors in this analysis, we disaggregated data per country and considered $AvgPub_k$ and $\widehat{AvgPub}_k$. We obtained that the trends described above for aggregated data are always confirmed for each country. We also observed an enormous decrease of the average number of publications when we consider the neighborhoods without hubs. This is a further confirmation that the distribution of the publications among institutions follows a power law. Finally, as usual, the number of alone publications performed by Egyptian hubs is quite significant.

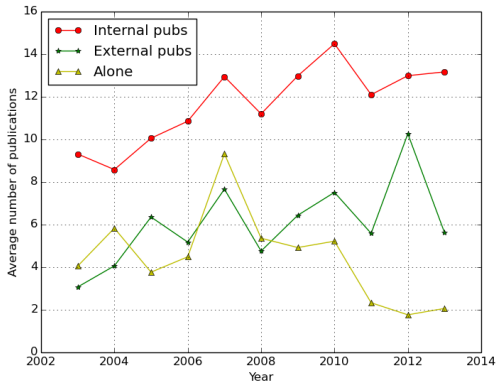

Figure 12: Average number of internal, external and alone publications for hub neighborhoods (after the hubs present therein have been filtered out)

A second investigation about neighborhoods regarded their average dimension. For this purpose, we computed $AvgDim$ over time. Obtained results are reported in Figure 13. From the analysis of this figure, we can observe that the average dimension of hub neighborhoods always increases. This implies that the number of institutions cooperating with hubs is increasing over time constantly. As usual, we disaggregated these data per country. Specifically, we computed $AvgDim_k$ over time for the four countries. Obtained results evidence that the average dimension of neighborhoods increases in all countries, although with some irregularities.

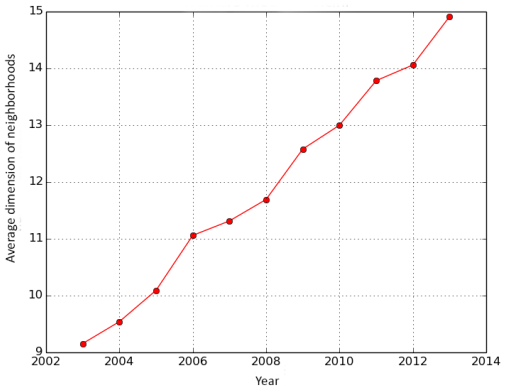

Figure 13: Values of $AvgDim$ over time

After this, we investigated the cooperation level among the institutions belonging to hub neighborhoods. We started by computing $AvgCFrac$ over time. Obtained results are reported in Figure 14. Their analysis shows that $AvgCFrac$ tends to increase over time, although with some irregularities. This implies an increase over time of the cooperation among institutions belonging to hub neighborhoods. We also disaggregated data per country. For this purpose, we computed $AvgCFrac_k$ for the four countries. Obtained results evidence that the four countries show very different behaviors.

Figure 14: Values of $AvgCFrac$ over time

A second measure about intra-neighborhood cooperation is $AvgCNbh$. We computed it over time. Obtained results are reported in Figure 15. From the analysis of this figure, we can observe that this parameter is significantly increasing over time, which is a further confirmation that cooperation among hub neighbors is increasing. In fact, its increase implies an increase of $NbhCNum_i$ and, since we have seen that $|nbh_i|$ increases over time, this implies a higher increase of the number of cliques. We disaggregated data per country. For this purpose, we computed $AvgCNbh_k$ for the four countries. Obtained results show that, in this case, the four countries present very different behaviors.

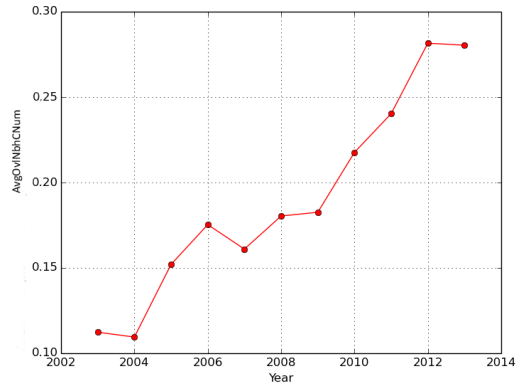

Figure 15: Values of $AvgCNbh$ over time

The final measure about intra-neighborhood cooperation regarded the average density $AvgDens$ of the nbh social networks. We computed this parameter over time and we reported obtained results in Figure 16. From the analysis of this figure, we can observe an increase of this parameter. This is a further confirmation, obtained via a different fashion (based on edge number, instead of on clique number), that hub neighbors tend to increase their cooperation over time. We disaggregated data per country and we computed $AvgDens_k$ over time for the four countries. Obtained results show that this parameter is significantly increasing over time for all countries.

