A Distributed Tool for Online Identification of Communities in Co-authorship Networks at a University

David Fernandes[1], Nuno David[1,2(✉)], and Maria João Cortinhal[1,3]

[1] University Institute of Lisbon – ISCTE-IUL, Lisbon, Portugal
nuno.david@iscte-iul.pt
[2] Dinamia-CET ISCTE-IUL, Lisbon, Portugal
[3] CMAF-CIO, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal

Abstract. Most universities have their public repositories of scientific publications available online. The data is made available raw or by department listing and does not provide the network of co-authorships that implicitly emerges from scientific collaborations among different departments. Sometimes, the network of co-authorships is computed within the institution, via standalone applications that have few or no functionalities to explore the structure of collaborations. The possibility of searching online and managing the network of scientific communities in the institution is a matter of management efficiency, both for the institution itself and other external collaborators. This paper explains a distributed architecture and a tool that uses data from an online institutional repository. The tool calculates and puts available online the co-authorship network that identifies research communities according to different algorithms. The tool reflects and identifies the emergent structure of communities, graphically analyses communities, exports, reports and follows up with the evolution of communities in time.

Keywords: Online institutional repositories · Community detection tools · Interdisciplinary collaboration · Co-authorship · Graph · Author · Publication · ABCD and MCL algorithms

1 Co-authorship Networks in ISCTE-IUL

Interdisciplinary collaboration fosters researchers to combine collective expertise and gain synergies. This may result in increased productivity, originality and higher-impact research. In academic institutions, most internal interdisciplinary, collaborative research emerges from ad hoc groups, based on personal relations. However, most institutions are structured into discipline-focused, encapsulated units, such as departments, schools and faculties. The segmentation of institutions into sub-structures of decentralized scientific areas, with higher or lower degrees of autonomy, is justified by organizational and financial efficiency. Without adequate instruments to foster collaboration, organizational segmentation is likely to limit the potential of collaboration among actors, either

R. Mugnaini (Ed.): DIONE 2020, LNICST 319, pp. 179–189, 2020.
https://doi.org/10.1007/978-3-030-50072-6_14

internal or external to the institution, making it more difficult to foster inter- and multi-disciplinary ideas and initiatives. Perceiving and analyzing collaborative research in academic institutions can be a way to develop academic policies for promoting further interdisciplinary research (Newman 2004).

Most universities have their public repositories of scientific publications available online. However, the network of co-authorships is usually computed within the institution, via standalone applications, with few or no online functionalities to explore and provide the structure of collaborations. In this work, we are interested in the development of tools that allow users and faculty boards to perceive the emerging social network structure of co-authorships and explore online the potential space of further collaboration in their institution. Automatically identifying communities enables detection of patterns of knowledge sharing within academic institutions, which otherwise would be imperceptible. In order to do so, we implemented a tailored framework that allows examining the evolution and the structure of the co-authorship network revealed by scientific publications of authors of ISCTE-IUL Lisbon University Institute (ISCTE-IUL)[1].

The framework identifies communities in the co-authorship network, according to different algorithms. It uses online data from the institutional repository database to construct the co-authorship graph, calculate the communities graphs, storing it into a documented oriented database and a making it available online to interested users.

Community detection algorithms is a graph clustering problem. Community detection is automatic, without human intervention, in the sense that the purpose of the algorithm is to mathematically find groups of nodes with higher likelihood of connecting each other than to nodes from other communities. However, given the availability of different algorithms in the literature, and the need to empirically analyze how and why users identify themselves with the generated communities, it is vital to provide online tools to facilitate the incorporation of new algorithms in the framework.

The purpose of this paper is to substantiate and describe the framework design and its ability to generate and provide online different networks based on different algorithms. Main requirements of design included: distribution, where users can access online the graph and analyse co-authorships from a web site; persistence, where the graph is analysed at a particular time, communities are detected and stored, making it possible to access it in a responsive way the evolution of the emergent networks through time; modularity, where new algorithms for community detection can be added and capture alternative views of communities.

