Journal of Informetrics 8 (2014) 728–737

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search

José Luis Ortega*

Cybermetrics Lab, CCHS-CSIC, Madrid, Spain

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 22 May 2014
Accepted 1 July 2014
Available online 27 July 2014

Keywords:
Bibliometrics
Academic search engines
Ego networks
Research impact
Co-authorship

ABSTRACT

The main objective of this study is to analyze the relationship between research impact and the structural properties of co-author networks. A new bibliographic source, Microsoft Academic Search, is introduced to test its suitability for bibliometric analyses. Citation counts and 500 one-step ego networks were extracted from this engine. Results show that tiny and sparse networks – characterized by a high Betweenness centrality and a high Average path length – achieved more citations per document than dense and compact networks – described by a high Clustering coefficient and a high Average degree. According to disciplinary differences, *Mathematics*, *Social Sciences* and *Economics & Business* are the disciplines with more sparse and tiny networks; while *Physics*, *Engineering* and *Geosciences* are characterized by dense and crowded networks. This suggests that in sparse ego networks, the central author have more control on their collaborators being more selective in their recruitment and concluding that this behaviour has positive implications in the research impact.

1. Introduction

Scientific collaboration is an intrinsic part of the research activity because it makes possible to share technical resources and to exchange new ideas that help to face new scientific challenges (Katz & Martin, 1997). However, and as argued by Latour and Woolgar (1986), the scientists' behaviour could be influenced by the type of research that they perform that makes possible producing different research products. In that sense, it could also be suggested that each research discipline generate a particular collaboration pattern. On the other hand, the current professionalization of Science has increased the number of scientific partners such as in large and dense institutional research groups, due to mobility of doctoral students or visiting professors, and thanks to the ubiquitous telecom facilities. The role and involvement of these partners differ considerably when it comes to carry out a research project. In this way, the disciplinary differences and the multiple roles that each actor acquires in a research process are key elements to understand co-authorship activity in relation with the research impact and productivity (Narin & Whitlow, 1990). Even more, the existence of multiple and diverse interactions between all the contributors could emerge complex partnership structures which properties could also affect the research performance. This new point of view based on networking structures and helped by social network analysis (SNA) techniques can puzzle out how the collaboration structures are able to configure the production and the quality of the scientific research

* Correspondence to: Cybermetrics Lab, CCHS-CSIC, Albasanz, 26–28, 28037 Madrid, Spain. Tel.: +34 916022603.
E-mail address: jortega@orgc.csic.es

http://dx.doi.org/10.1016/j.joi.2014.07.001
1751-1577/

(Moody, 2004). Under this relational perspective, this study explores the importance of these structures in the academic success of research papers.

2. Related research

Scientific collaboration can be expressed in several forms, but perhaps the best way to identify a collaboration process in Science is the co-authorship of research papers. For this reason, bibliometric studies have addressed this issue profusely going from global or national views (Braun & Glänzel, 1996; Leydesdorff & Wagner, 2008) to the individual level (Melin, 2000; Newman, 2004). Regarding that level, different approaches have dealt with the relevance of the collaboration to individual researchers. At macro level, large co-author networks were characterized by scale-free network properties and power law distributions that disclose preferential attachment in collaboration processes and its relationship with the scientific production (Barabâsi et al., 2002; Newman, 2004). Other studies were focused on the influence of the collaboration origins and types on the research production and impact with disparate results. There are thus numerous papers that found a positive relationship between the international collaboration and the research impact (Katz & Hicks, 1997; Narin & Whitlow, 1990), whereas others did not find any influence at all (Herbertz, 1995; Leimu & Koricheva, 2005). According to research disciplines, Gingras and Archambault (2006) detected different collaboration patterns between the Social Sciences and the Humanities, being Social Sciences closer to Natural Sciences than to Humanities. In more detail, Stefaniak (2001) and Lee and Bozeman (2005) evidenced that "expensive disciplines" such as Physics, Chemistry and Biomedicine were more collaborative, in spite of that no significant citation difference across disciplines were observed (Vaughan & Shaw, 2005).

One of the most interesting aspects, that curiously have been less studied, is the analysis of the structural properties of ego-centred networks in relation with their productivity and impact. For example, Eaton, Ward, Kumar, and Reingen (2002) found that the productivity is associated with centrality degree confirming that the scientific publishing is related with collaboration; Börner, Dall'Asta, Ke, and Vespignani (2005) presented several network measures that followed the changing impact of author-centred networks. Yan and Ding (2009) analyzed the Library and Information Science co-authorship network in relation to the impact of their researchers, founding important correlations, although these were not normalized per papers. Abbasi's team has profusely studied the relationship between scientific impact and co-authorship pattern. Thus, they discovered significant correlations between network indicators (Density and Ego-Betwenness) and performance indicators, as *g*-index (Abbasi, Chung, & Hossain, 2012) and citation counts (Abbasi et al., 2014), even if with different results. Whereas Hu, Rousseau, and Chen (2012) proposed several structural indicators to measure the impact at article level. McCarty, Jawitz, Hopkins, and Goldman (2013) attempted to predict the *h*-index evolution through ego networks, uncovering that this indicator showed a high transitivity, that is, it increased as new partners with high *h*-indexes were selected. However, no studies have focused on the relationship between impact and co-authorship networks across scientific fields, as well as analysing balanced ego-networks with a same size.

