Do researchers collaborate in a similar way to publish and to develop projects?

J. Clemente-Gallardo[a,b], A. Ferrer[a], D. Íñiguez[a,c], A. Rivero[a], G. Ruiz[a], A. Tarancón[a,b]

[a] *Instituto de Biocomputación y Física de los Sistemas Complejos, Edificio I+D-Campus Río Ebro, Universidad de Zaragoza, C/ Mariano Esquillor s/n 50018 Zaragoza (SPAIN)*
[b] *Departmento de Física Teórica, Universidad de Zaragoza, Campus San Francisco, 50009 Zaragoza (SPAIN)*
[c] *Fundación ARAID, Diputación General de Aragón, 50004 Zaragoza, Spain*

Abstract

The aim of this paper is to establish the similarities and differences between the way of collaboration and the production of researchers when dealing with publications or with the development of projects and whether the collaboration patterns change accross disciplines.

We have studied the networks of researchers formed through the collaborations in papers or in projects in a research institution (the University of Zaragoza) and we have analyzed a series of individual and global magnitudes. As a general result, we have observed that the *laws* governing the individual productivity are similar for the cases of publications and projects but, however, the behavior is different when analyzing more complex magnitudes such as the collaborations or other structural variables. We consider also the subnetworks defined by the researchers of the different disciplines and characterize their topologies and compare the corresponding collaboration patterns.

Because of the general approach, we expect most of the conclusions to be applicable to other universities or research centers.

Keywords: Academic performance, Complex networks, Co-authorship, Research projects, Knowledge areas

1. Introduction

Network analysis (Newman (2010)) is an important tool to analyze academic performance, especially from the beginning of this century (see for example (Barabâsi et al (2002)) as one of the seminal papers). It has been extensively used for bibliometric analysis, mainly through the study of co-authorship networks (a recent review of the literature of this topic can be seen in (Kumar S (2015))) and citation networks (Van Eck and Waltman (2014)).

In this article we will deal with paper co-authorship but also with the collaborations in the development of research projects in an institution. To construct the networks we will define the nodes to be the researchers of the institution and the links between

them will be given by the relations based on common articles[1] or funded research projects. Network analysis allows us to see global properties in contrast with statistical analysis based on individual or local production. In this way, we can analyze a large amount of data, identifying global properties such as communities, leaderships, time evolution, clustering, central and peripheral groups, etc. In (Álvarez et al (2015)) a software platform designed to perform this analysis was presented and the University of Zaragoza (Spain) was chosen to exemplify its power.

One of the main advantages of our framework is the possibility of studying, at the same time, the networks of publications and funded research projects. It is well known (see (Katz and Martin (1997))) that studying co-authorship relations is only one of the methods to analyze collaboration between researchers. Studying the structure of project collaboration provides an alternative view whose similarities and differences with the co-authorship one must be analyzed. The main problem for this is that while data concerning research papers are now accessible for almost all the areas in several platforms[2] there are not many public databases about funded research projects and the researchers involved in them.

Thus, while for publication networks there exists a large number of studies showing their structure, we could find only a few works studying project networks. Furthermore these studies use in general a small amount of data and hence they only consider small networks (Bellotti (2012); Lemes Alarcão and Sacomano Neto (2016); Miguel et al (2012)).

The University of Zaragoza (UZ) is a medium-size European university with around 35000 students and more than 4000 staff researchers. UZ decided some years ago to create a corporate database containing all the research data of its researchers, under the name of Sideral. It contains all research data, of publications and research projects, after 1990, with all details about them. Particularly subtle issues as disambiguation of the researcher names in publications is performed semi-automatically with human supervision. By using this corporate database, exhaustive in articles and projects, we have been able to create networks of researchers based on their relation as coauthors of a paper or as collaborators in a funded research project. In (Álvarez et al (2015)) we presented the platform and the main qualitative conclusions of a first analysis of our university, with a preliminar comparison of the topological properties of the networks. In the present paper we consider a quantitative analysis of some properties of the research and projects networks focusing on their differences. The study includes around 90,000 articles, 30,000 projects and 7,000 researchers of the UZ from the nineties up to now. We will also consider the subnetworks defined by the researchers of the different areas (Sciences, Life Sciences, Arts and Humanities, Engineering & Architecture and Social Sciences) and compare their properties. Thus, we will obtain the differences in the collaboration patterns of the different areas.

Particularly, with these tools we intend to answer three questions:

[1] Please notice that we will talk indistinctly of articles, papers or publications but we mean any kind of JCR indexed publications

[2] It is important to remark, nonetheless, that in many cases, an additional work is needed because a correct identification of authors is not fully automatic due to differences and bugs in signatures and affiliations

- Is it possible to compare the two networks (papers and projects), from a quantitative point of view? In particular, what is the evolution of papers and projects in time, and what is the scaling of productivity of the nodes in the two networks?
- Are there differences in the form researchers collaborate when publishing a paper and when developing a research project?
- Do these differences depend on the area? Are they the same for Sciences and for Arts, for instance?

We are aware of the fact that our conclusions refer only to UZ. Nonetheless, because of the general approach and the large number of researchers considered, we expect most of the conclusions to be applicable to other universities or research centers.

The structure of the paper is as follows. In Section 2 we summarize the main aspects of the framework introduced in (Álvarez et al (2015)) and its main implications in what regards the analysis of productivity of researchers, at the level of paper production and at the level of funded research projects. Section 3 presents the main contributions of the paper: the results of a quantitative analysis for the case of the University of Zaragoza. Finally Section 4 presents the main conclusions of the paper.

2. Academic production as a complex system

Let us review now the main aspects of our framework (see (Álvarez et al (2015)) for a more detailed presentation). Consider an academic institution U and the corresponding set of researchers $\{R_k\}$ who belong to U. We will consider the interaction between the researchers from two different points of view:

- co-authoring: two nodes (researchers) are connected if they have co-authored a scientific paper. As we are considering all areas of knowledge, and for the sake of simplicity, we will not consider the different authoring patterns used in the different areas (in mathematics and theoretical physics, for instance, it is frequent to sign papers in alphabetical order while in (bio)-chemistry, first authors are usually the graduate students while last authors are the directors of the project). Hence, we consider this to be an undirected graph.
- project-collaboration: We will be considering two different types of project networks. In the first case, already considered in Álvarez et al (2015),we shall say that two researchers are connected if one is the Principal Investigator (PI) and the other collaborates in the same (funded) research project. In this case, there is a natural order in the graph, since the PI is clearly playing a different role than the other researchers. We will consider thus that the project graph is a directed graph where the PI's are the source for all the links corresponding to each project, that join him/her with the rest of researchers. We will call this type of network the **directed project network**. But we will also consider an alternative description where the researchers which collaborate in a research project define an un-directed graph, analogous to the network defined by the researcher co-authoring a paper. We will call this type of network the **un-directed project**

network. As we will see below, both descriptions capture different properties of the collaboration, and provide complementary information about it. In any case, both types of project network can be compared with the paper one, and their similarities and differences can be analyzed. This is our main goal in this paper.

