Social and Technological Networks: Project (3) An Analysis on the Informatics Research Network Community Structure

Thomas Souliotis
The University of Edinburgh
Teammate: Giorgos Mousa.

***Abstract*—People in universities tend to collaborate more and more with people outside of their own domain. Furthermore, this is obvious nowadays as technology has advanced and many different people across different specialties are required to collaborate so as to solve modern problems.**

However, as it is expected people who belong to the same scientific domain tend to work together in many different projects and papers, as collaboration is an important attribute and requirement for an academic person.

Finally, in this project we try to examine these relations between people who belong to different institutes of the school of informatics in the University of Edinburgh. As part of that, various techniques and protocols are followed as part of the data analysis.

I. Introduction

In this report, we try and analyze the collaborations of the people across the institutes of the school of informatics in the University of Edinburgh.

First of all, some preliminary ideas are presented which actually guide the overall process of this project. We try and provide the fundamental ideas and models on which the analysis will be based on later. Moreover, a theoretical model of the expected relationships is presented, which is later analyzed further by examining the given datasets and performing the proper data processing.

Finally, a conclusion is provided which shows that people from the same institutes work together a lot more than with others from other institutes. However, that is not completely definititive and multiple metrics and algorithms show that people tend to work also with people from different institutes and external partners.

II. Preliminaries

In this part some basic ideas and protocols will be shown that are fundamental for the further understanding of this report. Also, the basic tools that were used in the process of this project will be stated and documented so as for anyone to be able to understand and possibly test the results presented.

First of all, we will consider as nodes the people and as edges the connections if a common publication of a paper has taken place. Having that in mind, the whole analysis follows, and gives as the opportunity to study the connections and the institutes as part of graph analysis algorithms.

In addition to that, as part of this project, the dataset that was used is the one suggested by the project description. This dataset consists of two csv files, each containing data about university members(their names and position) as well as the collaborations of different people who are either members or external partners of the university.

Moreover, the above datasets were analyzed with various algorithms which will be presented in more extent later in this report. What needs to be stated now is that all programs where written in Python 3, while the necessary packages so as to run the code will also be stated later.

Furthermore, an important issue for someone to understand this report is to have some knowledge -at least fundamental- of some graph algorithms. More specifically, some one should be used to metrics regarding graphs, while in the meantime a good knowledge of community detection/clustering algorithms is also required.

In the next sections the ideas which were briefly presented here, will be analyzed in more extent.

III. Theoretical Approach

A main objective of this project was to determine whether with proper analysis of the data, we could conclude if the Institutes of the school of informatics constitute different communities/clusters. Intuitively, this is a rather expected result as someone could expect that people from the same institute will collaborate a lot more with people from the same institutes. However, while this is partially true, the outcomes will show that it is not definitive. In order to study the above problem at first, we had to decide on the algorithms that would give as the wanted results. First and foremost the K-Means[1] algorithm was used. For this algorithm, the data will be divided to k different clusters with k being given as input by the user. The way it is done is, by choosing k centers, $c_1, ..., c_k$ and trying to minimize the sum of squared distances of nodes to their clusters($C = \{C_1, .., C_k\}$):

$$\Phi_{kmeans}(C) = \sum_{j=1}^{k} \sum_{a_i \in C_j} d^2(a_i, c_j)$$

Of course this is a NP-problem and many different algorithms with the most well-known being Lloyd's[2], have been developed. This algorithm is an heuristic one and the actual process

that converges and finds the solution will not be presented here as it is not the subject of this report.

Other than that, Affinity Propagation algorithm[3], is also used as a clustering analysis algorithm. This algorithm does not require to predefine the number of clusters in which the system will be divided, and is mostly used when we expect many clusters of not the same size.

Finally, Spectral[4] and Agglomerative[5] clustering algorithms are also used. The first one is used in case of few clusters while the second is expecting many. Both of them, have as input the expected number of clusters. Again, the exact process of how these algorithm reach to the result will not provided to this report as it is not the report's purpose to inform about these algorithms.

