ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/291086304

Analysis of Co-Authorship Ego Networks

Conference Paper · January 2016

DOI: 10.1007/978-3-319-28361-6_7

CITATIONS
37

READS
730

4 authors:



Valerio Arnaboldi
California Institute of Technology
50 PUBLICATIONS **2,244** CITATIONS
SEE PROFILE



Robin Dunbar
University of Oxford
782 PUBLICATIONS **59,019** CITATIONS
SEE PROFILE



Andrea Passarella
Italian National Research Council
326 PUBLICATIONS **11,356** CITATIONS
SEE PROFILE

Marco Conti
Italian National Research Council
363 PUBLICATIONS **12,884** CITATIONS
SEE PROFILE

All content following this page was uploaded by Valerio Arnaboldi on 08 November 2017.

The user has requested enhancement of the downloaded file.

Analysis of Co-Authorship Ego Networks*

Valerio Arnaboldi[1], Robin I. M. Dunbar[2,3], Andrea Passarella[1], and Marco Conti[1]

[1] IIT-CNR, Via Moruzzi 1 - 56124, Pisa. {name.surname}@iit.cnr.it
[2] Department of Experimental Psychology, University of Oxford, South Parks Road - Oxford OX1 3UD, United Kingdom. robin.dunbar@psy.ox.ac.uk
[3] Department of Information and Computer Science, Aalto University School of Science, Konemiehentie 2 - FI-02150, Espoo, Finland

Abstract. The availability of co-authorship data from large-scale electronic databases is paving the way for new analyses on human collaboration networks. The complex network of co-authorships can identify specific features that characterise the behaviour of researchers, and impact on their production and performance. In this paper, we analyse a large sample of data regarding scientific publications from Google Scholar. The aim of our analysis is to study a fundamental aspect of co-authorship networks, i.e. the structure of authors' ego networks. Specifically, we highlight the existence of a hierarchical organisation of these networks in a series of concentric circles, quite similar to that found in general human social networks. In addition, we highlight some properties of the correlation between the ego network structure and the authors scientific productivity, measured in terms of h-index.

Keywords: Collaboration Networks, Ego Networks, Optimal Team Size, Scientific Productivity, h-index

1 Introduction

Co-authorship networks represent the patterns of human collaborations in the production of scientific knowledge. The analysis of these networks is gaining momentum, due to the increasing availability of co-authorship data from electronic databases like Google Scholar, Scopus or Microsoft Academic Search, and to the gradual shift in science from an individual based model (i.e. where individuals have the main role in the production of knowledge) to a teamwork model [18].

In this paper, we study the structural properties of a large number of co-authorship *ego networks* extracted from Google Scholar. In this context, the ego network of a given author is the network formed by linking that author with all its co-authors, and weighing each link with a measure of strength of their collaboration. The main goal of this work is understanding whether similar structures found in general human ego networks are also present in co-authorship

* This work was partially funded by the EC under the H2020-INFRAIA SoBigData (654024) project.

ego networks. The main intuition is that the structure of human ego networks is determined by cognitive and time constraints that limit the number and intensity of social relationships each person can maintain. Scientific collaboration can be seen as one specific type of human social relationship, which is in principle also affected by cognitive and time constraints. Therefore, we ask ourselves if those constraints result in similar structures also in the co-authorship networks.

The structure of ego networks is a well-investigated topic in the area of social networking. The effect of cognitive and time constraints leads, in human social networks, to the formation of a typical hierarchical structure around each ego. According to this structure, *alters* (i.e. people with whom the individual or ego is connected) are arranged into inclusive groups. The size of these groups averages 1.5, 5, 15, 50, 150 respectively, with a scaling ratio of $\sim$3 [6,19]. In this hierarchy, Dunbar's Number (150) represents the average limit on the total number of social contacts that humans can actively maintain in their networks [5]. The structural pattern found in ego networks is consistent among different social environments, including online social networks (OSNs) such as Facebook and Twitter [6], and other types of social organisation, such as the land armies of many countries [19]. This indicates a natural hierarchical grouping of everyday social structures, optimising the cognitive processing of within-group interactions [19].

Although the properties of ego networks have been largely studied in the literature, there are still no conclusive results about the relation between the hierarchy found in ego networks and the collaboration strategies in organisations and team work in general. To the best of our knowledge, there are no studies that characterise the hierarchical structure of ego networks in collaboration environments and that analyse the relation between this structure and scientific productivity. In addition, most of the experiments on collaboration networks are related to small-sized samples, often in controlled and limited environments.

