EVALUATION OF THE PERFORMANCE OF A MODULARITY-BASED CONSENSUS COMMUNITY DETECTION ALGORITHM

Fabio Morea [1] and Domenico De Stefano [2]

[1] Area Science Park, Trieste, Italy (e-mail: fabio.morea@areasciencepark.it)

[2] Department of Social and Political Sciences, University of Trieste, Piazzale Europa 1, Trieste, Italy (e-mail: ddestefano@units.it)

ABSTRACT: This paper presents a novel consensus community detection (CCD) performed adopting a modularity-based community detection algorithm that exploits the concept of consensus over N independent trials to generate robust communities and to aggregate marginal nodes into a single community. The algorithm is tested on a class of artificial networks with built-in community structure that can be made to reflect the properties of real-world networks. Preliminary results show that CCD outperforms a single run of the original algorithm in terms of Normalised Mutual Information (NMI), number of communities and community size distribution, and provides an effective tool for community detection in real-world networks and a way to overcome the dependence on random seed of modularity-based algorithms.

KEYWORDS: network analysis, community detection, consensus.

1 Introduction

Community detection algorithms are a powerful tool for understanding the inner structure of natural and social complex systems that can be represented as networks [1]. This is generally an unsupervised learning task, as real-world networks often have no intrinsic labels; thus, the community structure found depends on the definition of "community" that is embedded in the community detection algorithm.

Modularity-based algorithms (a common choice in many fields of research) rely on the definition of community as a set of nodes that are more densely connected to each other than to the rest of the network. Modularity measures the degree to which the nodes within a given community are more densely connected to one another than to nodes in other partitions: the higher the modularity, the better the partition. Such algorithms have the advantage of being fast and providing easily interpretable results but have a relevant issue: using a "greedy" maximization approach, the composition of communities and the number of communities is different at each run.

The intrinsic variability of results may be acceptable if the subsequent analysis is focused on the global structure of the network. However, when the need to be interpreted in terms of individual nodes (e.g. research questions in the form of "do vertices V1 and V2 belong to the same community?") a more robust approach is required.

A common approach in such cases is to repeat the community detection algorithm several times, and to select as "best result" the iteration providing the maximum value of modularity.

We developed a different approach, based on the concept of Modularity-based Consensus Community Detection (MCCD), consists of the following steps:

1. **Independent trials**: The Louvain community detection algorithm [2] is repeated Ni times, with a randomly chosen fraction α of edges assigned a small, but non-zero weight W_0 and a randomly assigned resolution γ. This ensures that the resulting network does not lose connectivity, but edges associated with W_0 are more likely to be assigned to different communities at each iteration.
2. **Consensus**: The consensus algorithm counts how many times a pair of vertices V_i and V_j are assigned to the same community, and assigns a proportion of membership $P_{Vi} \in [0,1]$. Vertices that are strongly connected to one another are always assigned to the same community and have $P_{Vi} = 1$; lower values of P_{Vi} indicate that the vertex is not strongly connected to its neighbours, and it may be assigned to two or more communities with some degree of confidence.
3. **Pruning**: Nodes with $P_{Vi} < 0.5$ and trivially small communities (i.e. communities that have a number of nodes or a weight below a given threshold) are assigned to community "0".

2 Methodology

We evaluated the performance of the MCCD algorithm on artificial benchmark networks with built-in community structure defined by [3] and commonly named LFR after their Authors (Lancichinetti, Fortunato, Radicchi). These networks are characterised by a power-law distribution of the degree of the nodes and the size of the communities, a common feature of real-world networks. We generated a family of 9 benchmark networks with N = 1000 nodes, $\tau_1 = 2$, $\tau_2 = 3$, average degree = 10, and μ values from 0.1 to 0.9. All networks have 37 communities, with community size varying from 20 to 50. Lower values of mixing parameter μ indicate that the communities are clearly separated and are therefore easily identified by community detection algorithms; on the contrary, high values of μ are related to networks with fuzzy communities that are hard to identify.

To measure the performance of the clustering algorithm, we calculated the normalised mutual information (NMI) between the built-in partition of the graph and the one detected by the algorithm as a function of μ. Moreover, we compared the *number of communities* and the *community size distribution* of MCCD results with the original network.

3 Results and discussion

Comparative analysis (*Figure 1)* shows that the MCCD algorithm consistently outperforms the repetition of Louvain algorithm. For low values of μ (clearly defined communities) NMI is close to 1.0, with small differences between the methods; as μ increases (fuzzy communities), the consensus algorithm identifies communities that are more closely related to the original community structure. The parameter α does not significantly influence the values of NMI.