Other relevant and original aspect of this study is exploiting Microsoft Academic Search (MAS) data, a recent academic search engine that contains aggregated data of researchers and institutions. In fact, very few papers have studied this service. Jacsó (2011) was the first to critically explore the functioning of the database, while Hand (2012) described its potentialities. Ortega and Aguillo (2014) used a comparative approach to analyze its performance as research assessment tool. To the best of our knowledge this is the first attempt to analyze this tool to build a bibliometric view.

3. Objectives

Three main questions are laid out in this work:

- Is there any relationship between the scientific performance of a researcher, expressed by bibliometric indicators, and their collaboration patterns, measured with network indicators? And, even more, could this relationship be measured in isolated and balanced ego-networks using data from Microsoft Academic search?
- Could these network indicators significantly vary according to the research areas? That is, if are there research disciplines where the collaboration patterns change due to their intrinsic research activity? And, are these differences related with the research impact?
- Is there any influence between the different types of collaboration (local, national and international) regarding to the research impact?

4. Methods

4.1. Data

MAS is a scientific web database which gathers bibliographic information from the main scientific editorials (Elsevier, Springer) and bibliographic services (CrossRef). It roughly contains 38.9 millions of documents and 22 million of profiles (Microsoft, 2014) that are automatically created for the each of the authors of these papers. Among other features, MAS presents a personal profile which provides not only the author's list of publications but also relevant bibliometric indicators (publications, citations), the disciplinary areas of interest and other rosters showing the most frequent co-authors, preferred

journals and a few keywords. The reason for choosing this tool is because the publication data are already aggregated at author level, which makes easier the structural analyses on the collaboration patterns among academic authors. In addition, this source permits to analyze ego-networks independently of their research fields and with the same size. Nowadays, it is the only source that provides free access to this information in the public Web.

The collection process of co-author networks was implemented following a snowball sampling procedure of three steps (Goodman, 1961). This technique is appropriated when it is intended to track communication flows among network members. First, a random list of 500 MAS author IDs was randomly taken. Then a SQL script was written to extract the co-authors list from each profile. As many authors supply a large list of partners, many of them collaborating only in one or two papers, only the ten most active colleagues from each profile were considered. This also avoids that the ego network structure be affected by the size and every network can thus be compared in a more exact form. Then a new list of IDs from these contributors was built again and for each one the ten most important co-authors were extracted as well. This was done to know the specific relationships between the partners of the initial 500 profiles. The final list of profiles consists of 32,213 authors and 51,054 links. Data extraction process was carried out during July 2013.

However there are several weaknesses that should be taking into account before starting any in-depth analysis on MAS data (Ortega & Aguillo, 2014):

- *Duplicated profiles*: It is estimated that an 11% of profiles are duplicated (Ortega, 2014). This does not seriously affect to the initial sample of 500 authors, but it indeed does to the subsequent list of co-authors because it is frequent that similar profiles could appear as co-authors. Duplicated co-authors with the lowest number of papers were removed.
- *Poor updating*: other important problem is that MAS only updates their databases once a year which could cause that some partners are yet not computed by MAS. However, this data absence is not very significant and it would only affect to infrequent partners. Besides, it includes profiles of inactive authors whose activity has now ceased. In this case, only profiles with an activity after 2010 were selected.
- *Disciplinary assignation*: subject matter classification is based on journals, in this way one author is assigned to one or another discipline according to the journals where he/she has published. This would cause that an author, who mainly publishes in multidisciplinary journals, could be misclassified in fields far away from his/her research line. However, the solution of this problem is complex but its influence, I think, is not decisive for the results.

For all these reasons it is strongly recommended to perform a previous cleaning process to remove and correct these problems, mainly with duplicated authors. In this case, each name was split in name, second name or abbreviations, surname and second surname. Next, these parts were crossed to identify coincident profiles which were later confirmed exploring their interests and affiliations. This matches names with or without abbreviations, for example, *Michael A. Smith* and *Michael Smith*; inverse names such as *Smith Michael* or incomplete names such as *M. Smith*. MAS also shows parsing problems, thus names such as *Michael Smitha* or *Michael SmithMichael* were manually corrected. When similar profiles were identified, these were merged and computed as two collaborations with the same author. As profile's names are extracted from papers, it is not possible that two profiles contain the same paper. Therefore, publication and citation from similar profiles could be summed up.

MAS developed its own subject classification system to group profiles. It is organized in 15 main categories subdivided in additional lower level categories (Microsoft, 2011). As a profile can be classified in different main categories, in several cases it was necessary to explore the publication list to assign a unique research area.

4.2. Bibliometric indicators

From each personal profile the following bibliometric indicators were collected from MAS.