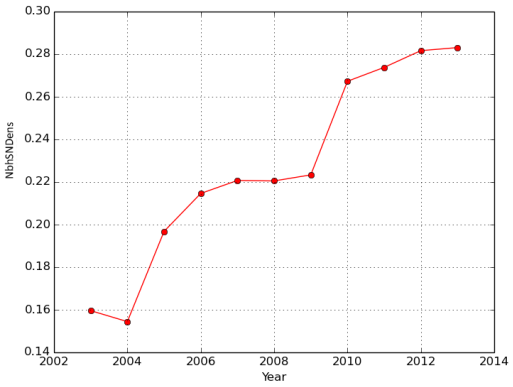

Figure 16: Values of $AvgDens$ over time

6 Discussion

In the Introduction, we have outlined the main innovations of our approach. Furthermore, in Section 2, we have examined related literature. Now, after having examined our approach in all details, and after having seen its behavior on a real case study, we can provide a more detailed presentation of its main features and novelties w.r.t. the previous ones.

First, differently from most of the previous approaches described in Section 2, which focus on authors, our approach is centered on institutions.

One of the specific goals of our approach, i.e. hub detection and characterization, is novel in the literature. As matter of fact, to the best of our knowledge, the only paper investigating hubs is [11]. However, in [11], the definition of hubs is centered on authors and centrality measures, and is much simpler than the one we adopted in this paper.

Our approach also aims at investigating the similarities and the differences of the research scenarios in a set of countries of interest. This is another contribution provided by it, which is generally not found in the previous approaches proposed in the past.

Also the techniques employed to carry out investigations are very different from the ones adopted in the past. In fact, past researches in this field were centered on the concept of centrality, whereas our approach employs more specific and ad-hoc data structures and parameters.

Furthermore, to better evaluate cooperation among involved institutions, we have employed the concept of clique and we have defined the clique social network, i.e., a specific support social network in which the dimension of a node is directly proportional to its tendence of cooperating with the other ones. We have also introduced some metrics to quantitatively evaluate the difference between two or more clique social networks.

Moreover, we have carried out a deep study of hub neighbors by introducing several metrics for quantitatively analyzing and comparing them.

As a further specificity of our approach, we have deepened our investigation about its main features, as well as about the similarities and the differences of research scenarios, by disaggregating data not only per country but also per research area and per pairs (country, research area).

In our evaluations, we considered not only the number of publications but also the corresponding quality by taking both their impact factor and their citation number into account.

Last, but not the least, we provided a re-definition of the Herfindahl index (largely used in the past research in Biology and Economics) to measure how much, in a given country, research activities are guided by few hubs or distributed among many institutions.

Another interesting issue could regard the comparison of our approach with some commercial systems, like Elsevier Pure, Elsevier Fingerprint Engine and Elsevier Scopus.

Elsevier Pure supports an institution in the definition of the optimal research and cooperation strategies, in assessment activities and in making business decisions. Pure aggregates information regarding the research activities of a given institution stored in different, both internal and external, sources. Furthermore, it ensures that data guiding strategic decisions is trusted, comprehensive and accessible in real time. It has an underlying centralized system, which is very versatile and supports the construction of reports, the evaluation of performances, the management of researchers' profile, the construction and the maintenance of research networks, the expertise detection, etc. Pure can be integrated with Elsevier Fingerprint Engine for stimulating the cooperation among researchers.

Elsevier Fingerprint Engine mines scientific documents ranging paper abstracts, funding announcements and awards, project summaries, patents, proposals/applications, etc., to create an index of weighted terms called Fingerprint visualization. The construction of Fingerprints is made through Natural Language Processing techniques and through the support of suitable thesauri. By aggregating and comparing the Fingerprints of people, publications, funding opportunities and ideas, Elsevier Fingerprint Engine mines metadata to detect connections among people, publications, funding opportunities and ideas. The thesauri adopted by Elsevier Fingerprint Engine make this last tool well suited in life science, engineering, earth and environmental sciences, arts and humanities, social sciences, mathematics and agriculture. Elsevier Fingerprint Engine can be integrated with Pure, to create expertise profiles aiming at helping cooperation, with Expert Lookup, to identify referees and potential conflicts of interest, and with Elsevier Journal Finder, to find the journal most suited to publish a given article.