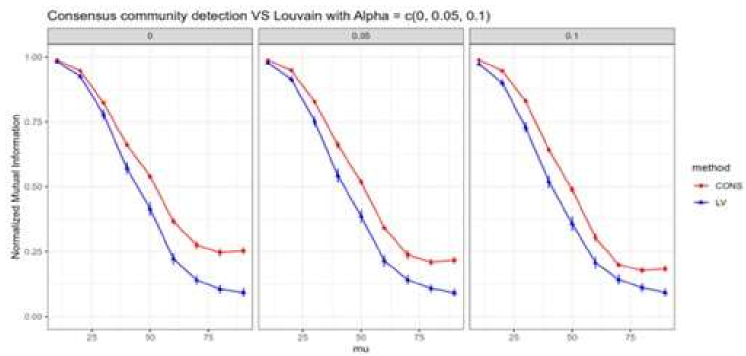

Figure 1: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network with 1000 nodes constructed with the following parameters: degree exponents $\tau_1 = 2$ and $\tau_2 = 3$, average degree $k_{avg} = 20$, maximum, degree $k_{max} = 10$, minimum community size $c_{min} = 20$, maximum community size $c_{max} = 50$.

As of community size, for high values of μ the Louvain algorithm partitions the network in larger communities, while MCCD, with appropriate values of parameter α produces a more robust results as shown in *Figure 2*.

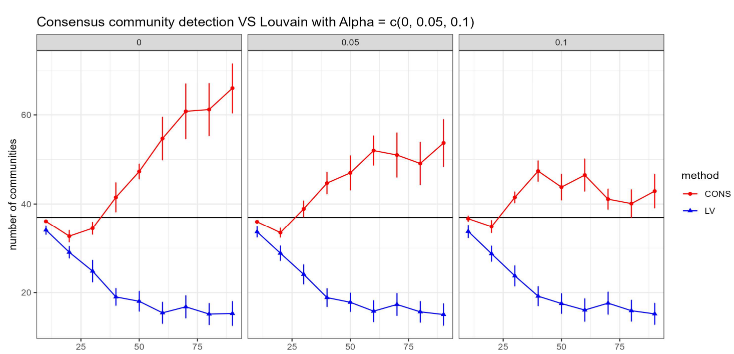

Figure 2: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network (same parameters as in previous figure)

Community size distribution varies significantly at each trial of the Louvain algorithm, and is consistently improved by applying MCCD, as shown in Figure 3. In the plot a single trial is represented by a vertical line, and a marker for each community. The black horizontal lines highlight the original community size distribution between 20 and 50. The Louvain algorithm (*blue lines*) identifies fewer, larger communities than the original ones, while the MCCD algorithm produces results that are much closer to the original community size distribution (*red lines*) and have higher NMI scores.

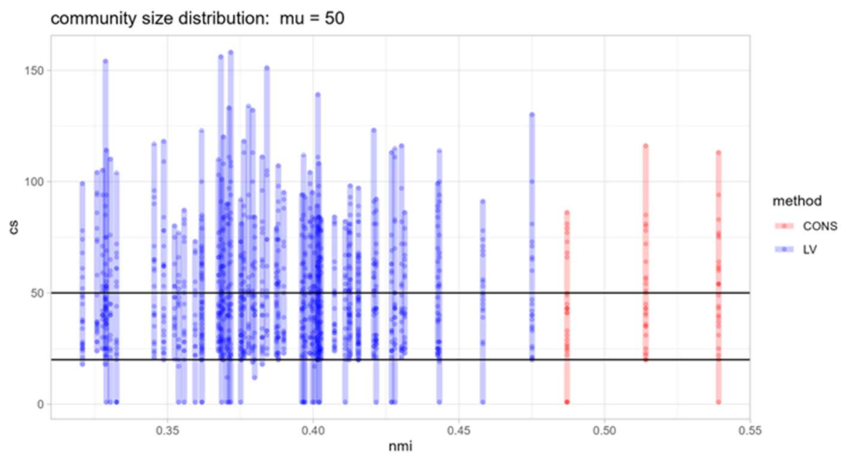

Figure 3: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network (same parameters as in previous figures).

Further research should focus on of the influence of parameters (independent trials, consensus and pruning), on LFR benchmark of different size and structure, as well as the application of MCCD to real world networks of different types.

References

[1] JIN D., YU Z., JIAO P., PAN S., HE D., WU J., YU P.S. ZHANG W. 2023. A Survey of Community Detection Approaches: From Statistical Modeling to Deep Learning, *IEEE Transactions on Knowledge & Data Engineering*, **35**.

[2] BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*.

[3] LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. 2008: Benchmark graphs for testing community detection algorithms. *Physical review E*, **78.4**.