- *Co-authors*: Total number of co-authors mentioned and with a profile in MAS.
- *Cit./Pap.*: Total amount of citations received by each author divided by the number of papers.
- *Papers*: Articles co-authored by the researcher with any of his/her first 10 co-authors.

Additionally the affiliation of these 10 most active co-authors was also specified. The purpose is to classify each collaborator as local, national or international partner. In this sense, local collaboration refers to co-authors from the same institution, national involves authors from the same country and international refers to foreign countries affiliations. Each contributor only can belong to one group to reinforce the independence of these variables; therefore local partners are not included into the national category. MAS uses a normalized list of organizations that are assigned to each author, avoiding the cleansing and correction of this information. Its disadvantage is that it has difficulty with authors that have worked in different organizations, because this information corresponds with the last working place. Other limitation is that the slow updating provokes that authors recently moved still show the old affiliation. These mistakes are difficult to detect and could influence the results on collaboration types.

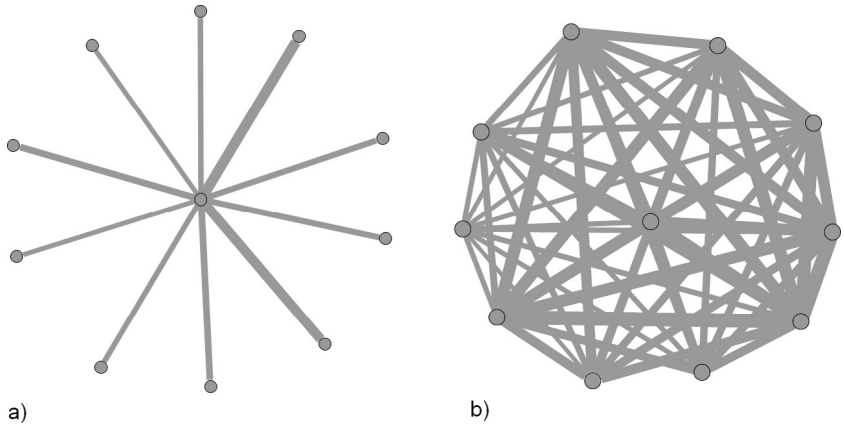

Fig. 1. Two opposite examples of an ego network. The *star-shape network* (a) and the *sphere-shape network* (b).

4.3. Egonets

Egonets or ego networks are those networks that have a central point, ego, with the network arranged around this point, describing the relationship of this ego with its environment. The size of these networks is measured by the number of steps of any node to reach the centre or ego. In this study, only 1-step neighbourhood networks are analyzed which include all the ties between the ego and their partners as well as the links among these same partners.

Fig. 1 shows two extreme examples of an ego network. The first one (a) describes a *star-shape network* with few links and isolated co-authors with limited collaboration between them and where the ego author dominates the network. These networks describe solitary authors that occasionally collaborate with other (usually) isolated authors. According to network indicators, they are characterized by a high betweenness centrality, a low average degree and a poor average clustering coefficient. In contrast to the first, the *sphere-shape network* (b) shows a dense and highly connected network in which all the members collaborate between them and where the central node has lost importance in the network. These structures are usual in large scientific groups where every paper is authored by the entire research team. These networks present a high average degree and a strong clustering coefficient while the betweenness centrality and the distance show low values.

To analyze each ego network and to measure the collaboration topology of each network, several structural indicators were used. These were grouped in two categories: The first one contains measures at the local level in order to observe the relationship of a node with its environment.

- Weighted degree centrality (k): It measures the number of lines incident to a node (Freeman, 1979). A variation is the weighted degree centrality, which calculates the weight of each line, indicating the strength of each relationship. In this study the weighted degree centrality enables to count the number of written papers between two researchers and measure the collaboration degree of an author with their partners.
- Freeman's Betweenness centrality (CB): It is defined as the capacity of one node to help to connect those nodes that are not directly connected between them (Freeman, 1980). This measurement enables us to observe the importance of an author in his/her own context. A high betweenness centrality would hence show that a researcher mediates among all their partners and he/she achieves an important role in the research performance of their contributors, being a bridge among their colleagues. Whereas, a low betweenness centrality could be a symptom that an author does not control the network and that the ties between their partners are independent to his/her will.
- Clustering coefficient (C_i): this indicator measures in what extent a node establishes a perfect cluster, in which all the partners are connected between them. It is calculated as the proportion of observed triads by the possible ones, where triad is a complete interconnected cluster of three nodes. This measure indicates the propensity of an author to create close groups with their partners. A high clustering coefficient means that one author has a dense and interrelated network of co-authors, while a low clustering coefficient reports a weak network of distant collaborators.

The second group consists of several global indicators at network level that were used as well to analyze the influence of the whole network in the research activity of an author.