The structure of the paper is as follows. In the next section we recall the concept of co-authorship networks. In Sect. 3 we describe the framework architecture and in Sect. 4 the implemented algorithms. In Sect. 5 we describe the results of the authorship community detection in ISCTE-IUL, according to two algorithms, and a give brief comparative analysis of results. Finally, in Sect. 6, we present the conclusions.

[1] ISCTE-IUL is a public university with approximately 10 000 students (https://www.iscte-iul.pt). It comprises four schools, 8 research centers and more than 500 professors and researchers.

2 Co-authorship Networks

Co-authorship networks are social networks that have been widely studied to determine the structure of scientific collaborations (Jackson 2008). Unlike citations, co-authorships involve a temporal and collegial relationship, and consequently places them more squarely in the realm of social network analysis (Liu et al. 2005).

A co-authorship network can be modelled by an undirected graph in which nodes and edges represent authors and co-authorship relationships, respectively. Besides the number of shared scientific publications, there are other factors that are important to shape collaboration patterns among authors. For instance, if one article has two authors and another article has ten authors, the authors in the first article should be considered more connected than those of the second article. To express this relationship magnitude we considered a weighed network as follows.

Let G = (V, E, W) be the co-authorship graph G where V is the set of nodes (authors), E is the set of edges (co-author relationships between authors), and W is the set of weights w_{ij} associated with each edge connecting a pair of authors (vi, vj). The weight of each edge (vi, vj) is then given by:

$$w_{ij} = \sum_{k=1}^{K_{ij}} \frac{1}{(n_k - 1)} \tag{1}$$

where K_{ij} and n_k represent the number of scientific publications co-authored by at least vi and vj, and the total number of co-authors of such publications, respectively. According to (1), each co-authored publication adds to the co-authorship relationship the factor $\frac{1}{(n_k - 1)}$ (Newman 2001). From now on, these weights will be named as Newman attractiveness.

3 Framework Architecture

Data for constructing the network originates from Ciência-IUL (2018). Ciência-IUL is the institutional repository of scientific publications produced by the authors of ISCTE-IUL. The information present in Ciência-IUL must be transformed into a network of co-authorships where community identification algorithms can be applied. The overall framework is intended to provide users with online, easy access through a web browser, as illustrated in Fig. 1.

The framework architecture contains four modules, according to Fig. 2: co-authorship graph generation, community identification, database and website. The construction of the co-authorship graph is a time-consuming process. Thus, once the graph is generated it is persisted in the database. The modular nature of the solution allows future replacement of either module without compromising the use of the remaining modules. For example, a new community identification algorithm can be added, which uses the persisted graph in the database without any changes to the modules responsible for generating the graph.

Data are collected from Ciência-IUL through a REST[2] API provided by Ciência-IUL and transformed into a graph with authors as vertices and co-authorships as edges

[2] Representational State Transfer (REST).

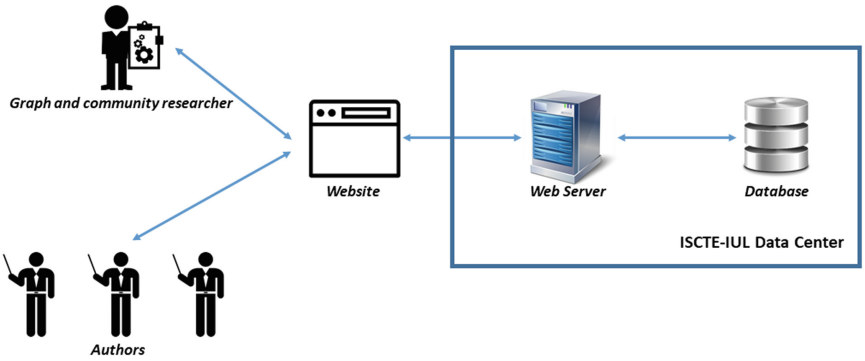

Fig. 1. Framework architecture.