Out of this large network, we can extract the subgraphs corresponding to particular departments, research groups, Institutes, etc. Hence, it allows us to consider the Institution at the micro, meso and macrolevel.

2.1. The publication network

2.1.1. The different metrics

Once the nodes and the links are defined, we must consider different metrics to assign weights to the links of the graph. There is no simple mechanism to define such a metric and many different choices are possible. In the present version of our platform, we consider three:

- a constant value for each paper. This is the simplest case but it does not take into account the quality of the publication.
- the JCR impact factor of the journal where the paper is published, at the year of publication. This may be a good metric if we consider a case where all the researchers belong to the same area, as in the case of a research institute or university department. But, at the same time, it is not a good metric to compare researchers working in very different areas, since the absolute value of the impact factor of the journals of ISI areas changes significantly. For instance according to JCR 2014, an impact factor of 1.6 belongs to the first quartile of the area *Physics Mathematical*, while it is in the last quartile of the area *Biochemistry and Molecular Biology*.
- a discretized version of the NJP metric (Normalized Journal Position, introduced in (Bordons and Barrigón (1992)) and discussed in (Costas and Bordons (2007)) in relation to the h-index) with respect to the position of the journal in the corresponding ISI area. We consider a simpler version based on the JCR-quartiles in the year of the publication:
  - assign 4 points to the paper published in a Q1 journal,
  - 3 points to the paper published in a Q2 journal,
  - 2 points for papers in Q3 journals
  - and 1 point to papers in Q4 journals.

  If the journal belongs to more than one category, we choose the one with the highest position. It is immediate that this metric is just a discrete version of NJP, where we consider just four intervals instead of the NJP value. Both choices (our scheme and NJP) solve the problem of the absolute value of the impact factor for different JCR areas, since all areas are weighted in the same way.

In all three cases, we introduce the possibility of considering the number of co-authored papers in the weight of the link. Thus, the weight of the papers with a large number of authors may be shared by the co-authors (see Grauwin and Jensen (2011) for discussions of co-authoring metric definitions).

Notice that the most popular measures defined for researchers (based on citations and the h-index, for instance, see (Alonso et al (2009)) and references therein) are not implemented yet in our platform. Despite its popularity, and its usefulness to compare researchers of the same area and age, they exhibit two problems that we are trying to understand and control:

- different areas have different citation patterns. Therefore, in some of them the number of entries in the bibliography of a paper may be 10 and in some other, the average bibliography section may contain 100 entries. Of course, this fact gives a different effective value to citations in their publications, and therefore a normalization mechanism is required. We are planning to add this feature to the next version of our platform, although besides this normalization problem,
- this is a measure which changes constantly in time (even for a fixed set of publications) and therefore the topology defined on the network by such a metric would not correspond to stable properties.

Future versions of our platform will allow us to include also the citation-based metrics and therefore the comparison between the different approaches.

2.1.2. The normalization of the weights

We can also consider the normalization of the graph, by splitting the measure chosen among its authors in the graph. In particular, we can divide the weight w_n corresponding to a given paper n among the square of the number of co-authors of the paper contained in the graph ($N_{int}(n)$, while $N_{tot}(n) = N_{int}(n) + N_{ext}(n)$, $N_{ext}(n)$ corresponding to the authors of the paper who are not contained in the graph). As there is a total number of $N^2_{int}(n)$ links for each paper and the total effect of every paper published in the institution must be equal to its weight w_n, the weight of each link becomes

$$L_n = \frac{w_n}{(N_{int}(n))^2} \tag{1}$$

Of course this is only one possible normalization criterion. Our choice coincides with the *Third collaboration network* discussed in (Batagelj and Cerinsek (2013)) if a weight $w_n = 1$ is assigned to all papers. If $w_n \neq 1$, the total contribution of each paper to the graph and to the degree of each researcher does not coincide with theirs, although they are proportional and the idea used for the normalization criterion is similar. Notice that to make it consistent we must consider always a "self-link" for each of the authors. As it does not offer any topological information, we shall omit it when representing the networks graphically, unless the node is disconnected from the rest of the graph. Notice also that the information is anyway contained in the map since the diameter of the disks representing the nodes are proportional to its degree, which is computed as the sum of all the links starting at the node and weighted with the quality index considered. Thus the omission of the self-links is harmless.

Notice that the weight of the links of the graph also affects the degree of the researchers. Indeed, the effect of the paper n on the degree of the researcher j who is one of its co-authors, becomes:

$$g_j = \frac{w_n}{N_{int}(n)}. \tag{2}$$

Notice that, defined in this way, papers published with external collaborators contribute more to the degree of the researcher than the papers co-authored with people in the institution (because of the normalization factor). Thus the resulting degree does not allow us to compare the total scientific production of two researchers directly, we can just compare their influence on the institution. On the other hand, papers signed with external researchers do not provide stronger connections with the other nodes of the graph (which represent the researchers of the institution).

2.2. The directed project network

We can consider a similar construction for the case of research grants or funded projects in general. The choice of the network type is not as simple as in the case of the publication network (see, for instance, (Maggioni et al (2014))). Our choice in (Álvarez et al (2015)) was to assign a direction to the links connecting the nodes, from the Principal Investigator of the project to all the researchers collaborating with him (denoted as CI's). The result is a directed graph, as we mentioned in the previous section.

2.2.1. The metric

Again, we consider the assignment of a weight to the different links. The simplest choice is to fix the weight as the total amount covered by the corresponding grant. Although other choices may also be considered, taking into account normalization criteria per area which we plan to include in future releases, we judge that our choice represents a good equilibrium between quality and simplicity.

2.2.2. The normalization

As in the publication network, the weight of the links are normalized by the number of researchers of the institution collaborating in the project, i.e., the weight of the link L_n is equal to the part of the funds €(n) of the project n that corresponds to each researcher of the team (we assume for simplicity that they are equally distributed but this is not necessary):

$$L_n = \frac{€(n)}{N_{int}(n)} \tag{3}$$

Notice again that we use N_{int} for the normalization and not the total number of researchers of the project (i.e. we do not include the researchers external to the UZ). This has similar consequences to the ones commented above for the publications network.

Again, the assignment of weights to the links of the graph also affects the degree of each node (i.e., each researcher). The contribution of a project n to the strength of node j is the total amount of the grant if the node represents the PI, and zero otherwise.