- *Average weighted degree centrality*: it is the average of the weighted degree centrality of the entire network. It notes the overall collaboration strength between the network members. A high average weighted degree centrality thus shows a densely collaborative group and indirectly identifies a highly productive group.
- *Average clustering coefficient*: it is a measure that shows the proportion of nodes that tend to group together. Mathematically, it is the proportion of closed triads by open triads in the whole network. This measure is important to detect "small world" phenomena in collaborative networks and it is an indicator of the level of participation between members. A high Average

Table 1
Correlation matrix between bibliometric and network indicators (Spearman's correlation coefficient).

	Variables	Co-authors	Cit./Pap.	Papers
Network indicators (local level)	Weight Degree Centrality	**0.625**	0.024	**0.901**
	Clustering Coefficient	**0.167**	**−0.313**	**0.421**
	Betweenness	**−0.183**	**0.303**	**−0.438**
Network indicators (global level)	Average degree	**0.184**	**−0.301**	**0.399**
	Average Weight degree	**0.511**	**−0.202**	**0.749**
	Density	**0.182**	**−0.294**	**0.388**
	Average Clustering Coefficient	**0.130**	0.042	**0.164**
	Average path length	**−0.163**	**0.304**	**−0.401**

Values in bold are significantly different from 0 with a significance level alpha = 0.05.

Clustering coefficient shows that all the partners of the network establish many links between them, while a low average describes contributors separate from one another.

- *Average path length*: it is the average of the minimum number of steps that a node needs to reach any other one in the network. The average of these entire shortest paths is used to calculate the efficiency of a network and shows the closeness degree of the authors in a collaboration network. A large path length means that each contributor needs many steps to reach other partners, making clear that these colleagues are unaware themselves. However, a short path describes a close team of co-authors.
- *Density*: it is the proportion of the maximum possible number of incident lines between nodes. It is used to measure how compact a network is. High rates of density show compact and close networks, while low values are symptoms of weak and isolated networks.

4.4. Statistics

Principal Component Analysis (PCA) was used to plot the bibliometric and structural indicators to observe their relationships and to detect the disciplinary distribution patterns of the authors (Hotelling, 1933). The objective is to visually observe if there is any relationship between these indicators and their observations from a disciplinary view. The aim of the PCA is to reduce the dimension of p variables to a set of new variables (principal components) which contain the highest amount of information from the original variables. These components or factors are uncorrelated between them, because the first one has the highest amount of information, and the second one has the information that the previous one does not contain and so on. To simplify the components' structure and therefore to make its interpretation easier and more reliable, it is usual to apply rotations to the components. The most popular rotation method is Varimax because makes that each component represents only a small number of variables.

Kruskal–Wallis H test (1952) detects if n data groups belong or not to the same population. This statistic is a non-parametric test, suitable to non-normal distributions such as the power laws observed in bibliometrics.

Dunn's post-test (1961) compares the difference in the sum of ranks between two columns with the expected average difference (based on the number of groups and their size). It is used after applying the Kruskal–Wallis or the Friedman test. The Dunn's test resolves which samples are different and groups them in sets named *A*, *B*, etc.

These statistics were calculated with the Excel plug-in XLStat 2008, while the network indicators were worked out using Gephi 8.0.

5. Results

5.1. Correlations

Table 1 presents a correlation matrix between the structural indicators and the bibliometric ones. These correlations are calculated using Spearman's correlation coefficient because many of these variables are not normally distributed and a non-parametric coefficient is more suitable and robust than the Pearson's coefficient. Correlations with statistical significance (95%) are signed in bold. Weighted degree centrality and average weighted degree centrality are the network indicators that best correlate with the bibliometric ones. Weighted degree centrality shows high correlation with Papers ($\rho = 0.901$) and co-authors ($\rho = 0.625$). This relationship is explained because the degree centrality counts the number of different contacts, this is, co-authors; while the weight is calculated from the number of co-authored papers. This is similar with Average weight degree that shows its best correlation with Papers ($\rho = 0.749$) and Co-authors ($\rho = 0.511$) as well.

However, it is interesting to notice the significant, although tiny, relationship between Citations per Papers (Cit./Pap.) and several network indicators, mainly with Clustering Coefficient ($\rho = -0.313$) and Betweenness centrality ($\rho = 0.303$) at individual level. The negative correlation with the Clustering Coefficient shows that the citation impact is negatively related with the fact that an author is embodied in compact and dense networks as in the case of well-defined research groups. Contrarily, the positive correlation of the Betweenness centrality means that authors connected to sparse networks but in which he/she has a central and relevant position increases positively his/her research impact.

Table 2
Differences between research disciplines according to network and bibliometric indicators (individual level), grouped by Dunn's post test.