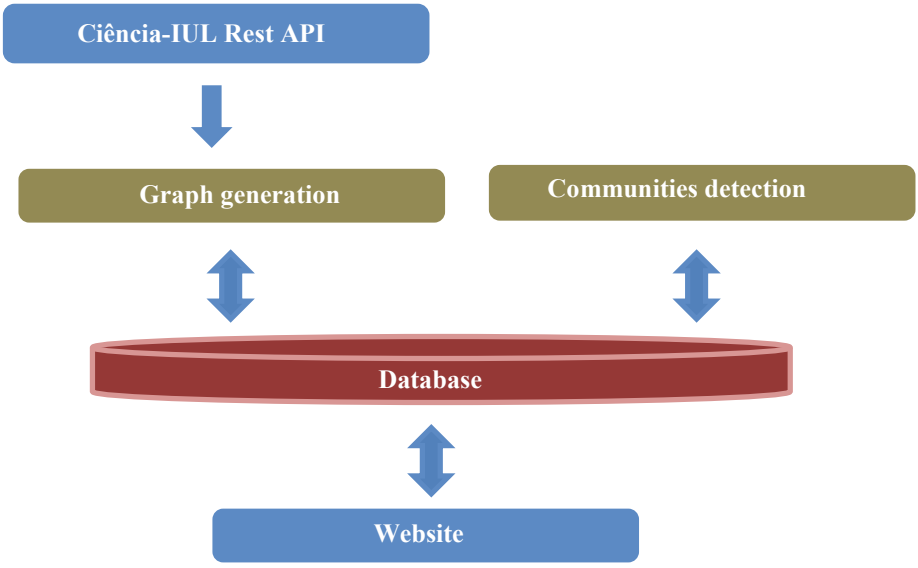

Fig. 2. Solution architecture.

between them. The REST API offers what are called endpoints to connect, gather information or perform some functionality in the Internet. Such web-services allow requesting systems to access and manipulate textual representations of web resources using a uniform, predefined set of stateless operations.

Like most publication repositories, Ciência-IUL API does not search for entities that have changed in a given period. This requires searching for all entities every time a co-authorship network is calculated. This limitation implies unbearable processing times, which would preclude providing online the networks in a responsive way. The solution was to construct a non-relational database (Fig. 3) with MongoDB (MongoDB 2017) that stores every graph. To consume the Ciência-IUL endpoints and create the

co-authorship network we used the Node.js software. (Node.js 2017). The framework is thus able to consult and compare communities calculated over time.

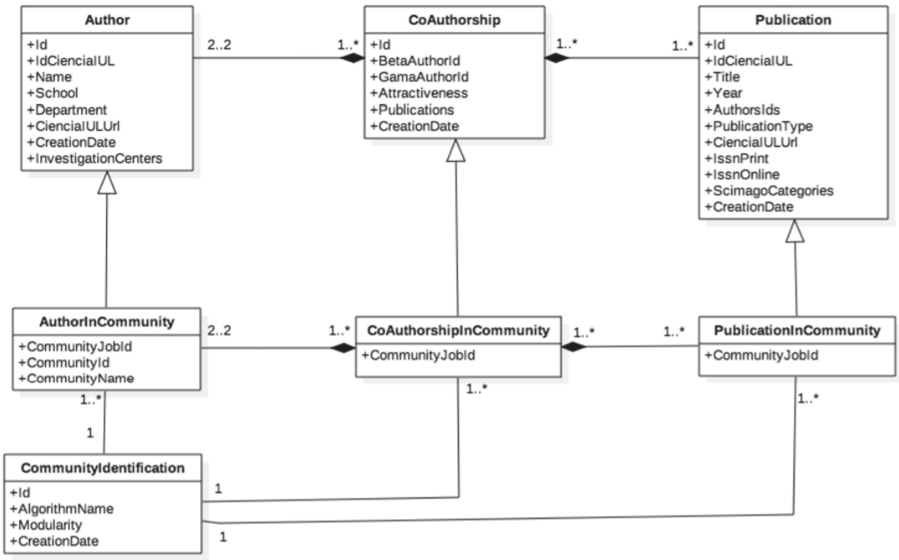

Fig. 3. UML data entities.

Once the graph is generated – a simple network with no identified communities – it is persisted in the database and may then be accessed through the graphical interface, in the website. Community identification algorithms can be applied and their results also stored, persisted in the database and accessed through the same website.

The Community Identification component is where community identification algorithms are executed. To the extent that it is a separate component, which will have its own timing and logic, one can easily add or remove algorithms to the solution. There are currently two algorithms in place, ABCD and MCL.