Elsevier Scopus is the greatest database of abstracts and citations of scientific literature. It encompasses scientific journals, books and conference proceedings. Scopus supplies several functionalities. In particular, Scopus supports the search of documents, authors, affiliations and several forms of advanced search. It also allows the definition of alerts regarding search, documents and authors, the browsing of resources, the creation of personalized lists of documents, the export of data to reference managers, the discovery of the documents citing selected articles, the visualization of the list of references included in an article, the analysis of search results, the comparison of journals, the quick visualization of the citation impact and the scholarly community engagement for an article, the analysis of the citation trend for an article, the analysis and tracking of an individual's citation history including total citations and document count, the computation of the h-index of an individual. Finally, Scopus has a comprehensive suite of metrics to facilitate evaluations and provide a better view of research interests.

Differently from Scopus, which bases most of its features on article citations, our approach is based on co-authorships. Furthermore, Scopus is more focused on single authors or single institutions, whereas our approach focus mainly on cooperations among authors or institutions. In its main

objectives, our approach is more similar to Pure than to Scopus. However, Pure finds cooperation and network information based on text analysis performed by Fingerprint Engine. By contrast, as previously pointed out, our approach is based on co-authorship information.

Furthermore, our approach introduces the concept of hub, which is fundamental to help innovation managers in their decision making activities. This concept is not directly present in Pure and Fingerprint Engine. Only after several computations of the information directly provided by these systems, followed by a strong human intervention, it could be possible to derive (at least partial) information on hubs.

Analogously to Scopus, also our approach introduces several metrics to evaluate the level of the research activities in a scenario of interest, although the metrics used by the two systems take different pieces of information into account.

Once hubs have been detected, our approach allows a deep analysis of their main features and their relationships. For instance, it can indicate if there is a strong cooperation among the hubs of a given country or among the hubs (possibly of different countries) that operate in a given research area. Furthermore, it can investigate the characteristics of the hub neighbors and how they can be influenced by the hub themselves for the different countries and research areas of interest. Interestingly, our approach can incorporate in its metrics also citation counts and impact factors (i.e., the main parameters used by Scopus) to obtain more refined results.

7 Conclusion

In this paper, we have proposed a new SNA-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 promoting innovation in the countries of interest. Our approach is based on the publications performed by all the research institutions of the countries of interest, as registered in the Web of Science repository. We applied it to four North African countries (i.e., Algeria, Egypt, Morocco and Tunisia). Furthermore, we considered several related scientific approaches and three commercial systems, and specified the analogies, the differences, the strenghts and the weaknesses of our approach w.r.t. them.

This paper is certainly an important end point but, at the same time, it can be also considered a starting point for future research efforts. In particular, we plan to employ analysis techniques about 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 this latter. Moreover, we plan to investigate the possible application of classification techniques to derive hub profile in different countries. Furthermore, 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. Finally, we plan to extend our approach in such a way that it may handle related data sources, such as the ones concerning patents. This would allow us: *(i)* to extract knowledge patterns about patent inventors and their cooperation; *(ii)* to verify the presence of "power inventors" in a country; *(iii)* to verify the existence of a backbone and of possible cliques among them.

Acknowledgments

The authors thank the research center I-CRIOS (the Invernizzi Center for Research on Innovation, Organization and Strategy) of Univertità Bocconi that provided data for their analysis. They also thank Prof. Franco Malerba, Prof. Roberto Mavilia and Prof. Fabio Landini, who helped them very much to understand the innovation management aspects and the implications of the knowledge patterns found by this research. This work was partially supported by Aubay Italia S.p.A.

References

[1] Mongodb. `https://www.mongodb.org/`, 2016.

[2] Python. `https://www.python.org/`, 2016.

[3] Web Of Science. `http://wokinfo.com/`, 2017.

