Master in Data Science

Complex and Social Networks Laboratory

Finding and assessing community structure

Raffaele D'Agostino
raffaele.d.agostino@estudiantat.upc.edu

Àlex Font
alex.font@estudiantat.upc.edu

Project Report

Academic Year 2025/2026

December 17, 2025

Contents

1 Introduction

The analysis of community structure is a fundamental task in the study of complex networks, allowing us to identify groups of nodes that are densely connected internally while being sparsely connected to the rest of the network. The main objective of this laboratory is to evaluate and compare the performance of various community detection algorithms available in the R `igraph` library, such as Louvain, Walktrap, and Label Propagation.

In this work, we focus on assessing the quality of network partitions using two complementary approaches. First, we utilize the `clustAnalytics` package to evaluate the statistical significance of the detected communities. Second, we implement a custom evaluation metric based on the Jaccard Similarity Index. This metric allows us to quantify the agreement between a computed clustering and a reference partitions, which may be either a known ground truth or a high-quality consensus clustering.

The experimental analysis is conducted on four distinct datasets:

1. **Zachary's Karate Club**: A benchmark social network with a known ground truth partition.
2. **Synthetic Network**: A computer-generated graph with a scale-free degree distribution and built-in community structure.
3. **ENRON Email Network**: A real-world communication network processed as a weighted graph.
4. **Immuno Network**: A biological network representing the interaction structure of an immunoglobulin protein (from the `igraphdata` package). It consists of 1316 nodes representing amino acids, where edges indicate spatial proximity.

The report describes the implementation of the Jaccard-based comparison tools, presents the results obtained for each network, and discusses the strengths and limitations of the different detection methods encountered during the analysis.

2 Methods

In this section, we describe the criteria designed to evaluate clustering quality by comparing a computed partition against a reference clustering. The reference clustering acts as a "ground truth" or a high-quality proxy (e.g., the partition with the best internal scoring function values).

2.1 Selection of the Reference Clustering

To evaluate the performance of community detection algorithms, it is crucial to establish a reliable reference clustering against which comparisons can be made. Our selection strategy relies on the statistical analysis of internal quality metrics (computed as the mean over $N = 10$ independent runs to account for stochasticity) and varies depending on the availability of a known ground truth.

We evaluated each algorithm using a set of topological scoring functions provided by the `clustAnalytics` package. We selected one scoring function for each different quality criterion: based on internal or external connectivity, a combination of both or based on a network model. For each metric, we identified the optimal direction:

- **Higher is Better:** Clustering Coefficient (Based on internal connectivity).
- **Lower is Better:** Expansion (Based on external connectivity).
- **Lower is Better:** Conductance (Combine internal and external connectivity).
- **Higher is Better:** Modularity (Based on a network model).

Based on these scores, the reference clustering was selected according to two criteria:

Case 1: Ground Truth Available (e.g., Karate, Synthetic). When the true community structure is known, we included the ground truth partition in the evaluation table. We compared the metrics of each algorithm directly against those of the ground truth.

Case 2: Ground Truth Unknown (e.g., Enron, Immuno). In the absence of a ground truth, we adopted a "consensus" approach based on the ranking of quality metrics. For each algorithm, we counted how many metrics achieved the best score (highest modularity, lowest expansion, etc.) compared to the others. The algorithm with the highest number of top-ranking metrics was selected as the *Proxy Ground Truth*. This data-driven reference was then used to compute Jaccard similarity scores for the remaining algorithms.

2.2 Chosen metrics

In this section we briefly describe the metrics used to evaluate the quality ("significance") of the community partitions.Following the notation in the slides, let $C \subseteq V$ be a community of size n_C, with m_C internal edges and f_C frontier edges (edges with one endpoint in C and the other in $\bar{C}$).