2.3. The un-directed project network

The choice of the project network as a directed one captures several important features of the type of interaction the researchers are subject to when collaborating in a research project. Indeed, it does describe the flow of money covering the expenses of the project, which is distributed only through the PI. This makes the node of the PI to be different from the other nodes, since it is the only one responsible of the funds assigned to the group. This was the property we considered in Álvarez et al (2015) to fix the project network to be directed. Nonetheless, the choice is not perfect, as there are structural properties of the network that are affected by this choice, which explains well the flow of funding, but not necessarily the other aspects of the scientific work. Actually, part of that work related to the project is producing also papers as output, and therefore it makes sense to considered an un-directed network to represent it. Hence, we are considering two complementary descriptions of the project network, aiming to capture a larger set of properties in the combined model.

2.3.1. The metric

Again, we consider the assignment of a weight to the different links. The simplest choice is to fix the weight as the total amount covered by the corresponding grant. From that point of view both projects networks are weighted in the same way.

2.3.2. The normalization

As in the publication network, the weight of the links are normalized by the number of researchers of the institution collaborating in the project, i.e., the weight of the link is equal to the part of the funds of the project that corresponds to each researcher of the team. As now there are $N_{int}(n)^2$ links inside the graph, the normalization is equivalent to the one chosen in the publication network, but we replace the publication metric by the total amount of funds. Hence, the normalization of each link becomes now:

$$L_n = \frac{€(n)}{N_{int}(n)^2} \tag{4}$$

Notice again that we use N_{int} for the normalization and not the total number of researchers of the project (i.e. we do not include the researchers external to the UZ). This has similar consequences to the ones commented above for the publications network.

Again, the assignment of weights to the links of the graph also affects the degree of each node (i.e., each researcher). The contribution of a project n to the strength of node j is now the corresponding share of the total amount of the grant.

3. Quantitative results

The study uses all the data present in the UZ corporate database (Sideral) from the origin up to 2017 both for publications and projects. There are a few more ancient dates, but the information is exhaustive only from the late nineties. For this reason, despite the networks are calculated with all the information, when we present yearly evolutions of some variables, we will do it only between 2001 and 2017.

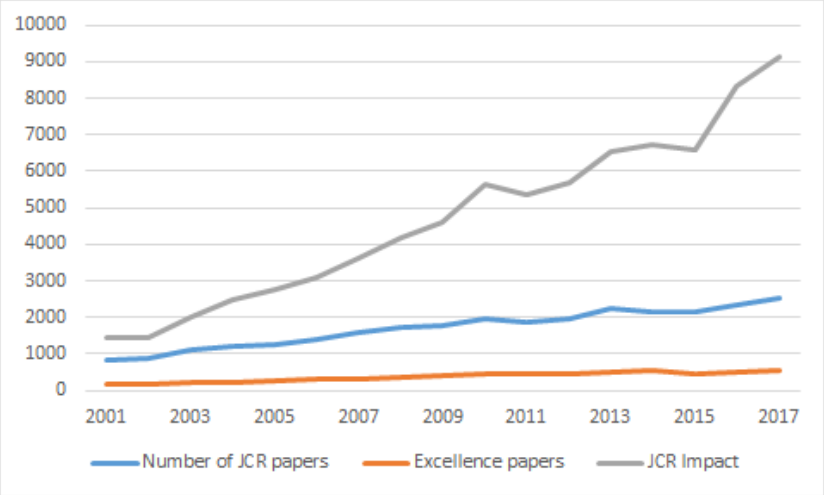

Figure 1: Evolution of paper production in the UZ between 2001 and 2017. We represent the total number of JCR papers, the corresponding JCR impact, and the number of those paper which are classified in the first decile of their corresponding JCR areas.

The whole set of data includesa a total of 93827 articles with 6336 authors affiliated to the University of Zaragoza as well as 29080 projects with their PI from the UZ, involving 6768 researchers of this university. Almost all the researchers in the article network are contained in the project network. The excess of researchers of the second network corresponds, mostly, to technicians, hired with project funds but that may not appear in the publications.

3.1. Global considerations

To have an idea of the global numbers of the system we can represent the paper production and funds obtained in projects by all the researchers in UZ, as we can find in Figs. 1 and 2.

In the paper production plot, we represent the total number of JCR papers, the corresponding JCR impact, and the number of those papers which are classified in the first decile of their corresponding JCR areas.

In the project production plot, we represent in two different curves the funds corresponding to pure research projects and the funds corresponding to applied research developed mainly with Industry.

It is easy to see how the production had a high growing in the first years of the century and how the economic crisis has had a very negative impact from 2009 on, especially in the evolution of project funds.

3.2. Scaling of Scientific Production and Funds

Let us start by analyzing the scaling laws followed by both networks, the papers network and the research projects network. For each case a few considerations are in order

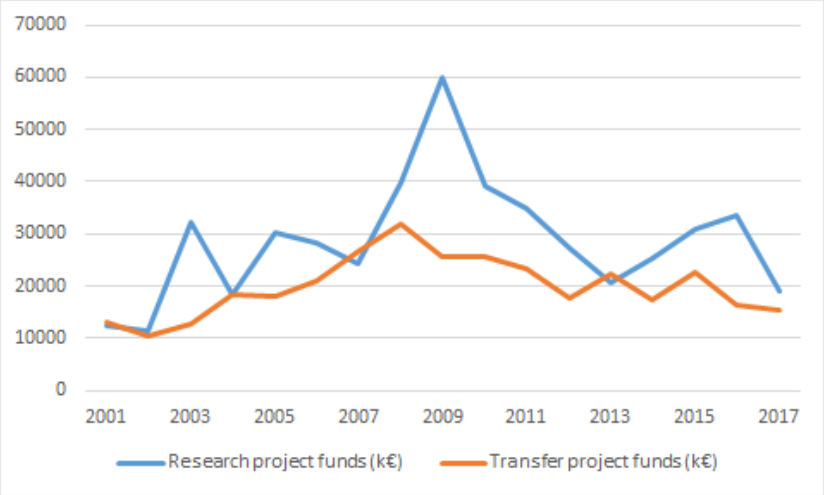

Figure 2: Evolution of the funds obtained by projects in the university between 2001 and 2017. In blue appears the total amount of funds obtained from research projects and in orange the amount coming from transfer projects with companies.

- In the paper network, we can consider two different metrics to weight each link:
  - the quartile metric, presented above, where the paper is assigned a weight of 4-3-2-1 points depending on the quartile the journal belongs to
  - we can also consider the constant metric, where the weight of each article is always 1 (in this case the degree of each author is the number of signed papers at the UZ and hence a measure of his/her total production).
- As the topology of the network is not relevant in this analysis we restrict, for the sake of simplicity, to the directed project network. For that case, we also consider two different types of weights
  - the weight can be considered to be proportional to the project funds divided by the number of Members. Therefore each link has a strength proportional to the funds *traveling* from the PI to the CI.
  - The second possibility is to consider the same total weight for each project (equal to one), and therefore the weight of each link corresponds to one divided by the number of members.