[4] A. Abbasi. h-Type hybrid centrality measures for weighted networks. *Scientometrics*, 96(2):633–640, 2013. Springer.

[5] A. Abbasi, J. Altmann, and L. Hossain. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5 (4):594–607, 2011.

[6] 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.

[7] F. J. Acedo, C. Barroso, C. Casanueva, and J. L. Galan. Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, 43 (5):957–983, 2006.

[8] J. Adams, K. Gurney, D. Hook, and L. Leydesdorff. International collaboration clusters in Africa. *Scientometrics*, 98(1):547–556, 2014. Springer.

[9] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. APS.

[10] T. Alnuaimi, J. Singh, and G. George. Not with my own: long-term effects of cross-country collaboration on subsidiary innovation in emerging economies versus advanced economies. *Journal of Economic Geography*, 12(5):943–968, 2012. Oxford University Press.

[11] 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.

[12] C. Autant-Bernard, S. Chalaye, E. Gagliardini, and S. Usai. European Knowledge Neighbourhood: Knowledge Production in EU Neighbouring Countries and Intensity of the Relationship with EU Countries. *Tijdschrift voor economische en sociale geografie*, 108(1):52–75, 2017. Wiley Online Library.

[13] 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.

[14] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. American Association for the Advancement of Science.

[15] A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002. Elsevier.

[16] 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.

[17] 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.

[18] L. Branstetter, G. Li, and F. Veloso. The rise of international co-invention. *The Changing Frontier: Rethinking Science and Innovation Policy*, pages 135–168, 2015. University of Chicago Press.

[19] S. Breschi and F. Lissoni. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4):439–468, 2009. Oxford University Press.

[20] Y. Chen, C. Ding, J. Hu, R. Chen, P. Hui, and X. Fu. Building and analyzing a global co-authorship network using google scholar data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1219–1224, Perth, Australia, 2017. International World Wide Web Conferences Steering Committee.

[21] Z. Chinchilla-Rodriguez, A. Ferligoj, S. Miguel, L. Kronegger, and F. de Moya-Anegón. Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics*, 93 (3):699–717, 2012.

[22] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review Part E*, 70(6):066111, 2004.

[23] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[24] R. De. Optimal conditions for innovation: Firm-level evidence from Kenya and Uganda. In *Proc. of the Annual African Economic Conference*, Addis Abeba, Ethiopia, 2014.

[25] T. Dehdarirad and S. Nasini. Research impact in co-authorship networks: a two-mode analysis. *Journal of Informetrics*, 11(2):371–388, 2017. Elsevier.

[26] B.D.P.F. e Fonseca, R. Sampaio, M.V. de Araujo Fonseca, and F. Zicker. Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems*, 14(1):34, 2016. BioMed Central.

[27] J.L. Furman, M.K. Kyle, I.M. Cockburn, and R. Henderson. Public & private spillovers, location and the productivity of pharmaceutical research. Technical report, National Bureau of Economic Research, 2006.

[28] E. Giuliani, A. Martinelli, and R. Rabellotti. Is Co-Invention Expediting Technological Catch Up? A Study of Collaboration between Emerging Country Firms and EU inventors. *World Development*, 77:192–205, 2016. Elsevier.

[29] M. Goedhuys. Learning, product innovation, and firm heterogeneity in developing countries; Evidence from Tanzania. *Industrial and Corporate Change*, 16(2):269–292, 2007. Oxford University Press.

[30] R.M. Grant. Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, 7(4):375–387, 1996. INFORMS.

[31] Z. Griliches. Introduction to "output measurement in the service sectors". In *Output measurement in the service sectors*, pages 1–22. 1992. University of Chicago Press.

[32] R. Hanneman and M. Riddle. *Introduction to social network methods.* `http://faculty.ucr.edu/~hanneman/nettext/` , 2005. University of California, Riverside.

[33] V.A. Hill. Collaboration in an Academic Setting: Does the Network Structure Matter? *Center for the Computational Analysis of Social and Organizational Systems (CASOS) technical report*, 2008.

[34] S. Kazi, Q. Rajput, and S. Khoja. Study of Evolving Co-Authorship Network: Identification of Growth Patterns of Collaboration Using SNA Measures. In *IEEE 11th International Conference on Semantic Computing*, pages 488–493, San Diego, CA, USA, 2017. IEEE.