The graphical interface, through the website, allows to visualize online the co-authorship graph, as well as the communities identified in a commercial web browser. Its essential features are the visualization of the graph with the co-authorship network, the visualization of a graph with several identified communities (all vertices of the same community are the same color), the ability to query the communities or author details by accessing the vertex(es) that it represents and a set of descriptive metrics (such as author numbers, number of publications, number of co-authors, etc.). Exporting graphs in standard format to other general-purpose tools, such as GEXF - Graph Exchange XML Format (GEXF Working Group 2018), is also provided.

The database has the structure shown in Fig. 3. Each co-author has two authors and at least one publication, represented by the Publication entity. Whenever an instance of publication is created its Scimago scientific categories are added in the ScimagoCategories property. A publication may be in several different co-authorings. Each of these entities has a set of properties that characterize it in the context of the co-authoring

network. Note that this database is not intended to be a copy of Ciência-IUL. In this context, in order to know more details about an author or a publication the repository Ciência-IUL may be consulted.

Whenever communities are identified at a given time the community graph is stored in the derived entities AuthorInCommunity, CoAutorshipCommunity and Publication-InCommunity. The context of community identification, as well as the algorithm used, are stored in the CommunityIdentification entity.

4 Unfolding Co-authorship Communities

The detection of communities in graphs is a problem with vast literature and it is normally known as a graph clustering problem (Liua et al. 2014). On a broad sense, a community is a group of nodes with higher likelihood of connecting each other than to nodes from other communities. How to determine that a node belongs to a community and not to another one is central in this problem. We used Newman's concept of attractiveness, in order to give us the likelihood of two nodes belonging to the same community. We implemented two communities detection algorithms, the Markov Cluster Algorithm (MCL) and Attractiveness-based community detection (ABCD). It should be stressed that other algorithms can be implemented, providing the online user with different ways to calculate communities.

4.1 MCL

The Markov Cluster Algorithm (MCL) follows the principle that a cluster of nodes has many edges inside it and few connections to other clusters (Dongen 2000). This means that if two nodes, u and v, are in the same cluster, the probability (the Newman force) that a path between u and v has external nodes to the cluster should be low. Therefore, a random walk between u and v has very little probability of leaving their cluster, in other words, their community. Random walks in a graph can be described by means of Markov chains in which the sequence of variables in the chain is represented by a sequence of probability transition matrixes.

Once the initial matrix is determined, the algorithm proceeds iteratively until the stopping criteria is verified, that is, until there is no difference between two consecutive transition matrices. In each iteration, a new matrix is calculated using two operators: Expansion and Inflation. The Expansion changes the transition probabilities so that they reflect the introduction of intermediate edges in the random walk from any vertex j to any vertex i. The Inflation operator, in turn, increases the difference between the highest and lowest transaction probabilities and, in this way, it reinforces the attractiveness of the stronger edges and reduces the attractiveness of the weakest edges.

Stijn van Dongen created MCL and provided a tool to use it (Dongen 2017). We used his implementation in our application of MCL. We used it with the operation Expansion with $p = 2$ and with the operation Inflation with $r = 2$.

4.2 ABCD

Attractiveness-based community detection, ABCD (Ruifang Liua 2014) is an algorithm used to detect communities in weighted graphs. It relies on two main concepts: density of a group of nodes and attractiveness between a group of nodes. It is an algorithm for agglomeration of nodes, where they are put in the same group until some condition is met and the agglomeration stops. It begins with as many groups as nodes. As the number of nodes in a group grows, the denser it becomes and the more difficult it is to merge with another. Two groups of nodes are merged together whenever their attractiveness is higher than their own density. The attractiveness is based on the weight of the edges, which is measured by the Newman attractiveness. A group density is the sum of the weights of its nodes, and the weight of a node is the average weight of its edges. In this way, there is a direct relation between authors and co-authorships.

5 Results

In this section, we summarize and analyse the scientific collaborations among ISCTE-IUL authors. Our purpose in this section is not fundamentally theoretical but to describe, as a proof of concept, the kind of information a user can access and explore through our distributed framework. We considered all the ISCTE authors that have co-authored at least a scientific publication with another ISCTE author from 1975 to March 2017, which comprises a 42 years interval. It is worth to remark that only per-reviewed scientific contributions were considered in this study.