  With respect to these measures, we consider the degree of each researcher (i.e., the diameter of its disc in the map). In the first case, the degree of each node is proportional to the total funds of the projects in which the researcher participates. In the second case, the degree is proportional just to the number of projects where the researcher participates.

We start the analysis with the normalized cumulative histogram for the degree of the nodes (Figure 3), with respect to the four metrics presented above

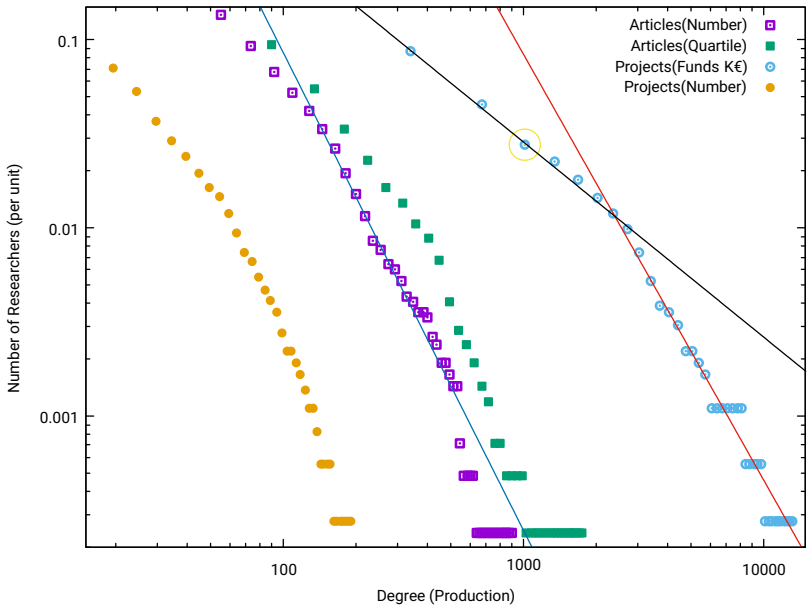

Figure 3: The X axis represents degree, and the Y axis, the number of researchers with a degree larger than or equal to this value (per unit). We use bilogarithmic scale to see the global behavior, and to control the fit quality. The lines represent fits to the function $f(x) = ax^{-b}$, a straight line in bi-logarithmic scale.

- articles weighted by quartile (i.e., quality of the journal is considered)
- articles weighted by number (i.e. total production)
- total projects funds (i.e., *quality* of the project is considered)
- total projects number.

In order to see in a single plot all the points in a clear way, we plot them in bi-logarithmic scale, and we expect approximately straight lines in the large X region (Clauset et al (2009)). We consider the degree of the node (i.e., the production of papers or projects, weighted by quality or not) in the X axis, and the percentage number of researchers with a degree equal to or larger than X in the Y axis. All curves start at $X = 0, Y = 1$ because every researcher has 0 or more degree (we do not show this point in the figure due to the logarithmic scale).

For instance, in the *article number* case (violet squares), X represents articles. We have, for example a point at $X = 200, Y = 0.0146$: that means that the 1.46% of the researchers have published 200 or more papers.

For project funds (cyan circles), we are able to see that the 2.7 percent of the researchers have generated projects (as PIs) which add up to more than one million euros (for clarity, it is marked in Fig. 3 with a yellow circle).

In order to study the behavior of curves, we can make a fit to the function $f(x) = ax^{-b}$. This fit is not perfect, and the values of a, b depend on the region of points used

for the fit. In any case, we can claim that the type of function chosen will be correct if the original curve is a straight line in the fitted region (remember we are using a bilogarithmic scale).

In general the behavior is similar for all curves, but there are important differences if we look at them in detail:

- For the total production metrics (not taking into account the quality of the contributions and represented by the violet squares and yellow points in the figure) we get very similar graphs, curving for large X in a similar smooth way. The slope of the central region is around 2.2 for both cases (blue line in the figure).

- For both quality metrics, article-quartile (green solid squares) and project-funding (cyan circles), the behavior is similar, but now we see two different regions. For small X, we have a straight region with a 1.04 slope (black line), and for large X, the slope is 2.36 (red line).

We conclude that there are no discontinuities when we consider the total production of papers: certainly it is less probable to have published 100 articles than 10, but the probability is inversely proportional to the number. However if we consider the quality of the papers or projects, there are two different behaviors: the low quality region, accessible to a large number of people, where their number decays smoothly, and a new region accessible only to a reduced number of people, where the level of results decay in a much faster way. In other words, the number of people who produces a large number of articles or projects is decreasing smoothly as one could expect but, if we consider quality aspects, we find that, for low levels, the decay is smooth up to a certain point while after that it becomes much stronger. We can define this zone as the excellence region, and it represents the most relevant subset of researchers for the impact of the university from an international point of view. The change in the slope of the line is much more significant in the project curve, but it can also be seen in the publication one. We can conclude that the different scaling is associated to excellence and it is qualitatively similar in the paper and in the project cases.

3.3. The different areas of knowledge

In this section we will analyze different aspects of the publication and projects networks taking into account the different macroareas of knowledge in which the university is divided, namely:

- Science: traditional scientific disciplines as Chemistry, Physics, Mathematics, Geology

- Life Sciences (shortened as Health in the figures): Medicine, Veterinary Science, Biology, Biochemistry, etc

- Social Sciences and Law (shortened as Social): Psychology, Sociology, Economy, Law, Education, etc

- Arts and Humanities (shortened as Arts): Filologies, History, Philosophy, etc

- Engineering and Architecture (shortened as Engineering): all areas of Engineering and the relatively recent areas of Architecture (the Architecture degree is offered in our university since 2010)

- The researchers who are not administratively attached to any department in the university but belong to their staff, as some of the investigators working at the different research Institutes are include in a category called ”Others”. We will not use this category in the analysis of this section.

Researchers are considered to belong to a given area if the department to which he/she belong is contained in it. Thus, researchers belonging to the Department of Theoretical Physics, as some of the authors of this paper, are considered to belong to the macroarea ”Science”, while researchers of the department of Mechanical Engineering, belong to the macroarea ”Engineering and Architecture”.

Let us start showing the behavior of the different areas in what regards the publication of papers in journal indexed in JCR and its evolution. We can see in Fig. 4 how the total JCR impact of the publication of the different areas has evolved in the last years (period 2001-2017).