Disciplines	Profiles	Betweenness centrality			Clustering coefficient			Cit./Pap.		
		Mean rank	Groups		Mean rank	Groups		Mean rank	Groups	
Physics	80	184.86	A		309.66		B	115.05	A	
Engineering	54	198.11	A		290.57		B	176.31	A	
Geosciences	11	217.50	A	B	296.59		B	197.86	A	
Chemistry	31	242.73	A	B	261.39	A	B	206.11	A	
Computer Science	56	250.57	A	B	242.05	A	B	260.21		B
Medicine	151	256.47	A	B	238.11	A	B	291.93		B
Biology	75	283.15		B	211.21	A		328.62		B
Material Science	7	308.07		B	186.50	A		244.71	A	
Economics and Business	17	362.32		B	129.76	A		370.68		B
Mathematics	9	368.67		B	110.61	A		278.67		B
Social Science	3	411.00		B	192.67	A		384.00		B

5.2. Disciplinary differences

The second objective of this study is to know if there is any difference between disciplines when they come to shape their collaboration networks, and if these differences are related with the impact. The aim is to observe if each discipline generates different structural pattern according to the way in which they develop their research activities. Kruskal–Wallis test was used to detect statistically significant differences between disciplines and the Dunn's post-test was employed to group bilateral differences. Kruskal–Wallis test has the limitation that does not properly work with too small samples. Due to this, two categories with fewer profiles were removed (*Arts & Humanities* and *Agriculture Science*). Table 2 groups profiles by research categories and describes the average rank of each discipline according to two network indicators, Betweenness centrality and Clustering coefficient. The ratio citation per paper is also included to compare these indicators with the research impact. It is detected that these categories are mainly grouped into three clusters. The first group (*A*), including *Physics*, *Geosciences* and *Engineering* profiles shows low Betweenness centrality and a high Clustering coefficient. This means that these research areas consist usually of large research groups with a high production and an important participation of all their members, which causes a high collaboration degree and a strong cohesion of their co-authorship networks. A second intermediate and mixed group (*A*–*B*) is formed by *Medicine*, *Chemistry* and *Computer Science*, which shows average values of Betweenness centrality and Clustering degree. Finally, *Biology*, *Material Science*, *Economics & Business*, *Mathematics* and *Social Science* (*B*) are the research areas that reach the higher ranks in Betweenness centrality and low values in Clustering degree. These figures describe those researchers that establish few collaboration ties, frequently with isolated and individual authors.

Regarding Cit./Pap., the distribution is similar to the observed one in Betweenness centrality, which matches with the previous correlation results. According to that, authors from *Economics & Business*, *Mathematics* and *Social Science* have larger number of citations per document; while *Physics*, *Engineering* and *Geosciences* researchers have lower ratios of citations by article.

A similar result is observed with the indicators at global level (Table 3). *Physics* and *Engineering* authors tend to belong to networks with a short Average path length (*A*) and a high Average degree (*B*). That is, to compact and dense groups set up by many and highly productive members. The intermediate group (*A*–*B*), with *Geosciences*, *Chemistry*, *Computer Science*, *Medicine*, *Biology* and *Material Science*, consists of areas in which the length and degree show average values. Finally, *Economics & Business*, *Mathematics* and *Social Science*' authors are members of small groups whose partners are very distant between them in the network (Average path length = *B*), with little collaboration and sparse productivity (Average degree = *A*). According to the number of citations per article, the distributions of Average path length, Cit./Pap and Betweenness centrality are similar as shown in Table 2 results.

Table 3
Differences between research disciplines according to network and bibliometric indicators (network level), grouped by Dunn's post test.

Disciplines	Profiles	Average path length			Average degree			Cit./Pap.		
		Mean rank	Groups		Mean rank	Groups		Mean rank	Groups	
Physics	80	192.89	A		299.49		B	115.05	A	
Engineering	54	193.64	A		277.07		B	176.31	A	
Geosciences	11	221.09	A	B	240.55	A	B	197.86	A	
Chemistry	31	237.50	A	B	262.44	A	B	206.11	A	
Computer Science	56	250.87	A	B	234.99	A	B	260.21		B
Medicine	151	256.42	A	B	250.70	A	B	291.93		B
Biology	75	279.33		B	224.09	A	B	328.62		B
Material Science	7	307.79		B	206.43	A	B	244.71	A	
Economics and Business	17	360.53		B	123.29	A		370.68		B
Mathematics	9	373.33		B	71.78	A		278.67		B
Social Science	3	407.33		B	103.00	A		384.00		B

Table 4
Differences between research disciplines according to international authored papers, grouped by Dunn's post test.

Disciplines	Profiles	Mean rank	Groups	
Biology	69	193.25	A	
Engineering	50	199.53	A	
Material Science	7	200.00	A	
Medicine	129	200.21	A	
Computer Science	51	204.83	A	
Economics and Business	15	204.90	A	B
Geosciences	10	205.40	A	B
Chemistry	30	252.17	A	B
Mathematics	9	260.89	A	B
Physics	78	314.75		B

5.3. Collaboration type

The third question is to know if there is any difference between the type of collaboration and the research discipline. Collaboration was measure as the number of papers authored by local, national or international co-authors. Kruskal–Wallis test and the Dunn's post-test were employed in this analysis as well. Results do not detect any statistical difference between local and national collaboration regarding to research disciplines. However, it was indeed found significant differences in the international collaboration case. The Dunn's post test (Table 4) tells apart a first group (*A*) with low international co-authored papers shaped by *Biology*, *Engineering*, *Material Science*, *Medicine* and *Computer Science*; an intermediate group (*A–B*) set up by *Economics & Business*, *Geosciences*, *Chemistry* and *Mathematics*; and finally, *Physics*, the discipline with the larger number of international co-authors (*B*). However, no relationship was detected between the research impact and the collaboration type.