Other than that, edge expansion[6] in terms of considering each institute as a set is also explored. More precisely we measure:

$$\alpha = min_{S \subseteq V} \frac{e^{out}(S)}{min(|S|, |\tilde{S}|)}$$

with S, being each of one of the 6 institutes(*'Institute of Language, Cognition and Computation', 'Institute of Perception, Action and Behaviour', 'Institute for Adaptive and Neural Computation', 'Laboratory for Foundations of Computer Science', 'Institute for Computing Systems Architecture', 'Centre for Intelligent Systems and their Applications'*) while V will have two possible values. Either it will be all the people belonging to the institutes, or in a more general case, all the people including external partners that collaborate with institute members. Generally, the second case, as we will notice has a lot more 'noise' due to the fact that the external partners are a lot more than those of institutes. Of course the relative difference between in-edges and out-edges is also presented as a perspective for more clear results.

Finally, a degree distribution of the nodes is also researched to check whether the nodes correspond to a Power Law[7] network, or if actually the degrees do not have big fluctuations as expected.

IV. Practical Implementation and Measuremenrs

As it was previously noted the implementation of the algorithms and all the measurements where done in python3 programming language with the suggested datasets as input. Moreover, some figures will be given which will support what will be shown here as a result. Also, a guideline.pdf is also provided for anyone interested to run the program. Also, as part of convenience from now on, we will refer to institutes *'Institute of Language, Cognition and Computation', 'Institute of Perception, Action and Behaviour', 'Institute for Adaptive and Neural Computation', 'Laboratory for Foundations of Computer Science', 'Institute for Computing Systems Architecture', 'Centre for Intelligent Systems and their Applications'*, as institute 1 to 6 respectively.

Now we have to specify that two kind of results and analysis are given. One is for the full graph, and another is for the full graph without the 'outsiders'(external partners-people not being part of the institutes). The second case is also noted as cleared FG, meaning it is the full graph cleared of nodes and edges adding too much noise. This noise is obvious in Fig. 1, and Fig. 2, which show the full graph and full graph cleared respectively, and show the big difference between those two.

