# A comparison of graph clustering algorithms

Jean Creusefond

# A comparison of graph clustering algorithms

Jean Creusefond
GREYC, Normandy University
jean.creusefond@unicaen.fr

## I. Introduction

The community detection problem is very natural : given a set of people and their relationships, can we understand the underlying structure of social groups? The applications are numerous in marketing, politics, social statistics, ...

Thanks to technological improvements and change of uses, many social networks of various sizes has been automatically extracted from recorded social relationships. The most obvious examples are the social network websites, but other networks are also studied as social networks : collaboration between scientists, who-talks-to-whom on the phone/on emails, etc. Automatic extraction made large networks available for study, and the community detection algorithms that we could evaluate with ease by watching the result on small instances before, can not be compared on real-world networks.

We therefore try to find common ground among the various clustering algorithms. Indeed, most of them share design similarities, as the underlying assumptions about the characteristics of communities or the general steps of the algorithm. Our experiments show to what extent these similarities imply similarities of results.

Our work is close to the one of Almeida *et al.* [1], that compared a good number of algorithms and a few quality functions. However, they did not directly compare clusterings, but only the result when a quality function is applied on them.

## II. Graphs and algorithms

We want to run available algorithms on all graphs, we select relatively small real-world graphs ($< 100K$ edges) to stay in reasonable computation times.

**Karate** [12], relationships in a karate club;
**Dolphins** [8], a community of dolphins;
**Football** [6], matchs between football teams;
**Netscience** [10], scientific co-authorships;
**Facebook** [7], friends in a facebook excerpt.

In this preliminary work, we compare 4 algorithms :

**Clauset** [3] : an optimisation for sparse graphs of the hierarchical modularity optimisation algorithm proposed by Newman [9], where communities are recursively merged if it corresponds to to merge improving the global modularity the most;

**Leading eigenvector** [10] : a spectral method based on the modularity matrix, the eigenvector of the largest positive eigenvalue is used to split the network in two. This method is applied recursively;

**Louvain** [2] : at first, vertex are communities. Each node then changes community affiliation to the one of its neighbour if it improves modularity. The process is repeated with communities as vertex until no improvement can be made;

**Conclude** [5] : a greedy algorithm using k-path centrality and the Louvain Method.

## III. Comparing clusterings

The Normalised Mutual Information [4] (NMI) is used to compare two clusterings (see Eq. 1). $I(A, B)$ is the mutual information of the two sets (an information-theoric measure of the the dependancy), that is normalised by the sum of the entropy ($H(.)$) of both clusterings.

## IV. Quality functions

Modularity [6] :

$$Q(C) = \sum_{c \in C} \left[ \frac{E(c)}{m} - \left( \frac{Vol(c)}{2m} \right)^2 \right] \quad (2)$$

Mean internal clustering coefficient :

$$Cl(C) = \sum_{v \in V} \frac{|(a,b) \in k^C(v), \{a,b\} \in E|}{\binom{|k^C(v)|}{2}} \quad (3)$$

With $k_v^C$ the set of neighbors of $v$ that are in the same community.

## V. Results

**Do algorithms output similar results?**

The average and standard deviation of the NMI are presented in the following table:

| | conclude | clauset | leading |
|---|---|---|---|
| louvain | 0.87±0.05 | 0.83±0.07 | 0.64±0.17 |
| conclude | 1.0±0.0 | 0.77±0.10 | 0.59±0.18 |
| clauset | | 1.0±0.0 | 0.64±0.18 |
| leading | | | 1.0±0.0 |

Louvain and Conclude are very similar, and that this trend does not particularly depend on the dataset (since the standard variation is low). It is surprising, since they make different assumptions about the characteristics of a community. They do, however, share the same agglomerative strategy.

Clauset is quite close to both of these methods, while having a different agglomerative strategy. But they all share some kind of greedy process explaining the closeness of these methods. On the other hand, leading eigenvector, due to its completely different approach of the cluster formation, produces results that are less similar to the others.

$$NMI(A,B) = \frac{2I(A,B)}{H(A)+H(B)} = \frac{2\sum_{c\in A}\sum_{c'\in B}|c\cap c'|log(\frac{|c\cap c'|*|V|}{|c||c'|})}{-\sum_{c\in A}|c|log(\frac{|c|}{|V|}) - \sum_{c\in B}|c|log(\frac{|c|}{|V|})} \quad (1)$$

**Figure 1:** *The normalised mutual information*

**Do quality functions behave accordingly?**

We rank the quality of each algorithm for each quality measure and each graph. The ranking process is done on a larger benchmark, with a total of ten algorithms. The global ranking is given here, to reflect the distance between algorithms. The first value is the modularity rank, while the other one is the mean internal clustering coefficient rank.

| | louvain | conclude | clauset | leading |
|---|---|---|---|---|
| karate | 1/1 | 5/2 | 6/8 | 4/9 |
| dolphins | 1/5 | 9/2 | 5/6 | 6/4 |
| football | 1/8 | 7/6 | 8/9 | 9/10 |
| netscience | 1/8 | 9/7 | 3/1 | 6/4 |
| facebook | 1/4 | 8/2 | 7/3 | 6/7 |

Surprisingly, Louvain and Conclude show almost symetrically opposed results. Indeed, Louvain is systematically the best method w.r.t modularity, while having a relatively bad internal clustering coefficient. On the other hand, Conclude is very good with the clustering coefficient and in the low end of modularity. It means that, while having clusterings that share a lot of common ground (one clustering can be easily recomposed from the other), they have different characteristics, measured by the quality functions. Clauset does an average performance, rarely diverging from around the middle of the rankings, and so does the leading eigenvector method.

## VI. Future works

We assume that, if a clustering algorithm give close results to another one, it is because they agree on underlying models. Quality functions can help to define what models are exactly used, by characterizing what a "good" clustering is and measuring which algorithm agrees with this characterisation.

But quality functions themselves often capture the same notion : inside/outside density ratio, difference with a null model, etc. Which means that they often are redundant. We intend to compare clusterings that rank high w.r.t. different quality functions, and to discover relationships between these measures.

Once a few central quality function are identified, we can check their compliance to a set of axioms such as in [11]. The non-compliance of one of these axioms may imply a counter-intuitive behaviour, and should warn the user of the possible biais.

## References


[1] H. Almeida, D. Guedes, W. Meira Jr, and M. J. Zaki. Is there a best quality metric for graph clusters? In *Machine Learning and Knowledge Discovery in Databases*, pages 44–59. 2011.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), 2004.

[4] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[5] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Mixing local and global information for community detection in large networks. *Journal of Computer and System Sciences*, 80(1):72–87, February 2014.

[6] M. Girvan and M. EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[7] J. Leskovec and J. J. Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[8] D. Lusseau, K. Schneider, P. Boisseau, O. J.and Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[9] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004.

[10] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), September 2006.

[11] T. Van Laarhoven and E. Marchiori. Axioms for graph clustering quality functions. *The Journal of Machine Learning Research*, 15(1):193–215, 2014.

[12] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.