Finally, network and bibliometric variables, beside the different collaboration types were put in context to appreciate their relationships and observe differences between research disciplines. PCA was used to plot these variables and authors in a graph as a way to visualize the previous results (Fig. 2).

The PCA model identifies two factors at 67.5%. The vertical one groups the production-related variables such the collaboration types – collaboration is measured in co-authored papers – and the weighted degree. The horizontal one describes the structural variables, arranging in the left the indicators that better describe sparse networks, while in the right the indicators belonging to dense and compact networks are grouped. The most interesting fact is that the impact indicator (Cit./Pap.) correlates negatively with the structural factor, which confirms the influence of the structural properties of an author in

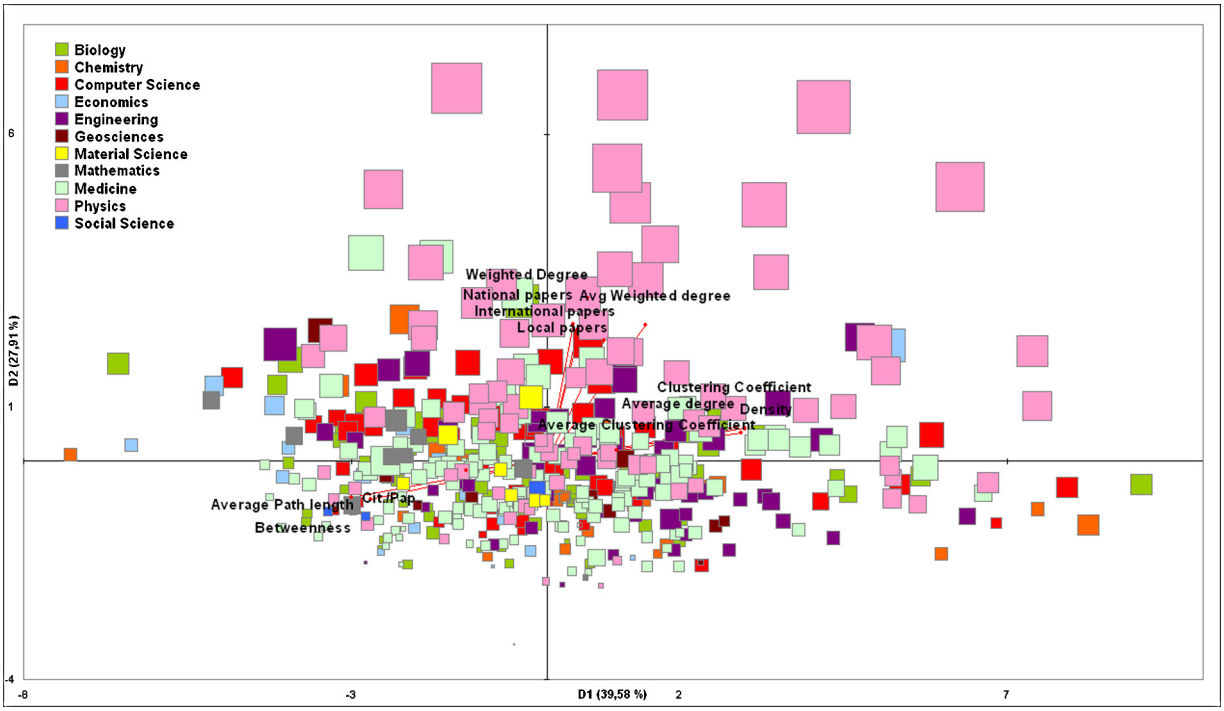

Fig. 2. PCA plot. Variance (67.5%) with Varimax rotation.

his/her research impact. Moreover, this variable is close to Betweenness centrality and Average path length, two indicators that positively describe the star-shape networks where the ego has few contacts but has full control on them, confirming that this type of authors have better research impact per document.

Fig. 2 also shows that there is not a differentiated pattern in the disciplinary distribution of authors, although it seems that the *Physics*' authors better correlate with the productive factor. It is also observable that *Mathematics*, *Material Science*, *Economics & Business* and *Social Science* are distributed near to Cit./Pap., Betweenness centrality and Average path length indicators which coincides with the results of *Disciplinary differences* section.

6. Discussion

Prior to the results interpretation, it would be relevant to state that this is one of the first scientometric analyses done with MAS data, so taking into account the reliability and consistency of the results, the use of this new web tool can be highly recommended for collaboration studies. Generalizing, the profiling web services such as MAS, Google Scholar Citations or ResearchGate can be trustworthy services that open a new perspective on the research evaluation and specifically on the collaboration studies. However, the fact that these new services are based on automatic procedures (i.e. autonomous citation indexes, web crawling and harvesting processes, parsing techniques) forces to a previous stage of cleansing data and an in-deep analysis about their sources. In this sense, further studies on the implication of these services for bibliometric analyses would be welcomed.