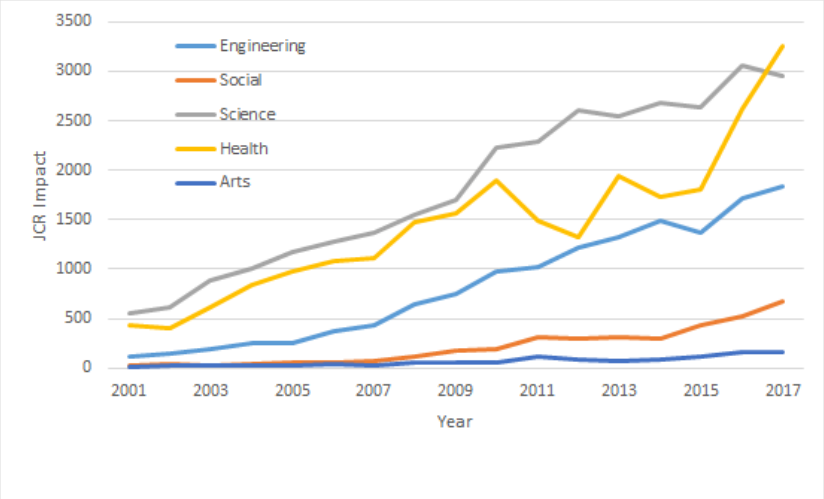

Figure 4: Evolution of the total JCR impact of the publications of the different areas

We can notice how the Science departments have been the traditional leaders of the JCR production but they have been recently surpassed by the Life Science departments with a very significant growth in the last three years. Social Sciences and Arts and Humanities stay at the bottom, since in these departments the publication scheme is more concentrated in different formats (books, national journals, etc) which are not included in JCR.

In Fig. 5 we present however the evolution of the number of articles, including both JCR and non-JCR papers, and now we can see how the differences between the distinct disciplines are much smaller.

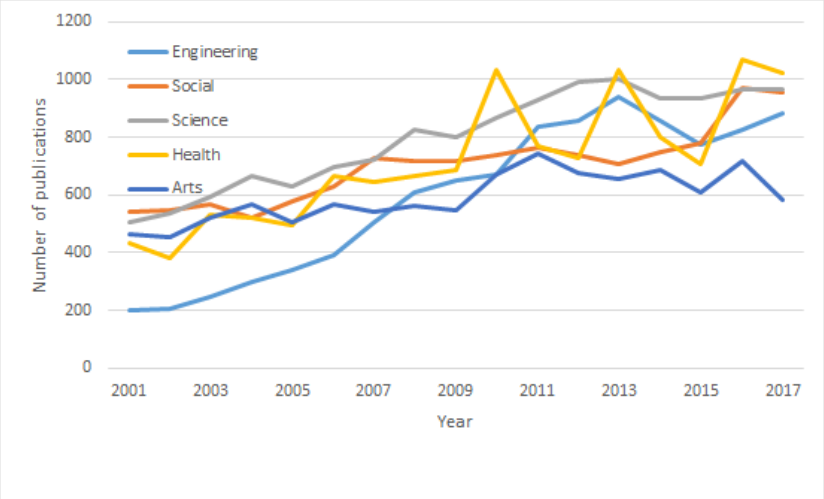

Figure 5: Evolution of the number of the publications (including non-JCR) of the different areas

With regard to the yearly evolution of the publications, we can see from those figures how it is a growing of the article number during the first years of the century and exhibits a plateau during the last 10 years, corresponding to the recent crisis. The Social Sciences and Law areas, on the contrary, have seen an important increase of JCR impact in the same period. This is probably due to a recent increment in the awareness of the researchers of these fields about the importance of publishing in indexed reviews, although it represents a quite particular issue in the Spanish academic system which may not have a direct translation to other countries. Indeed, the quality criteria approved by the National Agency for Quality Assessment and Accreditation of Spain, ANECA, include now the evaluation of ISI indexed publications for some of the subareas in the Humanities and Social Sciences fields. This may have affected the publication routines of the researchers of those areas.

If we consider the analogue of this problem in the project networks, we can plot the evolution of the funds obtained by the different groups. The result can be found in Fig. 6.

We see how the behavior coincides with the paper network in several aspects but now one can see how the project funds generation is dominated by the Engineering and Architecture area, mainly because of the bigger quantity of transfer activities developed by the Engineering departments for the industry. Here we can observe even better the effect of the economic crisis with a clear decrease of the funds after 2009, while on the last years we can appreciate some signs of recovering.

Another question that we have considered is how much the researchers of each area collaborate with members of the other areas of the university. In Fig. 7 we show the percentage of activity that is developed by publication co-authors that belong to the same area. In a similar way, in the same figure one can see the percentage of activity in projects that is generated internally to each of the areas. In both cases, we observe that,

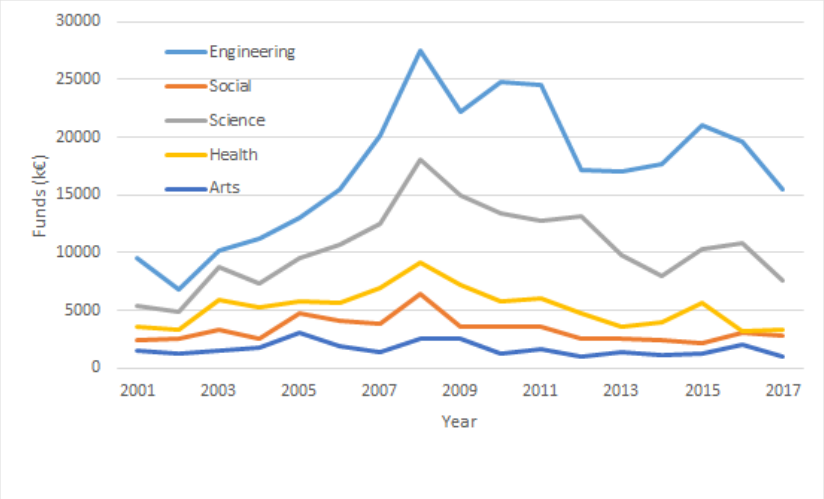

Figure 6: Evolution of the total amount of funds obtained by the different areas

in average, approximately 3/4 parts of the activity is realized internally, i.e. without the collaboration of members of other areas of the university. However, some differences can be appreciated between the behavior of the different areas and between the publication and project cases. From these criteria, the area with less external collaborations would be Social Sciences and Law, probably due to its own peculiarities, not needing the collaboration of other fields researchers to develop their projects or investigations. On the other hand, the dispersion of this indicator between areas is larger in the project case and it shows that the inter-area collaboration is bigger in the more technological areas, which in general have a larger necessity of technologies or modelization schemes from other fields than in the case of Humanities or Social Sciences.

A similar analysis can be made on other aspects of the network, as for instance the gender differences. We can consider the percentage of women co-authoring papers or being PI (Principal Investigator) of research projects in the different areas (Fig. 8). In the case of publications, the participation of women is similar to the percentage of women existing in each of the areas.

However, in the case of projects, the number of women being PI is significantly smaller, especially in the Engineering and in the Science departments. This can be due to the inertia of the past, because the majority of the elder professors in these areas are men. In the Social and Arts areas, this gap is smaller.