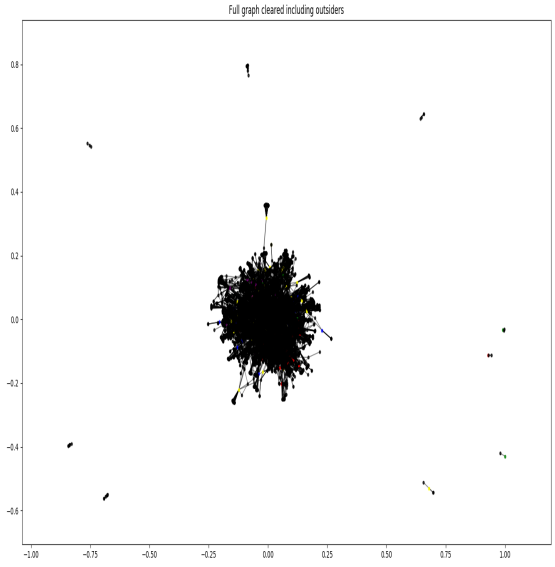

Figure 1. The full graph

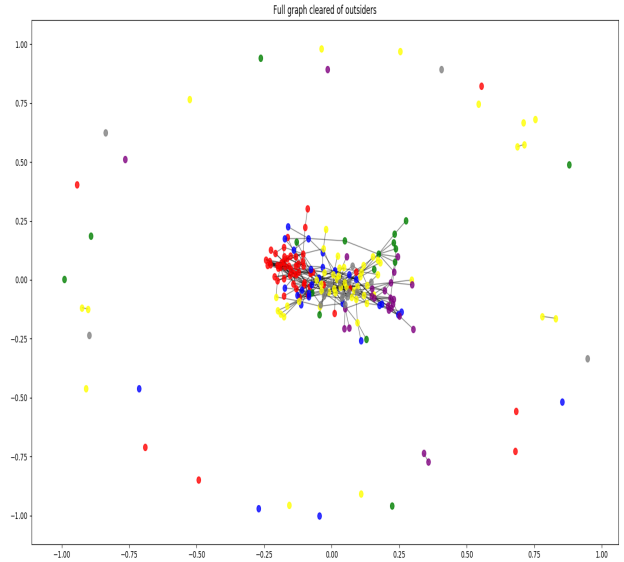

Figure 2. The full graph cleared from outsiders

As it is obvious from the above figures, the network has a main component and very few isolated nodes. Furthermore, the noise added in the full graph is obvious.

Regarding the clustering algorithms, what we have to say is that the results are relatively good. As shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6 these are the cluster results for k = 6 clusters(of course the affinity algorithm does not depend on k). What is more, despite the provided figures we can

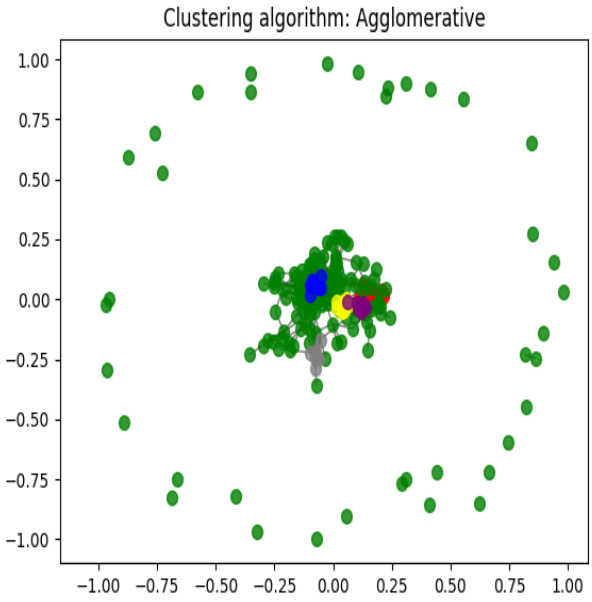

Figure 3. Agglomerative k =6

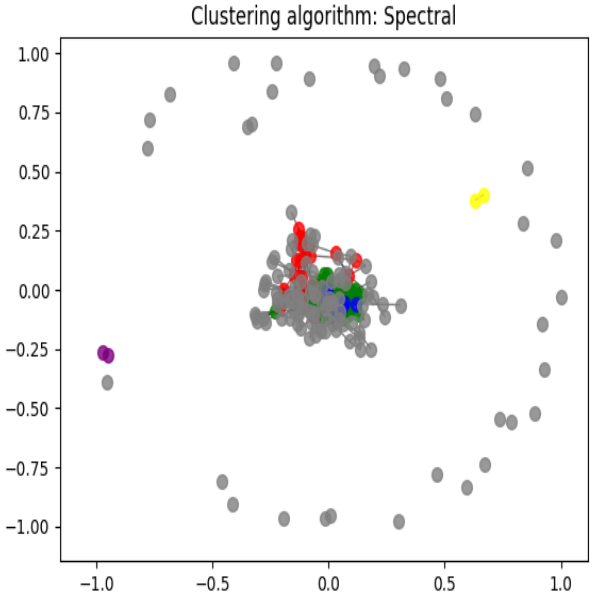

Figure 4. Spectral k = 6

observe by the execution of the code that for low values of k, the clustering is not that good, while when reaching high