5.1 The ISCTE-IUL Co-authorship Network

Co-authorship networks document scientific collaborations, where nodes are authors and a link represent the fact that two authors have written at least one scientific publication together. Thus, they are undirected graphs.

The ISCTE-IUL co-authorship network – see Fig. 4 – has 613 nodes and 1718 edges. To highlight multidisciplinary scientific collaborations the shade of each node depends on the school that each author belongs. Among the 613 nodes there are 92, 185, 195 and 141 nodes representing authors belonging to, respectively, the School of Technology and Architecture, the School of Sociology and Public Policy, the School of Social Sciences and the Business School.

Despite having 1718 edges, the ISCTE-IUL co-authorship represents 3766 scientific publications, since some pairs of authors have co-authored more than one scientific publication. However, the width of each edge in Fig. 4 does not represent the number of publications. It represents the Newman attractiveness force (see Eq. 1).

In Fig. 4 it may be observed that there is one school (Business School) in which, unlike other schools, authors share many co-authorships with other schools' authors but they do not form a strong block: they are a bridge that connects the remaining three schools. It could be theorized that authors of this school exert a knowledge sharing that tends to be more transversal to ISCTE-IUL.

By analyzing Newman attractiveness force, it was also possible to conclude that there is a huge difference among the top 5 co-authorships, 109 versus 20. Moreover, the

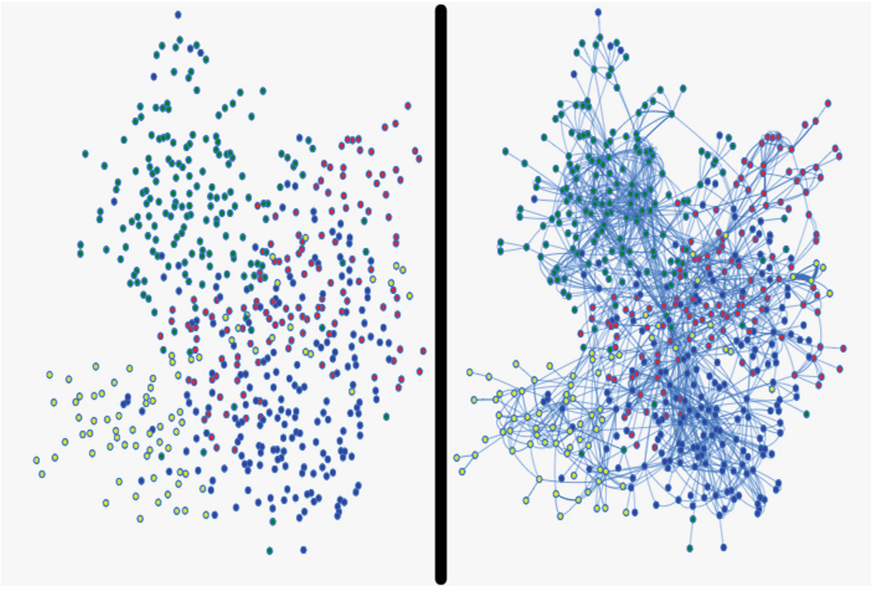

Fig. 4. ISCTE-IUL co-authorship network.

attractiveness of around 70% of co-authorships is equal to one, which reveals that the same pair of authors tend to share scientific publications only once. For this reason, the differences among the width of the edges in Fig. 4 are almost imperceptible.

5.2 Community Detection

As previously mentioned, two algorithms for community detection were implemented: MCL and ABCD. On what follows, communities with less than four authors were not considered and removed from this analysis. This was due to considerations of simplicity. However, this condition can be easily removed.

The MCL method identified 26 co-authorship communities in the network. Figure 5 displays the network with each different color representing each of the twenty six co-authorship communities.

From Fig. 5 we can observe that only one of the communities – the one represented by yellow nodes – stands out on the network. In fact, there is a single community with 39 authors and 121 co-authorships, whereas all the others have no more than 13 authors and 28 co-authorships. Moreover, about 65% (17 in 26) of the communities have less than 10 authors (Fig. 6).