[35] J. Kim, S. J. Lee, and G. Marschke. International knowledge flows: evidence from an inventor-firm matched data set. In *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, pages 321–348. 2009. University of Chicago Press.

[36] J. Kim and C. Perez. Co-authorship network analysis in industrial ecology research community. *Journal of Industrial Ecology*, 19 (2):222–235, 2015.

[37] C. Kiss and M. Bichler. Identification of influencersmeasuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008. Elsevier.

[38] H. Kretschmer and T. Kretschmer. A new centrality measure for social network analysis applicable to bibliometric and webometric data. *Collnet Journal of Scientometrics and Information Management*, 1(1):1–7, 2007. Taylor & Francis.

[39] 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.

[40] E.A. Leicht, P. Holme, and M. E. J. Newman. Vertex similarity in networks. *Physical Review Part E*, 73(2):026120, 2006.

[41] D.A. Levinthal and J.G. March. The myopia of learning. *Strategic Management Journal*, 14(S2):95–112, 1993. Wiley Online Library.

[42] F. Lissoni and E. Miguelez. Patents, Innovation and Economic Geography. Technical report, Groupe de Recherche en Economie Théorique et Appliquée, 2014.

[43] P. Liu and H. Xia. Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, 103 (1):101–134, 2015.

[44] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information processing & management*, 41 (6):1462–1480, 2005.

[45] L.M. Lubango. The effect of co-inventors' reputation and network ties on the diffusion of scientific and technical knowledge from academia to industry in South Africa. *World Patent Information*, 43:5–11, 2015. Elsevier.

[46] J.G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991. INFORMS.

[47] C. McGrath and D. Krackhardt. Network conditions for organizational change. *The Journal of Applied Behavioral Science*, 39(3):324–336, 2003. Sage Publications.

[48] E. Miguélez and R. Moreno. Research networks and inventors' mobility as drivers of innovation: evidence from Europe. *Regional Studies*, 47(10):1668–1685, 2013. Taylor & Francis.

[49] 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.

[50] M. Newman and E.A. Leicht. Mixture models and exploratory analysis in networks. *Proc. of the National Academy of Sciences of the United States of America*, 104:9564–9, 2007.

[51] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010. Elsevier.

[52] B. Oyelaran-Oyeyinka, G.O.A. Laditan, and A.O. Esubiyi. Industrial innovation in Sub-Saharan Africa: the manufacturing sector in Nigeria. *Research Policy*, 25(7):1081–1096, 1996. Elsevier.

[53] M. Pavlov and R. Ichise. Finding Experts by Link Prediction in Co-authorship Networks. In *Proc. of the International Workshop on Finding Experts on the Web with Semantics (FEWS 2007)*, pages 42–55, Busan, Korea, 2007.

[54] P.J.A. Robson, H.M. Haugh, and B.A.Obeng. Entrepreneurship and innovation in Ghana: enterprising Africa. *Small Business Economics*, 32(3):331–350, 2009. Springer.

[55] G.A. Ronda-Pupo and L.A. Guerras-Martin. Collaboration network of knowledge creation and dissemination on Management research: ranking the leading institutions. *Scientometrics*, 107(3):917–939, 2016. Springer.

[56] G. Rooks, L. Oerlemans, A. Buys, and T. Pretorius. Industrial innovation in South Africa: A comparative study. *South African Journal of Science*, 101(3-4):149–150, 2005. Open Journals Publishing.

[57] J. Singh. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5):756–770, 2005. INFORMS.

[58] J. Singh. Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96, 2008. Elsevier.

[59] M. Tortoriello and D. Krackhardt. Activating cross-boundary knowledge: The role of Simmelian ties in the generation of innovations. *Academy of Management Journal*, 53(1):167–181, 2010. Academy of Management.

[60] C.S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608–1618, 2005. Elsevier.

[61] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. 1994. Cambridge University Press.

[62] D. Wei, X. Deng, X. Zhang, Y. Deng, and S. Mahadevan. Identifying influential nodes in weighted networks based on evidence theory. *Physica A: Statistical Mechanics and its Applications*, 392(10):2564–2575, 2013. Elsevier.

[63] N. Zamzami and A. Schiffauerova. The impact of individual collaborative activities on knowledge creation and transmission. *Scientometrics*, pages 1–29, 2017. Springer.