Modularity. Modularity compares the observed number of intra-community edges with the expected number under the configuration (null) model that preserves the degree sequence. In the unweighted case it is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \mathbf{1}\{C_i = C_j\},$$

where A_{ij} is the adjacency matrix, k_i is the degree of node i, m is the number of edges, and $\mathbf{1}\{C_i = C_j\}$ is 1 if nodes i and j belong to the same community and 0 otherwise. Larger Q indicates partitions with more internal connectivity than expected at random, hence *higher is better.*

Clustering coefficient. As an internal-connectivity score, we use the global clustering coefficient (transitivity), i.e. the fraction of connected triplets that are closed into triangles. Equivalently,

$$C = \frac{\#\text{closed triplets}}{\#\text{connected triplets}} = \frac{3\,\#\text{triangles}}{\#\text{connected triplets}}.$$

Higher transitivity suggests stronger triadic closure inside groups, therefore *higher is better.*

Expansion. Expansion measures external connectivity by counting how many edges leave a community per node:

$$\text{expansion}(C) = \frac{f_C}{n_C}.$$

Small values indicate that few edges per node cross the community boundary, so *lower is better.*

Conductance. Conductance combines internal and external connectivity by normalizing the cut size by the community volume. With the same notation as above, the conductance of a community C is

$$\phi(C) = \frac{f_C}{2m_C + f_C}.$$

Low conductance indicates a well-separated community (a "bottleneck" cut), hence *lower is better.*

2.3 Jaccard index as a similarity measure

The Jaccard index is a standard statistic used to measure the similarity between sample sets. In the context of community detection, we use it to quantify the overlap between a specific cluster found by an algorithm and a reference cluster.

Given two sets of nodes A and B, where A belongs to a reference partition (e.g., Ground Truth) and B belongs to a computed partition, the Jaccard index $J(A, B)$ is defined as the ratio of the size of the intersection to the size of the union of the two sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

By definition, $0 \leq J(A, B) \leq 1$. A value of 1 implies $A = B$ (perfect match), while 0 implies the sets are disjoint.

To evaluate an entire partition, we first compute a pairwise Jaccard table containing the index for every pair of clusters (C_i^{ref}, C_j^{alg}). Then, for each reference cluster C_i^{ref}, we identify the "best match" in the computed partition:

$$J_{max}(C_i^{ref}) = \max_j J(C_i^{ref}, C_j^{alg}) \tag{2}$$

This step aligns each community in the ground truth with its most similar counterpart in the algorithm's output.

2.4 Weighted average of Jaccard indices

To derive a single global similarity score for the entire network clustering, we compute the weighted mean of the best-match Jaccard indices found in the previous step. The weight for each reference cluster C_i^{ref} is proportional to its size (number of nodes), ensuring that larger, more significant communities contribute more to the final score than small or trivial ones.

The Global Jaccard Similarity S_{global} is calculated as:

$$S_{global} = \sum_i \left(J_{max}(C_i^{ref}) \times \frac{|C_i^{ref}|}{|V|} \right) \tag{3}$$

where $|V|$ is the total number of nodes in the network.

Using a weighted mean is more reasonable than a simple arithmetic mean because it penalizes algorithms that fail to detect large, structurally important communities while potentially performing well on small, outlier groups. A simple mean would treat a tiny cluster of 2 nodes equally to a major community of 100 nodes, potentially skewing the similarity score.

2.5 Minimum of Jaccard indices

Similarity measures based on averages provide an overall assessment of how well two partitions match, but they may hide significant errors in individual communities. In particular, an algorithm can achieve a high global score while failing to correctly identify one or more important clusters, which could be problematic in practical applications.

To explicitly capture this worst-case scenario, we consider a *thresholded minimum* of the best-match Jaccard indices. We global similarity score is the minimum best-match Jaccard over all communities whose size exceeds a given threshold. This prevents very small or noisy clusters from dominating.

Let $J_{max}(C_i^{ref})$ be the be the best-match Jaccard index for the reference cluster C_i^{ref} and let $w_i = |C_i^{ref}|/|V|$ denote its relative size. The thresholded minimum similarity score is defined as:

$$S_{\min} = \min_{i:\, w_i \geq \tau} J_{max}(C_i^{ref}), \tag{4}$$

where τ is a minimum size threshold. We used a value of $\tau = 0.01$, meaning we consider only communities that are at least 1% of the network.