The communities provided by the ABCD algorithm, by contrast, are much more homogeneous, as it can be seen in Fig. 7 and Fig. 8: none of the communities has more than 29 authors and 37 co-authorships, and around 60% of them have no more than 10 authors.

This means that in both algorithms, 60% of the authors of the original network were discarded because they could not be putted in a community with more than four authors.

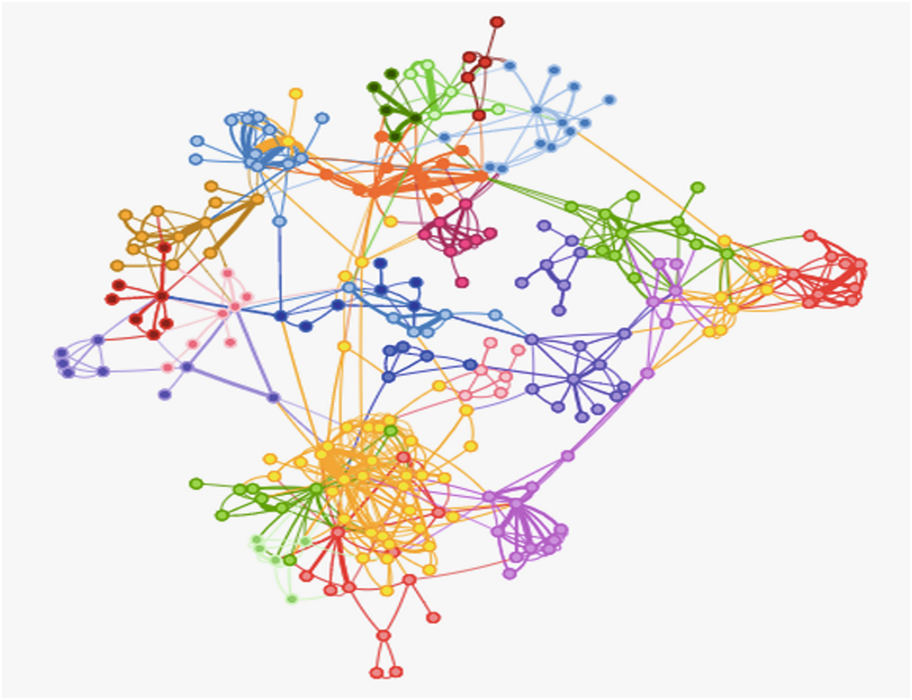

Fig. 5. Communities identified with MCL.

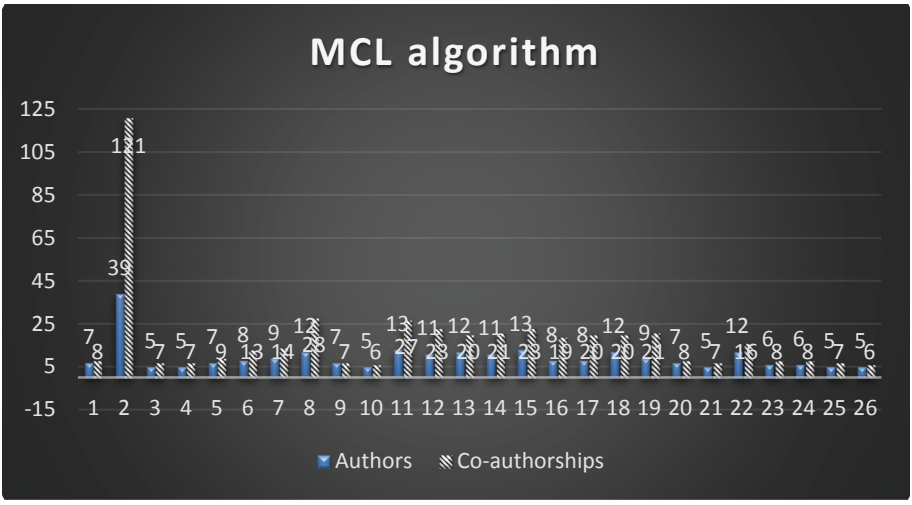

Fig. 6. Number of Authors and Co-authorships by community with MCL.