In this paper, we aim to bridge this gap by analysing the structural properties of a large number (313, 207) of co-authorship ego networks extracted from Google Scholar. We find that a hierarchical structure similar to that found in other types of human social ego networks is also present in co-authorship ego networks, suggesting that human cognitive and time constraints may play a significant role in shaping collaborations. In addition, we start to characterise the impact of these structures on scientific productivity. We study the correlation between the structural properties of the authors' ego network and their h-index. In particular, we find a significant correlation between the ego network size and the h-index, and we briefly discuss the main reasons that could explain this property.

2 Related Work

One of the most significant contributions to the analysis of co-authorship networks is the work on large-scale collaboration databases by Mark E.J. Newman (see for example [12, 13]). Newman applied standard social network analysis tools to complete co-authorship graphs (not limited to ego networks), and his results are the first indication that co-authorship networks have high clustering coefficient and small average distance between pairs of nodes, thus being "small-

world" networks. In addition, he found that statistics of individual authors can vary significantly in different fields.

Katz and Hicks [10] found that the number of co-authors has a positive impact on the number of citations of papers. Moreover, collaborations with foreign institutions bring to the highest impact. In a similar study, Guimerà and colleagues found that team "diversity" is one of the key aspects of success [8].

Abbasi and colleagues [1] analysed data related to five American Universities, by acquiring information regarding scientific publications from Google Scholar, ACM portal, and DBLP. They built complete co-authorship networks and calculated standard centrality measures (e.g. node degree, closeness, betweenness) for each node (author). They found that centrality measures have a high correlation with productivity (g-index). Moreover, authors embedded in clusterised parts of the network have lower performances than other authors.

Abbasi et al. also analysed ∼8,000 co-authorship ego networks from the Scopus database [2]. The results show that several ego network measures (e.g. density, Burt's constraint, ego-betweenness, effective size) correlate with author performances (g-index). This is in accordance with the results of the analysis by Ortega [14] on a large dataset collected from Microsoft Academic Search (MSA) with ∼32,000 authors, which also highlights differences in the properties of co-authorship ego networks in different research fields. These results also confirm similar findings on smaller co-authorship datasets [11].

Compared to these studies, our work is focused on the detailed analysis of the structural properties of ego networks, and their impact on scientific productivity.

3 Co-Authorship Data

3.1 Data Statistics

We analysed 313,207 author profiles accessed from Google Scholar in November 2013. For each author, we have information about the categories (research areas) manually indicated in their personal profiles as one or more freely assigned labels (for a total of 188,657 categories), and all the publications indexed by Scholar (for a total of 19,420,220 publications, of which 378,305 are patents). In addition, we have all the statistics calculated by Scholar, e.g. the total number of citations for each author and the h-index. We accessed the authors' profiles starting from a single category ("computer science"). Then, we accessed all the categories found in the visited profiles, iterating the procedure until no new categories were found. The distribution of the category size, depicted in Fig. 1a, clearly shows a power-law shape. The largest category, "machine learning", has been indicated by 8,122 authors. The distribution of the number of citations of the authors and that of their h-index are depicted in Fig. 1d and Fig. 1b respectively, while the distribution of the number of citations of the papers in our sample is depicted in Fig. 1c. The number of co-authors in the papers ranges between 1 and 150. The limit of 150 is imposed by Google Scholar. Papers with more than 150 co-authors are limited to 150. Nevertheless, the number of papers with 150 co-authors is only 6.5% of the total number of papers.

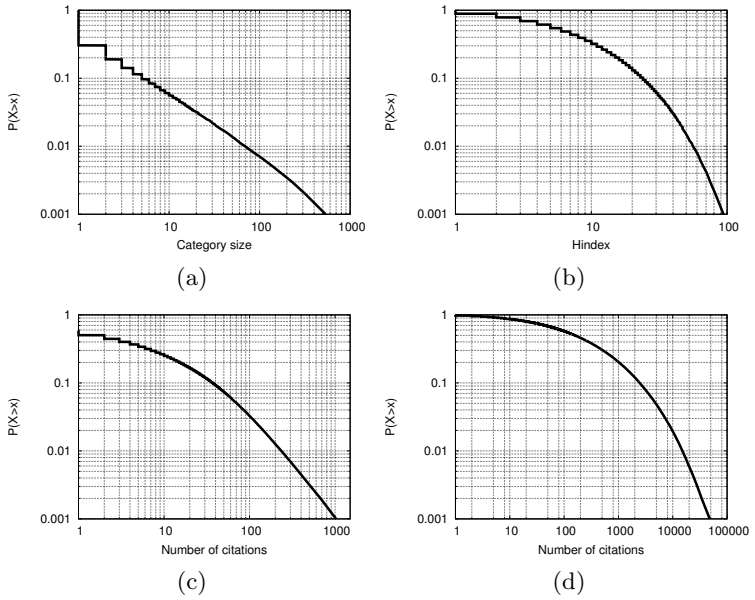


Fig. 1: CCDFs of category size (1a), author h-index (1b), number of citations per paper (1c), and number of citations per author (1d)