This metric works as a worst-case perspective by focusing on the least well-matched relevant community. Compared to average-based scores, the thresholded minimum provides a more conservative evaluation, showing if an algorithm makes large errors on communities that are not negligible.

2.6 Weighted harmonic mean of Jaccard indices

While the thresholded minimum focuses on the worst-matched community, it can be overly conservative, since the global score depends on a single cluster. To obtain a more balanced assessment, we also consider the weighted harmonic mean of the best-match Jaccard indices.

The harmonic mean penalizes low values more than the arithmetic mean, but still incorporates information from all communities. In this sense, it provides a compromise between average-case performance and sensitivity to poorly recovered clusters.

Compared to the weighted arithmetic mean, the harmonic mean gives more weight to communities with low Jaccard values, especially when they are large. Unlike the minimum, it still accounts for all communities, providing a more balanced global similarity score.

3 Results

In this section we present the results obtained on each dataset. Firs, we present the significance test for each clustering algorithm and chosen metric, than we give the results of the cluster quality computation.

3.1 Zachary's Karate Club Network

	Louvain	**labelprop**	**walktrap**	**betweenness**	**ground truth**
clustering coef	0.595 ± 0.029	**0.569 ± 0.034**	0.614	0.451	0.534
expansion	3.082 ± 0.464	**2.153 ± 0.446**	3.471	5.765	1.294
conductance	0.229 ± 0.034	**0.16 ± 0.034**	0.255	0.419	0.095
modularity	0.411 ± 0.014	0.395 ± 0.015	0.411	**0.362**	0.371

Table 1: Significance scores (mean ± standard deviation) of the community partitions obtained with different community detection algorithms on the Karate club network. Standard deviations are computed over 10 runs; entries reported without a standard deviation correspond to deterministic outputs in our setup. Bold values are the closest to the ground truth (in absolute difference) for each metric.

	Louvain	**labelprop**	**walktrap**	**betweenness**
GT1 best cluster (J)	L1 (0.69)	LP1 (0.69)	W2 (0.69)	B1 (0.56)
GT2 best cluster (J)	L3 (1.00)	LP3 (1.00)	W1 (0.72)	B4 (0.56)
Weighted mean (J)	0.706	**0.853**	0.706	0.559
Min best-match (J)	0.625	**0.688**	**0.688**	0.556
Harmonic mean (J)	0.698	**0.824**	0.705	0.559

Table 2: Karate Club: best cluster matches between the ground truth partition (rows) and each algorithm's partition (columns), based on the maximum Jaccard index per ground-truth community. The weighted mean aggregates the per-ground-truth best-match Jaccards using weights proportional to community size, as in the lab instructions; the minimum and harmonic mean provide alternative global summaries.

3.2 Barabasi-Albert Network

	Louvain	**labelprop**	**walktrap**	**betweenness**	**ground truth**
clustering coef	0.194 ± 0.002	0.17 ± 0.025	0.204	**0.196**	0.198
expansion	2.346 ± 0.132	1.409 ± 0.847	2.35	**2.26**	2.29
conductance	0.297 ± 0.02	0.175 ± 0.106	0.295	**0.283**	0.287
modularity	0.460 ± 0.009	0.293 ± 0.176	0.455	**0.463**	0.462

Table 3: Significance scores (mean ± standard deviation) of the community partitions obtained with different community detection algorithms on a Barabási–Albert graph. Standard deviations are computed over 10 runs; entries reported without a standard deviation correspond to deterministic outputs in our setup. Bold values are the most close to the gorund truth.