3.4. Topological differences of the networks

In this section we are going to analyze, from a quantitative point of view, the different networks introduced so far and study their topological properties. In order to do that let us consider the following parameters:

p1 The number of nodes in the network

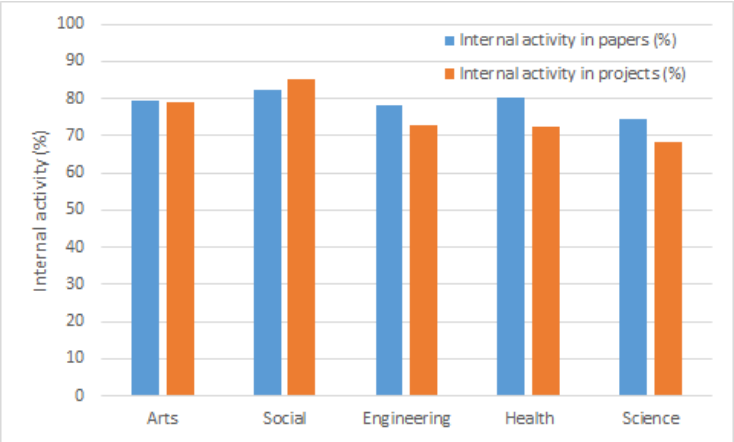

Figure 7: Percentage of internal coauthors and collaborators in the publications and projects of the different areas

p2 Clustering: Probability that two neighbours of a third party are at the same time neighbours of each other. To measure it, we use the notion of transitivity by Wasserman and Faust (1994), which computes the global clustering coefficient of the network.

p3 Assortativity: Preference for the nodes to attach to others that are similar in some way (in this case, having a similar number of contacts)

p4 Modularity: Presence of well-defined subgroups or clusters. The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random (Newman (2006))

p5 Giant cluster: number of nodes which belong to the *largest block*

p6 Average Path Length: the average distance between two nodes

p7 Diameter: the longest distance that can be made between two nodes

It is important to remark that the modularity of the network, the average path length and its diameter, are obtained considering only the giant cluster. We will come to this point later. We can summarize the properties of the three complete networks in Table 1. We see that the number of nodes is similar in all three cases, and therefore the comparison is reasonable. The slight differences arise from student or technicians that may be considered as part of the project but do not, necessarily, sign the research papers. In any case, the differences are below 7%.

From the topological point of view, we find remarkable differences:

- the clustering coefficients (p2) exhibit the largest difference between the paper network and any of the project networks, but also between the last ones. In-

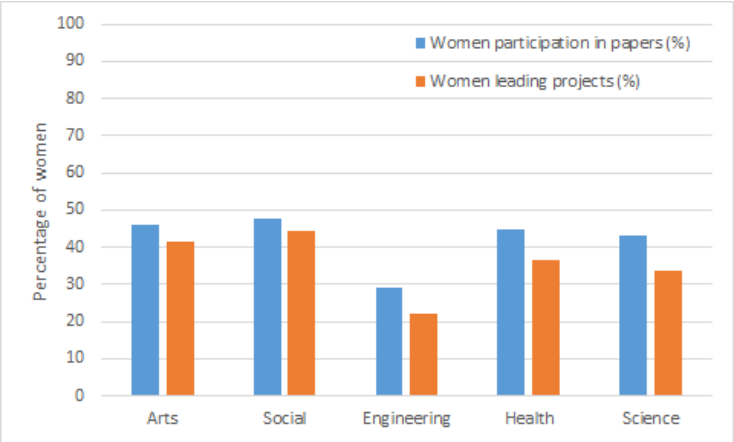

Figure 8: Percentage of women that are co-authors of publications or PI of projects in the different areas

deed, the paper network exhibit a clustering coefficient of 0.31, while the directed project network has a value of 0.13 and the un-directed one of 0.66. What do we learn from these numbers? Clearly, that the groups which collaborate for a project are much less promiscuous than those collaborating for a paper. Thus, if we consider the undirected graphs, if two researchers A, B are connected to a third one C, the probability of A, B, C to belong to the same project is much higher than the analogous situation signing a paper. The low value of the directed graph indicates the low probability of having more than one researcher acting as PI in a group. Combining both properties, we conclude that the groups formed to collaborate in a project are more stable than those created to publish a paper. Thus, when researchers collaborate to write papers together, they are more "promiscuous" and tend to collaborate more often outside their groups.

- the assortativity coefficients (p3) also reflect the above information, but not with the same intensity. Indeed, if we compare the undirected networks we see how the preference of the nodes to attach to nodes similar to themselves is higher in the project case. Thus, we can conclude that most nodes in the project network are connected only (85%) to those of their (stable) group, with whom they collaborate. In the case of the paper network the preference is only of 62%, what implies that they sign papers with people outside their groups (otherwise, they would be linked to nodes similar to themselves and the value would be higher). In the directed project network, the value is even smaller.

- Coefficient p4 (modularity) exhibits a similar behavior, but with less intensity. Thus, the project networks have larger values of the modularity coefficients indicate also that those networks have better defined subgroups, corresponding to a stable structure of researchers collaborating basically among them and less significantly with the rest of researchers.

- The Giant cluster of the networks (coefficient p5) also exhibit an important dif-

ference of the collaboration between the paper network and the project networks: even if large in both cases, the paper network contains a much larger set of isolated researchers (approx. 10%), who are outside the giant cluster and disconnected from the rest of the university; while in the project networks the quantity of these researchers is almost negligible (approx. 1%).

- Finally, the values of coefficients p6 and p7 allow for a comparison of the long-range properties. Comparing the undirected networks we notice how the project network is "more dense" than the paper one, with nodes closer (in average) to each other and smaller longest distances. Nonetheless, the much larger distances of the directed project network allows us to conclude that the larger density is the result of a much higher density of interconnections inside the groups (since their nodes have less connections with the nodes of other groups than in the case of the publication network) and not of a higher collaboration between groups. Indeed, if this was the case, the distances of the directed network would not have increase significantly with respect to the publication network distances.