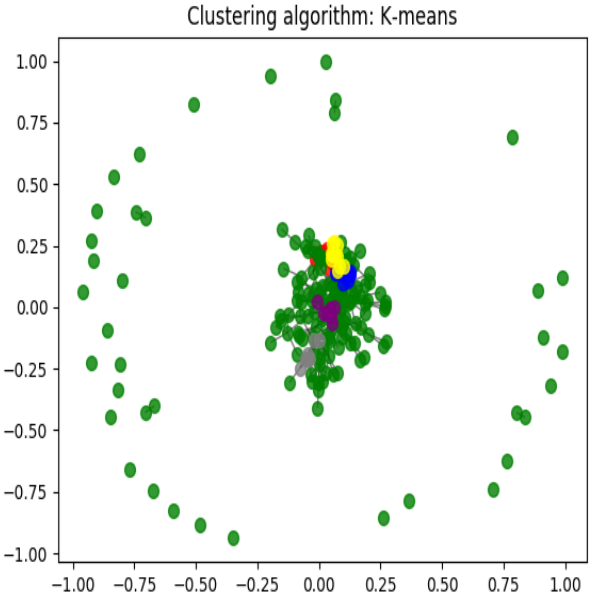

Figure 5. K - means, k = 6

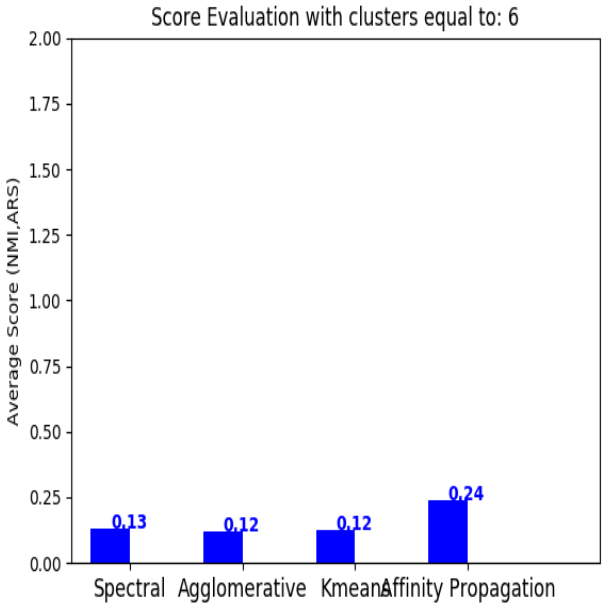

Figure 6. Comparison between the clustering algorithms.

values(greater than 6) the clustering results seem to converge. This result, shows us that despite the fact that the results are relatively good, however, all in all, we can say that the community detection is not as good as in other networks. The main problem here is the highly connected network, where many areas tend to have big overlaps. So, checking the results manually, we see that for some institutes with relatively

few outgoing edges the results are better, and the algorithms give better results. Finally, regarding the cluster analysis, as expected, the results are not good for isolated nodes, while spectral analysis gives better results than agglomerative and K-means, as the last two require large data sets and big number of clusters. Of course the affinity propagation analysis gives the best results when k=6. All these are shown in Fig. 7.

Despite the above clustering analysis what other metrics have given us are the following:

Institutes	Edges in	Edges out but in institutes	Edges completely out
1	180	26	2567
2	20	11	734
3	53	16	1524
4	123	37	2280
5	57	11	767
6	38	19	1033

What the above tabular shows, is edges beginning from each institute,and where they head to. From the above it is obvious that most edges are inside the institute they begin of. Of course if we consider outsiders too, then most of the edges are there. This has already been stated before and is the main reason we choose not to consider those nodes and edges in clustering analysis due to the too much noise. Moreover, the number of those nodes is more than 7000 while in institutes nodes are only 200. Thus, what we can conclude is that while there is an obvious collaboration between the institutes however still most collaborations are inside the same institute. However, a good amount of publications is done with people from different departments as it is obvious.

Now we will present the edge expansion tabular:

Institutes	Expansion among institutes	Expansion including outsiders
1	0.48148148148148145	49.68518518518518
2	0.6470588235294118	43.8235294117647
3	0.5714285714285714	55.0
4	0.6491228070175439	40.64912280701754
5	0.4583333333333333	55.36842105263158
6	1.0	26.473684210526315

What we extract from the above is the expected as before. When we do not have the outsiders then the edge expansion is not that big generally(of course for such a network the numbers are big), while when we incude them the value is really large.