3.2 Data Preprocessing

For the analysis presented in the rest of the paper, it is important to know the dates of the publications. Scholar derives them from the hosting web site of the paper or from the pdf file directly. For some papers, dates are wrongly read from fields related to author's personal information (e.g. the zip code of the author's affiliation). These cases, although rare, usually lead to inconsistent dates or a date far from the other publication dates of the author and are thus easy to be identified and corrected. We discarded dates in the future (i.e. after the date when the paper's data were accessed) and dates too far from the rest of the publication dates of the author. Specifically, for each author, we ordered its publication dates in a descending order and we calculated the difference between subsequent dates. We formed groups by putting together all dates such that the distance between two consecutive dates is lower than 20 years, and we consider as relevant publications of an author only those belonging to its largest group. Therefore, the first date (in temporal order) of this group is taken as the first publication date of the author. Publications with dates that are outside the obtained group are discarded from the analysis. The threshold of 20 years is chosen intuitively as the maximum period of inactivity of an author between two subsequent publications.

We further refined the data by selecting only the authors who published their first paper after the 1^{st} January 1900. This ensures that the analysed data are more homogeneous. Moreover, only authors who published for the first time at

Table 1: Properties of the first six clusters of categories found in Google Scholar ordered by number of authors.

Main Category	First 6 Sub-categories (by size)	n. of Authors	n. of sub-categories
machine learning	artificial intelligence, computer vision, data mining, image processing, robotics, software engineering	81,409	23,615
physics	nanotechnology, optimization, biochemistry, biophysics, chemistry, materials science	80,491	27,666
neuroscience	economics, psychology, education, innovation, cognitive neuroscience, enterpreneurship	59,609	24,219
bioinformatics	computational biology, genomics, molecular biology, evolution, genetics, conservation biology	59,428	19,802
ecology	climate change, remote sensing, gis, gis&t, hydrology, geology	47,519	21,965
molecular biology	microbiology, medicine, hiv aids, biotechnology, immunology, epidemiology	20,431	19,738

least three years before November 2013 were considered for the analysis, since after a three year period (coinciding with the duration of a Ph.D in most countries) the publication history of an author can be considered stable, whereas authors with a publication history shorter than three years could be still in a transient phase of their scientific career. After this preprocessing phase, we selected 285,577 authors. These authors represent the final sample of our analysis.

4 Clusters of categories

To better analyse co-authorship ego networks, we divided authors into separate research fields. To do so, we clustered them into different groups according to the categories that they declared in their profiles. Note that in Google Scholar categories are simple free-text strings that users associate to their profile. Therefore, it is not possible to simply group users by the same category. Instead, we applied a clustering algorithm on the bipartite graph G representing the relations between authors and categories. More formally, we defined $G = (A, C, E)$ as a bipartite graph where A is the set of authors, C is the set of categories, and E is the set of edges connecting elements of A to elements in C. An edge between an author $a \in A$ and category $c \in C$ exists if a has declared c in her profile. We applied a community detection algorithm (a greedy algorithm based on network modularity for bipartite graphs [4]) on G to group together similar categories. We obtained a set of 6,779 clusters of categories. Table 1 reports the properties of the first six clusters ordered by total number of authors (i.e. the sum of the number of authors of all the categories in the cluster). In the table, the main category of each cluster (i.e. the one with the highest number of authors) is placed along with the first six sub-categories (ordered by number of authors), the total number of authors in the cluster, and the number of sub-categories.

The results reported in the table confirm that the clustering algorithm is effectively able to group together categories belonging to the same research field.

Another important feature of the clustering algorithm is that it is able to correctly place duplicates of the same category in the same cluster (e.g. categories with small differences in the naming such as *Consumer Behavior* and *Consumer Behaviour*). This is possible since most of the authors indicating one of the two names in their profiles also indicate other categories that are in common with the other label, and this permits to identify a strong overlap between the two.