	Paper Network	Directed Project Network	Undirected Project Network
p1	6336	6768	6778
p2	0.31	0.13	0.66
p3	0.62	0.48	0.85
p4	0.64	0.80	0.7
p5	5695 (89.9%)	6681(98.7%)	6690 (98.7%)
p6	5.01	7.87	3.94
p7	10	16	8

Table 1: Values for the different parameters of the graphs of the complete networks

3.5. Topological differences of the subnetworks of the different areas of knowledge

Do these properties depend on the area of knowledge? From the corresponding subnetworks of researchers in the different areas, we extract the following results:

3.5.1. Paper networks

The data in this case is presented in Table 2. We see important differences in several aspects. The subnetworks of Art and Humanities is clearly the less hierarchical, it has the highest percentage of nodes which are similar to the others (see parameter p3, assortativity). The subnetwork of Life Sciences is the opposite case, it is where we find the smallest percentage of similar nodes (p3) and the lowest clustering coefficient (p2). The case of the modularity is interesting, but more difficult to study. If we look at the different networks in a graphical representation, we see remarkable differences between the Arts and Humanities case (Figure 9) and, for instance, the Science (Figure 10) and the Life Sciences (Figure 11) cases. We see that the Arts and Humanities network has a very small giant cluster (where the modularity is computed), compared to the other two. Therefore, when computing the modularity of the networks, the numerical value

	Sci.	Arts	Life Sci.	Soc. Sci.	Eng.	Others	Complete
p1	1149	740	1406	1050	1066	1103	6336
p2	0.4	0.58	0.3	0.43	0.37	0.5	0.31
p3	0.65	0.84	0.58	0.74	0.61	0.83	0.62
p4	0.75	0.44	0.58	0.51	0.71	0.85	0.64
p5	1050	454	1271	787	993	258	5695
p6	5.09	6.06	3.72	5.28	4.78	7.65	5.01
p7	13	19	9	14	13	21	10

Table 2: Values for the different parameters of the graphs of the paper networks

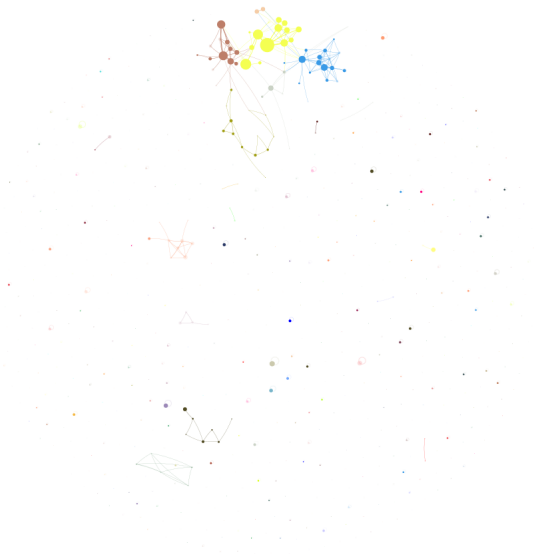

Figure 9: Graphical representation of the paper network for the Arts and Humanities subnetwork. The small rank of some nodes makes impossible to represent them, but it still provides a good representation of the structure of the network.

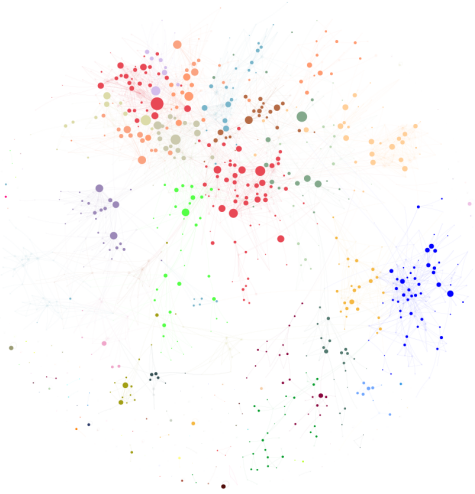

Figure 10: Graphical representation of the paper network for the Sciences subnetwork. The small rank of some nodes makes impossible to represent them, but it still provides a good representation of the structure of the network.

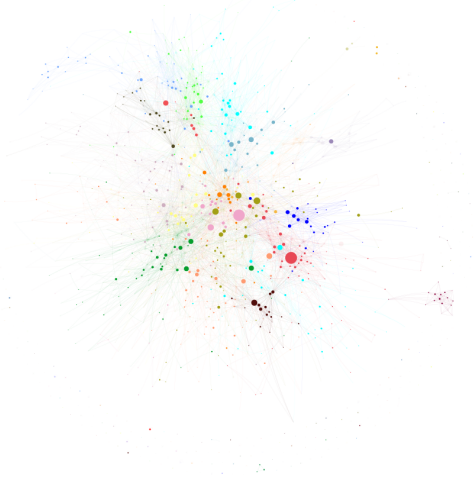

Figure 11: Graphical representation of the paper network for the Life Sciences subnetwork. The small rank of some nodes makes impossible to represent them, but it still provides a good representation of the structure of the network.

	Sci.	Arts	Life Sci.	Soc. Sci.	Eng.	Others	Complete
p1	1176	739	897	1026	1100	183	6768
p2	0.21	0.17	0.18	0.19	0.21	0.04	0.13
p3	0.48	0.44	0.47	0.51	0. 48	0.5	0.48
p4	0.82	0.85	0.8	0.76	0.75	0.91	0.80
p5	1148	713	838	981	1076	31	6681
p6	7.92	9.02	7.39	6.27	5.58	1.11	7.87
p7	14	20	12	12	12	7	16

Table 3: Values for the different parameters of the graphs of the directed project networks

for the Arts and Humanities subnetwork is higher than the value for the other two, while it is clear that the topology of the networks is much more modular in the other two cases. In the Life Sciences case we can easily identify the higher interconnection of the different communities. This high interconnection makes the Life Science area also the smallest in diameter compared to the other four (parameter p7). With respect to the total network, the area of Sciences is the one which better represents it with respect to most criteria.

It is also interesting that the low p7 value of the complete network suggests that there exist connections between groups of different macroareas, since it is smaller than any of the diameters of most of the subnetworks (excepting Life Sciences, which has only 15% of the nodes of the total network).

3.5.2. Directed Project network

The data in this case is presented in Table 3.

In this case we notice that all networks are much less homogeneous than their publications counterparts, and much more hierarchical. Modularities and diameters are similar to the publication networks (and to each other, excepting the large diameter of the Arts case), but the separation of nodes is larger (parameter p6). The total project network summarizes well the characteristics of all their subnetworks, excepting the diameter because of the lack of uniformity among the different areas. We conclude thus that the project network is more hierarchical in general, the networks are much less homogeneous, but on the other hand all the different areas are much more similar to each other than in the publication case.

Again, we see that the area of Arts and Humanities exhibits the most extreme values of the different areas: the lowest values for p2 and p3 and largest p4, p6 and p7. These numbers suggest a highly modular network, with few interconnections between the different subgroups and hence with large distances (in average) between nodes and large diameter. Again Science is the macroarea closest to the behavior of the total set, with the exception of the p2 factor, which is far larger than the total one.

3.5.3. Undirected Project network

The data in this case is presented in Table 4.