The obtained results enable to observe that the co-authorship is linked to the scientific production and scientific impact of a researcher, suggesting that the environment in which a scientist is located may influence in some extend on his/her research activity. In this way, not only the number of direct contributors is a performance indicator, but also the shape in which the partners are related between them can be a reflection of his/her research success. The positive and significant correlations between Betweenness centrality and Average path length with Cit./Pap., and their closeness in the PCA graph suggest that an author that just maintains bilateral collaborations (star-shape networks) with isolated co-authors has a better ratio of citations per paper. Contrarily, researchers immerse in dense networks (circle-shape networks), with a high Clustering coefficient and Density, achieve minor impact. These results fit with previous findings. Yan and Ding (2009) observed that betweenness centrality was the structural indicator that correlated closest with citations, while Abbasi et al. (2012, 2014) detected that *g*-index negatively correlated with density, pointing that sparse and isolated networks achieve better impact outputs.

However, these previous studies calculate structural indicators on an entire network, measuring the role of a researcher into his/her disciplinary network, and not into their closest group only. This provokes that authors describe different size ego-networks and therefore their structural indicators cannot be compared. Using one-step networks, and limited to the ten most active partners, is recommended because favours the comparison of similar size networks, makes possible to easily determine the main shape of the network and the role of the ego. On the other hand, it avoids external distortions caused by co-authors collaborations unaware of the ego as it happens in two or more steps networks.

Other methodological problem is that if these indicators were measured in a disciplinary network (Abbasi et al., 2012, 2014; Yan & Ding, 2009), authors located in the periphery of the discipline (interdisciplinary) will be underrepresented because many of their partners might be outside of the discipline. At last, some of these studies do not normalize citations by articles, causing that the relationship between impact and collaboration is influenced by production. Thus, correlations in this study are more precise because they are not only based on same size networks, but citations were standardized by papers as well. On the other hand, these tiny correlations also evidence that the research impact is affected by other important aspects such as language (Garfield & Welljams-Dorof, 1990), venue or research disciplines (Vaughan & Shaw, 2005).

Further understanding is able according to the Strength of weak ties theory (Granovetter, 1973), which suggests that weak links connect with external groups and therefore they provide more useful and newer information than the local partners, evidencing that many times these sporadic partners are stronger in information terms than the usual ones. So, the results can be interpreted in the sense that punctual collaborations with distant partners is most scientifically effective because it allows to obtain important information from unexpected sources which would enrich the research and, therefore, increase the research impact. On the other side, working in large research groups where all their members always collaborate in every paper and where is hard to see external contributors to this group, could be more secretive and with less fresh information. In spite of the much more productive environment, the gains in citations per paper are lower. This is true if the impact is measured as a relative indicator, avoiding the size factor of the number of papers. In absolute terms, cooperation in itself increases the production and therefore the likelihood to be cited.

From an efficiency point of view, these results fit with those from Bavelas (1950), who found that centred structures are more efficient. In this sense, authors that are the centre of their collaboration networks (high Betweenness centrality) have more control on it and enable them to manage the direction of the collaboration to obtain better results. On the contrary, authors that do not control the information flows of their partners have less influence on their research, which it could negatively affect their research success.

According to disciplinary differences, Kruskal–Wallis tests confirm that disciplines with a high Betweeenness centrality and Average path length are also the research areas that most citations per paper achieve. Thus *Mathematics*, *Social Science* and *Economics & Business* are disciplines characterized by sparse networks and isolated co-authors which enhance the research

impact for a low production; while *Physics*, *Engineering* and *Geosciences* contain dense and crowded networks that need a great production to achieve a significant impact. These results apparently are in conflict with previous results because they are not measured as proportion of citation per documents (Kousha and Thelwall, 2007) or they are not studied at author level (Vieira & Gomes, 2010). In this way, it is possible that the high proportion of co-authors in *Physics*, *Engineering* and *Geosciences* disciplines tends to an elevated production that not always gets impact.

No disciplinary differences were observed regarding the collaboration type. Only international collaboration produced disciplinary differences, where *Physics* and *Chemistry* describe an important presence of abroad partners, while *Biology* and *Medicine* are characterized by compact local groups (Newman, 2001). However, these differences are not related with the research impact, observing that disciplines with a high rate of international collaboration not necessary show good citation numbers. In spite of previous studies proving that the international collaboration increases the impact (Basu & Aggarwal, 2001; Narin & Whitlow, 1990), it is possible that this relation is less relevant when relative numbers are involved. For example, *Physics* and *Chemistry* show a high impact because they present a high collaboration rate and so a high productivity, but if this was measured as citations per paper this relationship will be blurred. On the other hand, the problems in the assignation of affiliations in MAS could distort these results, no detecting significant differences by collaboration type.