Hence, an explanation for this may be that the small number of authors in communities is due to the weak force of attractiveness of the co-authorships. Insofar as we have

Fig. 7. Communities identified with ABCD.

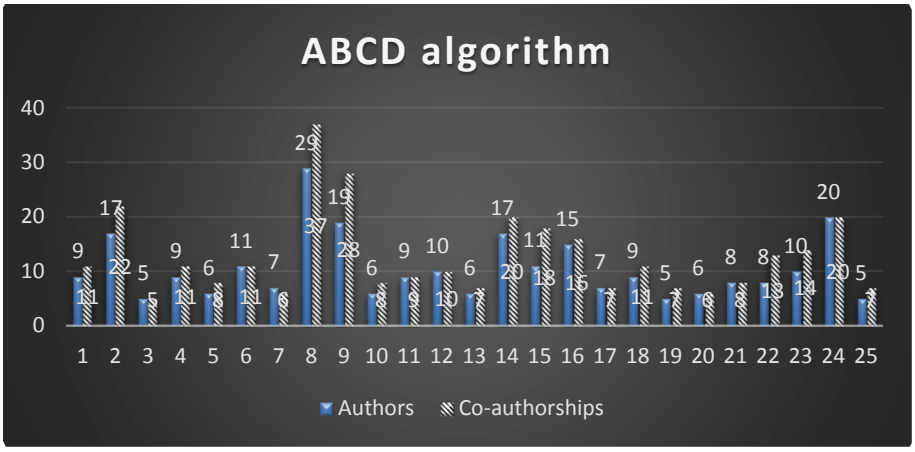

Fig. 8. Number of Authors and Co-authorships by community with ABCD.

determined that only a community with four authors is valid, it was expectable that many authors would be purged. From the 1755 co-authorships in the graph, 1225 (70%) have an attractive force equal or less than 1. Therefore, the algorithms generated small communities or even standalone authors, as solo communities.

6 Conclusions

The scientific information provided online by Ciência-IUL and the framework described were able to construct a co-authorship network and identify patterns of knowledge sharing in it. The distributed tool provides this information online and gathers important statistically information, like the total of authors, publications and co-authorships. Both the MCL and the ABCD algorithms found distinct communities using the force of Newman. However, this fact amounts to a challenge in the domain of community detection. Community detection is a formal, mathematical exercise. Once communities are automatically detected, we find a need to characterize the communities, that is, to provide the user with information in order to infer the identity or dynamics underlying the emergence of the community. In order to do so, some of the information provided may include the name, school, department and/or research center of each community member, the scientific areas of publications, the most frequent journals in the community etc. With this information the user will be able to compare the results of the different algorithms, as well as more easily confront the resulting patterns with his/her own expectations about the communities in which he/she thinks is integrated in. In this paper we describe a framework for easy and distributed access to this information which opens the doors to further modular improvements of the solution over the Web.

References

Ciência-IUL: Documentação da API Pública November 2018. https://ciencia.iscte-iul.pt/api/doc
Dongen, S.v.: Graph Clustering by Flow Simulation. University of Utrecht (2000)
Dongen, S.v.: MCL - a cluster algorithm for graphs January 2017. http://micans.org/mcl/
GEXF Working Group: GEXF File Format (2018). https://gephi.org/gexf/format/
Jackson, M.O.: Social and Economic Networks. Princeton University Press (2008)
Liu, X., Bollen, J., Nelson, M.L., Sompel, H.V.: Co-authorship networks in the digital library research. Community Inf. Process. Manage. **41**(6), 1462–1480 (2005)
MongoDB: MongoDB (2017, 2 6). https://www.mongodb.com
Newman, M.E.: Co-authorship networks and patterns of scientific collaboration. In: Proceedings of the National Academy of Sciences, pp. 5200–5205 (2004)
Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and central. Phys. Rev. E **64**(1), 01632 (2001)
Node.js: Node.js (2017, 2 6). https://nodejs.org
Ruifang Liua, S.F.: Weighted graph clustering for community detection of large social networks. In: 2nd International Conference on Information Technology and Quantitative Management, ITQM (2014)