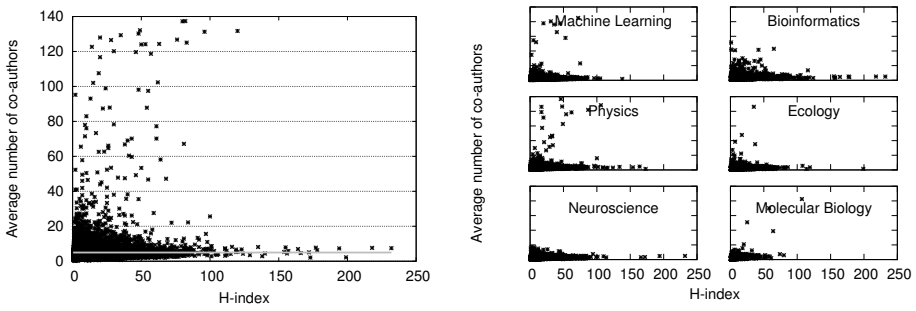
In some cases, clusters group together topics that would seem to be not that overlapping. This is the case for example of "economics" and "entrepreneurship", that are grouped in a neuroscience/psychology cluster. This is likely an effect of strong inter-disciplinary relationships among those topics. In other cases, the main category seems, at a first glance, not the best description of an area. This is the case, for example, for the "machine learning" cluster, that also contains "computer science". This is perhaps not so counterintuitive, as computer science is a very broad field and computer scientists may prefer to indicate more specific categories in their profile rather than a generic one.

We give here a more detailed description of the dataset leveraging the clusters that we obtained. In particular, we look at the relation between the h-index of the authors in the different clusters and the properties derived form their profile (i.e. the average number of co-authors per paper and their index of multidisciplinarity – defined in the following). To be able to perform a detailed analysis of the different clusters, we only considered the first six clusters in terms of size (number of authors) and we omitted the remaining clusters.

4.1 H-index vs. Average Number of Co-Authors per Paper

Figure 2a depicts the average number of co-authors per paper (considering all the authors) as a function of their h-index. The average number of co-authors per paper is distributed around 5 (the horizontal line in Fig. 2a), with higher variability for low values of h-index. Nevertheless, there are some authors with a very high average number of co-authors per paper with high h-index (up to 120). This could be due to the presence of authors working in fields where publications with many co-authors are common. When dividing the authors into the different clusters (Fig. 2b), it is clear that for some clusters (e.g. neuroscience) this effect disappears.

This first result tells us that although scientists may have a large number of collaborations, they tend to work in groups of 6 members (5 plus the ego). This is particularly true for authors with very high performance (i.e. with h-index $>$ 120). Notably, the result is consistent for all the clusters. This is in accordance with the results in the organisational literature, which identify 6 as the best team size for productivity [9]. Nevertheless, having an optimal team size is clearly not enough to perform well in scientific publications, as demonstrated by the large number of authors working on average with other 5 co-authors, but having a low value of h-index.



(a) (b)

Fig. 2: H-index as a function of the average number of co-authors for all the authors (2a) and the first six clusters (2b).

4.2 H-index and Mono- or Multi-Disciplinarity

We calculated a measure of multi-disciplinarity for each author as the similarity between the categories that each author declared in her Google Scholar profile. Specifically, we defined $O(C_i, C_j)$, the overlap coefficient between category C_i and category C_j as follows.

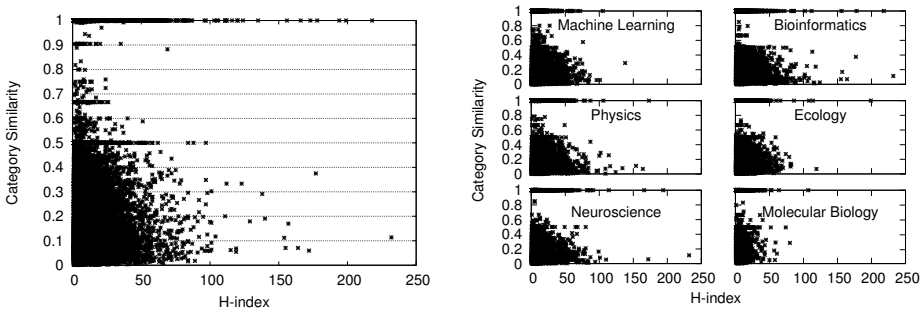
$$O(C_i, C_j) = \frac{|C_i \cap C_j|}{min(|C_i|, |C_j|)} \ . \tag{1}$$