7. Conclusions

From the obtained results, four main conclusions can be derived. The first one is that authors being part of sparse and thin networks, with isolated co-authors and an effective control of the network have a higher research impact; while researchers embodied in dense and bushy networks, collaborating with the same co-authors all the time and a poor control over the network information flows, obtain fewer citations per document.

The second conclusion is that these different networking behaviours are observed between disciplines. Being *Mathematics*, *Social Science* and *Economics & Business* the disciplines where sparse and tiny networks are more frequent; while *Physics*, *Engineering* and *Geosciences* are characterized by dense and crowded networks.

Finally, the third conclusion is that there are no relationship between the collaboration type, the research impact and the network structure. Both PCA as Kruskal–Wallis test have evidenced that the type of the collaboration have not any connection with bibliometric and network indicators.

A fourth conclusion is derived from the source used in this study. MAS could be a reliable tool for collaboration studies if and when their limitations are previous addressed, concretely the cleansing of duplicated profiles. In this context, results have to be interpreted taking into account the possible disciplinary and sources biases, common in any scientific database.

References

Abbasi, A., Chung, K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, *48*(4), 671–679.

Abbasi, A., Wigand, R. T., & Hossain, L. (2014). Measuring social capital through network analysis and its influence on individual performance. *Library and Information Science Research*, http://dx.doi.org/10.1016/j.lisr.2013.08.001

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3), 590–614.

Basu, A., & Aggarwal, R. (2001). International collaboration in science in India and its impact on institutional performance. *Scientometrics*, *52*(3), 379–394.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of Acoustical Society of America*, *22*(6), 725–730.

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, *10*(4), 57–67.

Braun, T., & Glänzel, W. (1996). International collaboration: Will it be keeping alive east European research? *Scientometrics*, *36*, 147–254.

Eaton, J. P., Ward, J. C., Kumar, A., & Reingen, P. H. (2002). Social–structural foundations of publication productivity in the Journal of Consumer Research. *Journal of Consumer Research*, *11*, 199–220.

Freeman, L. C. (1979). "Centrality in networks: I. conceptual clarification". *Social Networks*, *Vol. 1*, 215–239.

Freeman, L. C. (1980). "The gatekeeper, pair-dependency, and structural centrality". *Quality and Quantity*, *Vol. 14*, 585–592.

Garfield, E., & Welljams-Dorof, A. (1990). Language use in international research: A citation analysis. *Annals of the American Academy of Political and Social Science*, *511*(1), 10–24.

Gingras, Y., & Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, *68*(3), 519–533.

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, *32*(1), 148–170.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Hand, A. (2012). Microsoft Academic Search. *Technical Services Quarterly*, *29*(3), 251–252. http://academic.research.microsoft.com

Herbertz, H. (1995). Does it pay to cooperate? A bibliometric case study in molecular biology. *Scientometrics*, *33*(1), 117–122.

Hotelling, H. (1933). Analysis of a complex of statistical variables into Principal Components. *Journal of Educational Psychology*, *24*, 417–520.

Hu, X., Rousseau, R., & Chen, J. (2012). Structural indicators in citation networks. *Scientometrics*, *91*(2), 451–460.

Jacsó, P. (2011). The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*, *35*(6), 983–997.

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*(3), 541–554.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web-URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, *58*(7), 1055–1065.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673–702.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *BioScience*, *55*(5), 438–443.

Leydesdorff, L., & Wagner, C. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, *2*(4), 317–325.

McCarty, C., Jawitz, J. W., Hopkins, A., & Goldman, A. (2013). Predicting author *h*-index using characteristics of the co-author network. *Scientometrics*, *96*(2), 1–17.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, *29*(1), 31–40.
Microsoft. (2011). *Academic categories in Microsoft Academic Search*. http://social.microsoft.com/Forums/en-US/mas/thread/bf20d54a-ede2-48a9-8bbb-f6c1c1f30429 Accessed 31.03.14
Microsoft. (2014). *Microsoft Academic Window Azure Marketplace*. http://datamarket.azure.com/dataset/mrc/microsoftacademic Accessed: 31.03.14
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.
Narin, F., & Whitlow, E. S. (1990). *Measurement of scientific cooperation and co-authorship in EC-related areas of science. EC-Report EUR 12900*. Luxemburg: Office for Official Publications of the European Communities.
Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the United States of America*, *98*(2), 404–409.
Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5200–5205.
Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Witney, UK: Chandos Publishing. ISBN: 1843347911.
Ortega, J. L., & Aguillo, I. F. (2014). Microsoft Academic Search and Google Scholar citations: Comparative analysis of author profiles. *Journal of the American Society for Information Science and Technology*, *65*(6), 1149–1156.
Stefaniak, B. (2001). International co-operation in science and in social sciences as reflected in multinational papers indexed in SCI and SSCI. *Scientometrics*, *52*(2), 193–210.
Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the American Society for Information Science and Technology*, *56*(10), 1075–1087.
Vieira, E. S., & Gomes, J. A. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*(1), 1–13.
Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107–2118.