In this case we notice that all networks are much more homogeneous than in the previous cases. All areas are similar excepting again the p6 and p7 values of Arts and

	Sci.	Arts	Life Sci.	Soc. Sci.	Eng.	Others	Complete
p1	1205	766	970	1084	1136	1781	6778
p2	0.67	0.67	0.75	0.56	0.63	0.72	0.66
p3	0.79	0.73	0.88	0.66	0.72	0.79	0.85
p4	0.67	0.82	0.53	0.71	0.72	0.89	0.7
p5	1167	737	908	1029	1109	1082	6690
p6	3.53	5.01	3.58	6.39	3.26	6.47	3.94
p7	9	18	9	9	10	18	8

Table 4: Values for the different parameters of the graphs of the undirected project networks

Humanities, whose origin is the same as the previous one (independent subgroups with a large number of inner connections but few connections between groups). It is again interesting, though, that the lower p7 value of the complete network suggests that there exist connections between groups of different macroareas, since it is smaller than any of the diameters of the subnetworks.

4. Conclusions

We have studied academic networks formed by Papers or Projects as a Complex Networks system, focusing not only in individual but also in global, collaborative properties, using the data from the University of Zaragoza.

Regarding our first question in the introduction, as a general result, we have observed that the *laws* governing the individual productivity are similar for the cases of publications and projects but, however, the behavior is different when analyzing more complex magnitudes such as the collaborations or other variables depending in a structural way of the idiosyncrasy of the university. In particular, we have seen that the number of researchers reaching a certain level of production scales in a similar way for publications and projects. A relevant result that has appeared in the study of this scaling is that one finds two different behaviors if one considers production with or without a quality metric. When we do not use quality (taking into account only the number of publications or the number of projects), the number of people producing more than a certain quantity decreases smoothly. However, if we consider quality aspects (using JCR impact or project funds), the decay is smooth up to a certain point while, after that, it becomes much faster. We found in this case two different regions, the *low production* and the *large production* ones. We see a *barrier* between both of them, in the sense that the expected number of researchers in the large production zone is much smaller than what one would expect with a normal extrapolation from the low production region. We can conclude that this different scaling is associated to excellence and it is qualitatively similar in the paper and in the project cases.

Regarding the second question, when studying the article and project networks of collaboration, one can see a different structural behavior. In this paper we have introduced two different types of project networks in order to capture different properties of the set: the directed and the undirected one.

From our analysis we have identified important structural differences between the collaboration patterns for papers and for projects. Thus, when researchers collaborate to write papers together, they are more "promiscuous" and tend to collaborate more often outside their groups. The groups collaborating in a research project tend, on the other hand, to be more stable and to have less connections between groups. Hence, we can claim that the article network is more symmetric and homogeneous, and that the collaboration of researchers in papers is more *democratic* than the relation defined by the projects. In the project networks, we have a more hierarchical structure where *important* people share projects with *less important* people with little probabilty for PI-researchers to become regular researchers under another researcher lead. Notice that in order to obtain this conclusion, it is necessary to analyze, at the same time, both project networks, since the information extracted from them is complementary.

In what regards the third question, we have presented several properties of the system and their distributions among the different areas of knowledge. In our analysis, we have also been able to identify the evolution of papers and funds in time for all areas, the global behavior of the different areas of knowledge in the university and the areas with higher contributions in publications and projects. We saw how Arts and Humanities on one side and Life Science on the other become the most extreme cases for the areas. Collaboration in the Arts and Humanities happens in small and stable groups with few interconnections with others. In Life Sciences, on the other hand, collaboration is closer and networks become *more compact*. Again, from the complementarity of both project networks we can conclude that there exists collaboration between different macroareas, which make the distances for the global network smaller (on average) than those on some of (or even all) its subareas.

With the same tools, we have been able to analyze the impact of external collaborators on them or how gender differences are present in different degrees in each of the areas. Thus, we have seen for example how the Humanities or Social Science areas are the less likely to have external collaborations but, on the other hand, they are the less biased with respect to having women as head of a project.

Despite the fact that the study has been made on the data of the University of Zaragoza, we think that, even if the results could be quantitatively different, most of the conclusions could be applicable to other universities or research centers due to our general approach.

Acknowledgements

This paper has benefited from funding from the Spanish Ministry of Economics, projects FIS2015-65078-C2-2-P, MTM2015-64166-C2-1-P and FIS2013-46159-C3-2-P

References

Abreu M and Grinevich V, (2009) Gender patterns in academic entrepreneurship. Journal of Technology Transfer 42(4): 763-794

Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F (2009) h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics 3:273–289

Álvarez R, Cahué E, Clemente-Gallardo J, Ferrer A, Íñiguez D, Mellado X, Rivero A, Ruiz G, Sanz F, Serrano E, Tarancón a, Vergara Y (2015) Analysis of academic productivity based on Complex Networks. Scientometrics 104(3):651–672

Barabâsi A, Jeong H, Néda Z (2002) Evolution of the social network of scientific collaborations. Physica A: Statistical ... 311:590–614

Batagelj V and Cerinšek M (2013) On bibliographic networks. Scientometrics 96(3): 845-864

Bellotti, E. (2012). Getting funded. Multi-level network of physicists in Italy. Social Networks, 34(2), 215–229.

Bordons M, Barrigón S (1992) Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984–89). Part II. Scientometrics 25(3):425–446

Clauset A, Shalizi CR, Newman MEJ (2009) Power law distributions in empirical data. SIAM Review 51(4):661–703

Costas R, Bordons M (2007) The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics 1(3):193–203

Grauwin S, Jensen P (2011) Mapping scientific institutions. Scientometrics 89(3):943–954

Katz J and Martin B (1997), What is research collaboration?, Research Policy 26(1): 1-18

Kumar S (2015) Co-authorship networks: a review of the literature. Aslib Journal of Information Management 67(1): 55-73

Lemes Alarcão, A. L., and Sacomano Neto, M. (2016) Actor centrality in Network Projects and scientific performance: an exploratory study. RAI Revista de Administração E Inovação, 13(2), 78–88.

Maggioni MA, Uberti TE, Nosvelli M (2014) Does intentional mean hierarchical? Knowledge flows and innovative performance of European regions. Annals of Regional Science 53(2):453–485

Miguel, S., Chinchilla-Rodríguez, Z., González, C., Moya Anegón, F. (2012) Analysis and visualization of the dynamics of research groups in terms of projects and coauthored publications. A case study of library and information science in Argentina, Information Research 17(3) paper 524

Newman MEJ, Modularity and community structure in networks (2006) Proc. Nat. Acad. Sciences 103(23):8577-8582

Newman MEJ, Finding community structure in networks using the eigenvectors of matrices (2006b) Physical Review E 74(3):036104

Newman MEJ (2010) Networks: An introduction. Oxford University Press

Pons, P and Latapy, M (2006) Computing communities in large networks using random walks, Journal of Graph Algorithms and Applications (JGAA) 10(2):191-218

Van Eck NJ, Waltman L (2014) CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics 8 (2014) 802–823

Wallace DL (1983) Comment to "A Method for Comparing Two Hierarchical Clusterings". J American Stat Associ 78(383):569–576

Wasserman S and Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press