Then, we averaged the overlap for all the possible pairwise combinations of categories in each profile to obtain a measure of multidisciplinarity m_a, for each author a, defined as follows.

$$m_a = \frac{\sum_{\{C_i, C_j\} \in [S_a]^2} O(C_i, C_j)}{|[S_a]^2|} \ , \tag{2}$$

where S_a is the set of categories of author a, and $[S_a]^2$ is the set of 2-subsets of S_a. Authors with high values of m work in categories tightly connected to each other that share a high percentage of authors. This indicates research fields very close to each other or different terms used to describe the same field. On the other hand, a low value of m indicates that the author works in research fields that are far from each other, with a very low number of authors working in both fields. The authors in categories with very low values of similarity are often the few people working at the same time in those fields, thus representing bridges between them.

Figure 3a depicts the h-index of the authors as a function of the average similarity of their categories. Figure 3b depicts the same scatterplot but considering the first six clusters of categories separately. The figures show a more random pattern for low values of h-index than for high values. Interestingly, in the latter case, two strategies seem to prevail, and authors tend to: (i) be completely focused on one subject or (ii) spread the effort on subjects that are completely separate from each other. Other mid-way strategies are not adopted by the most



(a) (b)

Fig. 3: H-index as a function of the average category similarity for all the authors (3a) and the first six clusters (3b).

successful authors, and are only used by authors with lower h-index. This is visible also when the analysis is applied to the different clusters, with compatible patterns across the different fields. Of course, this aspect needs to be further investigated in the future to draw definitive conclusions.

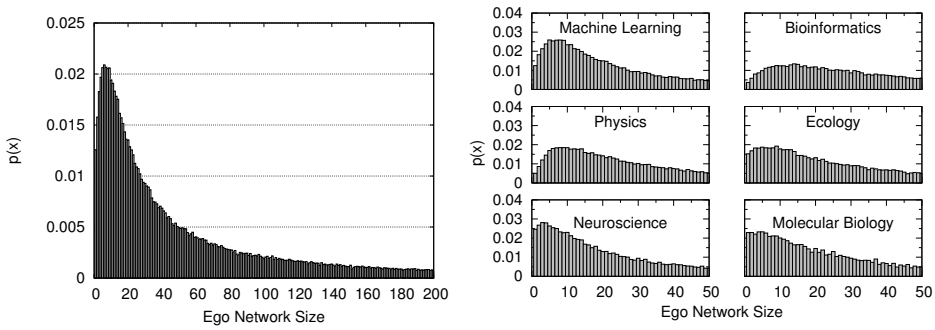
5 The Structure of Co-Authorship Ego Networks

5.1 Definitions

In a co-authorship ego network, an author is the *ego*, and the people with whom she co-authored at least a paper are her *alters*. Note that we use the most simple definition of ego network, without considering mutual connections between alters. To extract the set of alters from the metadata of the articles in Scholar, we looked at the co-authors in the papers of each author and we performed a lexicographic match of their names. Thus, for a given author, we cannot discriminate possible different co-authors with the same name, that are considered as one co-author only. However, intuitively the probability of this case should be very low. In different articles, the same authors may appear with slightly different names or abbreviations. To find these differences, we matched the names using either the first name and the surname or the first letter of the first name and the surname, case insensitive. Then, we defined the strength of a collaboration relationship between an ego i and one of her alters j, $ts_{i,j}$, as follows.

$$ts_{i,j} = \frac{1}{d_{i,j}} \sum_{p \in \{P(i) \cap P(j)\}} \frac{1}{k(p) - 1} \ , \tag{3}$$

where $d_{i,j}$ is the duration of the relationship between i and j, measured as the time span (in years, between the first article that the ego and her alter published together and the last article published by the ego), $P(i)$ is the set of papers where i appears as a co-author, and $k(p)$ is the number of co-authors of paper p. Intuitively, this measure of tie strength is a frequency of co-authorship, measured as the number of joint publications, where each publication is "weighted" by the



(a) (b)

Fig. 4: Density of ego network size for all the authors (4a) and the first six clusters (4b).

number of co-authors to estimate the level of collaboration between the ego and a specific alter in that collaboration. Being a frequency, the total number of (weighted) publications is divided by the number of years during which the ego and the alter are supposed to have collaborated. This definition of tie strength embodies the fact that a small number of co-authors generally implies a stronger collaboration between these than the case of high number of co-authors. We considered only the relationships with duration greater than six months to avoid possible bias due to approximation on too short relationships. Finally, for each author, we define the ego network size as the number of its co-authors. Note that, because we have defined a tie strength on each link, ego networks are *weighted* networks, which is the basis for analysing their structure.