Finally,from the degree distribution histogram we get the following diagram for the cleared graph:

Of course the histogram of the full graph is not presented here, but the values are shown in the python code.

In addition to that the average clustering coefficient for the whole Graph is 0.8081141306415582, while average clustering coefficient for the cleared Graph is 0.27507987299592396. Both values are considered to be really big, but especially for the whole graph the number is huge. What, the above values show, is that the network has significant clustering/community structure, due to the high cc (especially for the full graph including the outsiders).

V. Conclusions

In conclusion, what we have realized above is the true relations in the network of a the institutes in the school of informatics. The results may not be ideal, however a systematic way of trying to find the results and some important conclusions were given. What is more, some one should also

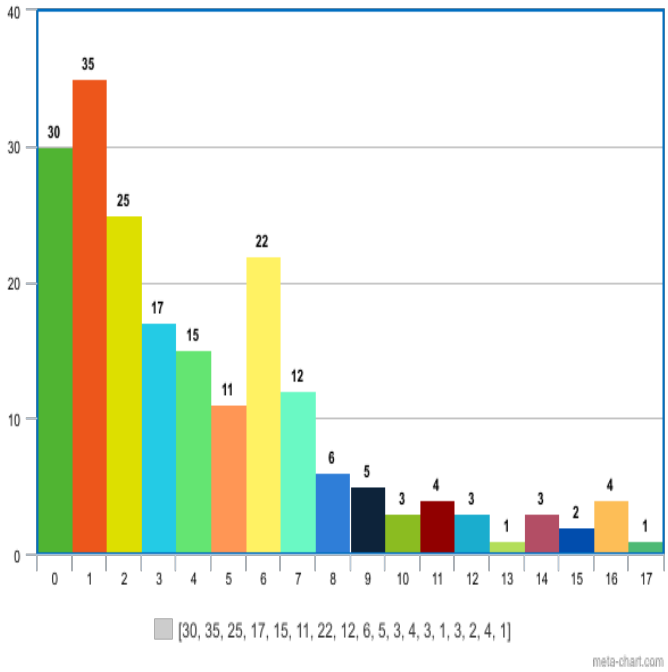

Figure 7. Histogram of full graph cleared.

check the code where a number of functions and results which were not presented here have been computed, due to the limited size of the requested report. Finally, in order to understand how the python code works, a guideline.pdf is also provided so as to be able anyone to run it.

VI. Acknowledgements

Special thanks to Giorgos Mousa, uun: s1687200, with whom we have thought of the problem together, and exchanged some very interesting ideas.

References

[1] MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07

[2] Lloyd, Stuart P. (1982), *Least squares quantization in PCM*, IEEE Transactions on Information Theory, 28 (2): 129137, doi:10.1109/TIT.1982.1056489

[3] Brendan J. Frey; Delbert Dueck (2007). *Clustering by passing messages between data points*. Science. 315 (5814): 972976. doi:10.1126/science.1136800. PMID 17218491

[4] Kannan, Ravi; Vempala, Santosh; Vetta, Adrian. *On Clusterings : Good. Bad and Spectral*. Journal of the ACM. 51: 497515. doi:10.1145/990308.990313.

[5] Zhang, et al. *Agglomerative clustering via maximum incremental path integral.* Pattern Recognition (2013).

[6] Goldreich, Oded (2011), *Basic Facts about Expander Graphs*. Studies in Complexity and Cryptography: 451464, doi:10.1007/978-3-642-22670-0-30

[7] Newman, M. E. J. (2005). *Power laws, Pareto distributions and Zipf's law*. Contemporary Physics. 46 (5): 323351. arXiv:cond-mat/0412004Freely accessible. Bibcode:2005ConPh..46..323N. doi:10.1080/00107510500052444