5.2 Ego Network Size

The average ego network size is 104.01. Remarkably, this value is not too far from Dunbar's Number, and compatible with other sizes found in offline [19] and online environments [6,7]. The probability density function of the ego network size, depicted in Fig. 4a, indicates that, although there are many large ego networks (note that we have limited the x axis, and the maximum size is 12,070) the highest density is around 5. Moreover, the density function resembles in its shape the function of the active network size found in social networks [3]. Figure 4b depicts the density function of the ego network size for the first six largest clusters of categories. The size varies sensibly among the clusters, with "machine learning" and "neuroscience" showing a more picked density than "bioinformatics" and "physics", in which the size is more uniformly distributed. This difference is probably due to the fact that in machine learning and neuroscience publications are typically done by a small group of authors, while in fields such as physics and bioinformatics, where sometimes large experiments involving a high number of people are needed, publications with many co-authors are more common.

5.3 Average Tie Strength vs. Ego Network Size

The average tie strength of ego networks is negatively correlated with ego network size ($r = -0.23$ considering all the authors, $r = -0.60$ when the variables are log-transformed). This is visible also when the clusters are analysed separately ($r = -0.20$ and $r = -0.60$ for machine learning without and with log-transformation respectively, $r = -0.24$ and $r = -0.62$ for physics, $r = -0.23$ and $r = -0.59$ for neuroscience, $r = -0.26$ and $r = -0.59$ for bioinformatics, $r = -0.22$ and $r = -0.53$ for ecology, and $r = -0.25$ and $r = -0.6$ for molecular biology). Interestingly, correlation values are quite homogeneous across research fields. This suggests that this property may depend on some general mechanism underpinning collaboration, rather than on the specific ways of collaborations in different fields, which may vary from topic to topic (see Figure 4(b)). Indeed, this result is in accordance with the general findings on social ego networks and in organisational studies, that show that the constraints on our social behaviour limit the number of alters that we can maintain at a certain level of intimacy, and people with larger ego networks must inevitably have, on average, relationships with lower levels of intensity [15].

Taken together, these results (distribution of ego network size, and correlation between ego network size and tie strength) provide a first indication that cognitive mechanisms similar to those governing typical social relationships may also determine the development of scientific collaborations. In particular, the negative correlation between tie strength and ego network size is a classical "capacity shrinkage" effect: as the total amount of cognitive resources we can allocate to scientific collaboration is limited, the more intense collaboration are for an author, the lower is the number of active collaboration they can sustain.

5.4 Ego Network Circles

Remember from Section 1 that one of the fingerprints of the general structure of ego networks is the presence of concentric layers, of a very well defined average size, containing alters with whom the ego maintains social relationships at a decreasing intensity. To further investigate possible similarities between generic ego networks in social environments and in scientific collaborations, we analysed whether such a structure can also be identified in co-authorship ego networks. Previous studies in other social environments (such as Online Social Networks) have shown that the typical number of layers is 5, with average sizes approximately equal to 1.5, 5, 15, 50, 150 [6]. Therefore, we applied two different clustering algorithms (k-means and DBSCAN) to the tie strengths of each ego network, with a number of cluster equal to 5, and analysed the sizes of the obtained layers. Note that, for each ego network, this means applying k-means and DBSCAN clustering on a mono-dimensional variable (the tie strength). For k-means, we used the `Ckmeans.1d.dp` algorithm [16], which ensures that that the clustering procedure converges to a unique and optimal solution. For DBSCAN, we implemented an iterative procedure to find the value of ϵ that gives a number of cluster equal to the chosen k. In addition, we fixed $MinPts$ parameter to 2

Table 2: Properties of ego network circles found by applying k-means on the set of tie strengths of each ego with a fixed k=5.

Cluster	**C1**	**C2**	**C3**	**C4**	**C5**
			Size		
All authors	2.0	6.3	15.8	37.9	116.8
machine learning	1.9	5.7	14.2	33.5	102.6
physics	2.0	6.4	16.1	38.6	119.5
neuroscience	1.8	5.5	13.3	31.1	100.1
bioinformatics	2.2	7.5	19.6	49.2	150.9
ecology	2.0	6.2	15.4	36.3	105.5
molecular biology	2.1	6.2	15.0	34.3	97.0
			Minimum Tie Strength		
All authors	0.77	0.40	0.20	0.09	0.02
machine learning	0.88	0.46	0.24	0.10	0.03
physics	0.83	0.43	0.22	0.10	0.02
neuroscience	0.75	0.38	0.20	0.09	0.02
bioinformatics	0.65	0.33	0.16	0.07	0.02
ecology	0.67	0.34	0.18	0.08	0.02
molecular biology	0.64	0.33	0.17	0.08	0.02

to eliminate possible noisy points in the data. This is the same technique used, among others, in [6].

The sizes of the resulting layers found by k-means (nesting the subsequent groups of alters identified) are reported in Table 2. The last circle (C5) may sometimes appear larger than the average ego network size because ego networks with less than 5 different values of tie strength were not considered. The results in Table 2 indicate that co-authorship ego networks have a very similar pattern to social ego networks. Specifically, the sizes of the circles are similar to those found in generic social networks. Although in some cases (typically for C4 and C5) the sizes are below the reference values (50 and 150, respectively), it must be noted that the actual values of ego network circles in *specific* social environments have been found to vary quite significantly around those reference values. The scaling factors between consecutive layers are approximately close to 3 (namely, they average 2.8 for all the authors taken together, 2.7 for machine learning, 2.8 for physics, 2.8 for neuroscience, 2.9 for bioinformatics, 2.7 for ecology, and 2.6 for molecular biology). This is another reference fingerprint of ego network structures in many other social environments. The table also shows the minimum tie strength for an alter to be part of each circle (averaged across all egos). Remember that, based on our definition of tie strength (which is essentially an annual frequency of co-authorship, normalised by the number of co-authors in each publication), a value of 1 would mean that the ego and the alter publish together a joint paper with 2 authors only every year. Interestingly, the results show that the intensity of co-authorship for the inner-most layer are quite close to this value of frequency, indicating indeed a particularly strong intensity of collaboration inside C1.

Table 3: Properties of ego network circles found by applying DBSCAN on the set of tie strengths of each ego with a fixed k=5.

Cluster	**C1**	**C2**	**C3**	**C4**	**C5**
Size					
All authors	2.8	6.4	11.2	19.3	152.8
machine learning	2.7	6.2	10.8	18.4	138.3
physics	2.9	6.5	11.5	19.8	146.7
neuroscience	2.8	6.3	11.1	19.2	139.8
bioinformatics	3.0	7.0	12.5	22.1	172.5
ecology	2.8	6.4	11.4	19.8	135.3
molecular biology	2.9	6.7	11.7	20.2	121.4
Minimum Tie Strength					
All authors	0.46	0.29	0.20	0.14	0.02
machine learning	0.49	0.31	0.22	0.14	0.02
physics	0.44	0.28	0.20	0.13	0.02
neuroscience	0.43	0.27	0.19	0.12	0.01
bioinformatics	0.36	0.24	0.17	0.11	0.01
ecology	0.41	0.26	0.18	0.12	0.01
molecular biology	0.36	0.23	0.16	0.11	0.01

The results found by DBSCAN, reported in Table 3, are similar to those obtained by k-means and are compatible with the pattern found in social networks [6]. However, DBSCAN tends to place more alters in the outermost circle (C5), where the density is much higher than in the other circles. As a consequence, C3 and C4 are smaller than C3 and C4 of k-means. This difference between the two clustering algorithms has been found also in Twitter and Facebook ego networks [6]. Note that the size of C1 is affected by the choice of having the parameter MinPts set to 2. This eliminates single element clusters, but allows us to detect noisy data. In addition, the somewhat larger values of C5 could be a by-product of our iterative procedure, which is in some (few) cases unable to produce the desired number of clusters, especially for very small networks. As these networks are not considered in the analysis, the values of C5 might be slightly biased towards larger ego networks. The scaling ratios obtained by DBSCAN clustering are also in this case close to 3 (3.4 for all the authors taken together, 3.3 for machine learning, 3.3 for physics, 3.3 for neuroscience, 3.4 for bioinformatics, 3.2 for ecology, and 2.9 for molecular biology). The values of the minimum tie strengths are slightly lower than those found by k-means, in particular for C1, which is again a side effect of the minimum cluster size we have imposed (2).

These results are another strong indication that the same cognitive and time constraints may regulate also scientific collaboration networks. If this results could be further generalised to other types of collaboration networks, this might be a first step towards a convergence between the results in social network analysis and in organisational research.

5.5 Author H-index and Ego Network Properties

Having characterised the structure of co-authorship ego networks, we now study whether there is any sensible correlation between the scientific productivity of authors, measured through the h-index, and structural elements of the co-authorship network. Specifically, we analyse the correlation between the h-index and the ego network size.

H-index and ego network size are highly correlated ($r = 0.56$ without transformations and $r = 0.79$ when both variables are log-transformed). The correlation values for the different clusters are $r = 0.47$ without transformations and $r = 0.79$ with logarithmic transformation for machine learning, $r = 0.56$ and $r = 0.80$ for physics, $r = 0.57$ and $r = 0.77$ for neuroscience, $r = 0.64$ and $r = 0.82$ for bioinformatics, $r = 0.58$ and $r = 0.80$ for ecology, and $r = 0.68$ and $r = 0.77$ for molecular biology. These results tell us that the larger the ego network of an author, the higher her h-index, with an exponential function controlling the relation between the two variables (because of the higher correlation of log-transformed variables). This could indicate that a high number of collaborations brings more information sources, thus increasing innovativeness and creativity of the ego. In the management literature, this is typically seen as one factor determining a successful career [17]. Another possible explanation of this correlation is that having more opportunities of collaborating facilitates more extensive productivity, and a higher impact on the respective scientific communities. Or, that authors with high h-index attract more and more collaborators, thus increasing their network size. Discriminating between these possible factors would require longitudinal studies on the evolution of the ego network and h-index over time, which is part of future work. Finally, also in this case, note that the correlation values are quite consistent across the different disciplines. Again, this suggests that the reason behind this property may be a general human behaviour underpinning scientific collaboration and productivity, rather than a feature of specific disciplines.

6 Conclusion

We analysed a large set of co-authorship ego networks from Google Scholar. These networks show a hierarchical structure based on the intensity of collaboration with the ego. This structure is similar to those found in other social environments, and known to be the result of the cognitive and time constraints in humans. The size and the scaling ratio between the sizes of the identified layers are quite similar to those found in other social networks, both offline and online. Moreover, the size of the ego network is negatively correlated with the average intensity of collaboration of the ego. This property is another fingerprint of ego networks in general social networks. In addition, we found a direct relation between these structural properties and the performance of each author. In particular, the size of an author's ego network is positively correlated with her h-index. This property does not depend on the specific research field where authors are active, suggesting that this might be a general feature of human scientific collaboration. We think that the results presented in this paper

contribute to bridge the gap between organisational studies of team work and social network analysis, giving interesting and new insights into the nature of collaboration networks and social (ego) networks in general.

References

1. A. Abbasi, J. Altmann, and L. Hossain. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4):594–607, 2011.
2. A. Abbasi, K. S. K. Chung, and L. Hossain. Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, 48(4):671–679, 2012.
3. V. Arnaboldi, A. Guazzini, and A. Passarella. Egocentric online social networks: analysis of key features and prediction of tie strength in Facebook. *Computer Communications*, 36(10-11):1130–1144, 2013.
4. C. F. Doormann and R. Strauss. A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, 5(1):90–98, 2013.
5. R. I. M. Dunbar. Neocortex size and group size in primates: a test of the hypothesis. *Journal of Human Evolution*, 28(3):287–296, 1995.
6. R. I. M. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella. The structure of online social networks mirrors those in the offline world. *Social Networks*, 43:39–47, 2015.
7. B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: validation of Dunbar's Number. *PloS one*, 6(8):e22656, 2011.
8. R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702, 2005.
9. R. A. Guzzo and G. P. Shea. *Handbook of industrial and organizational psychology*. 1992.
10. J. S. Katz and D. Hicks. How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, 40(3):541–554, 1997.
11. E. Y. Li, C. H. Liao, and H. R. Yen. Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530, 2013.
12. M. E. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101 Suppl:5200–5205, 2004.
13. M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–9, 2001.
14. J. L. Ortega. Influence of co-authorship networks in the research impact: ego network analyses from Microsoft Academic Search. *Journal of Informetrics*, 8(3):728–737, 2014.
15. A. Sutcliffe, R. I. M. Dunbar, J. Binder, and H. Arrow. Relationships and the social brain: integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103(2):149–68, 2012.
16. H. Wang and M. Song. Clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29–33, 2011.
17. M. F. Wiersema and K. A. Bantel. Top management team demography and corporate strategic change. *The Academy of Management*, 35(1):91–121, 2010.
18. S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.
19. W. X. Zhou, D. Sornette, R. A. Hill, and R. I. M. Dunbar. Discrete hierarchical organization of social group sizes. *Biological Sciences*, 272(1561):439–44, 2005.

View publication stats