	Louvain	**labelprop**	**walktrap**	**betweenness**
GT1 best cluster (J)	L1 (0.95)	LP1 (0.38)	W4 (0.88)	B1 (0.84)
GT2 best cluster (J)	L4 (0.83)	LP2 (0.44)	W2 (0.84)	B4 (0.80)
GT3 best cluster (J)	L3 (0.88)	LP1 (0.55)	W3 (0.87)	B3 (0.89)
GT4 best cluster (J)	L2 (0.92)	LP2 (0.50)	W1 (0.77)	B2 (0.84)
Weighted mean (J)	**0.896**	0.475	0.836	0.843
Min best-match (J)	**0.833**	0.375	0.769	0.800
Harmonic mean (J)	**0.894**	0.466	0.834	0.842

Table 4: Barabási–Albert blocks: best cluster matches between the ground truth partition (rows) and each algorithm's partition (columns), based on the maximum Jaccard index per ground-truth community. The weighted mean aggregates the best-match Jaccards with weights proportional to the ground-truth community sizes, while the minimum and harmonic mean provide alternative global summaries.

3.3 ENRON Network

	Louvain	**labelprop**	**walktrap**	**betweenness**
clustering coef	1.031 ± 0.005	1.05 ± 0.016	**1.085**	0.962
expansion	**223.603 ± 0.937**	293.252 ± 53.064	254.88	680.75
conductance	**0.211 ± 0.001**	0.286 ± 0.059	0.275	0.652
modularity	0.275 ± 0.001	0.228 ± 0.016	**0.278**	0.193

Table 5: Significance scores (mean ± standard deviation) of the community partitions obtained with different community detection algorithms on the Karate Club network. Standard deviations are computed over 10 runs; entries reported without a standard deviation correspond to deterministic outputs in our setup. Bold values are the best per metric (high is best for clustering coefficient and modularity; low is best for expansion and conductance).

	Louvain	**labelprop**	**betweenness**
W1 best cluster (J)	L6 (0.73)	LP5 (0.69)	B8 (0.55)
W2 best cluster (J)	L3 (0.56)	LP8 (0.70)	B3 (0.17)
W3 best cluster (J)	L1 (1.00)	LP1 (0.93)	B1 (0.67)
W4 best cluster (J)	L2 (0.13)	LP9 (0.56)	B15 (0.44)
W5 best cluster (J)	L3 (0.23)	LP2 (0.33)	B2 (0.20)
W6 best cluster (J)	L8 (0.86)	LP7 (0.81)	B2 (0.33)
W7 best cluster (J)	L5 (1.00)	LP4 (1.00)	B6 (0.65)
W8 best cluster (J)	L2 (0.54)	LP2 (0.45)	B5 (0.18)
W9 best cluster (J)	L2 (0.04)	LP11 (0.67)	B11 (0.38)
W10 best cluster (J)	L7 (1.00)	LP6 (1.00)	B5 (0.17)
W11 best cluster (J)	L2 (0.01)	LP2 (0.01)	B17 (1.00)
W12 best cluster (J)	L9 (1.00)	LP13 (1.00)	B24 (1.00)
W13 best cluster (J)	L2 (0.01)	LP2 (0.01)	B25 (1.00)
W14 best cluster (J)	L10 (1.00)	LP14 (1.00)	B27 (1.00)
W15 best cluster (J)	L8 (0.04)	LP7 (0.05)	B29 (1.00)
Weighted mean (J)	0.646	**0.665**	0.378
Min best-match (J)	0.0435	**0.333**	0.167
Harmonic mean (J)	0.302	**0.384**	0.278

Table 6: ENRON: best cluster matches using Walktrap as reference partition (pseudo–ground truth). For each Walktrap community (rows), the table reports the detected community with maximum Jaccard similarity (value in parentheses). The last rows report global summaries (weighted mean with weights proportional to reference community sizes, minimum, and harmonic mean); bold values indicate the highest score among algorithms for each summary measure.

3.4 Immuno network

	Louvain	labelprop	walktrap	betweenness
clustering coef	0.506 ± 0.001	**0.598 ± 0.015**	0.511	0.503
expansion	0.44 ± 0.02	2.057 ± 0.255	**0.419**	0.438
conductance	0.046 ± 0.002	0.216 ± 0.025	**0.044**	0.046
modularity	**0.872 ± 0.002**	0.758 ± 0.02	0.861	**0.872**

Table 7: Significance scores (mean ± standard deviation) of the community partitions obtained with different community detection algorithms on the Immuno network. Standard deviations are computed over 10 runs; entries with no standard deviation correspond to deterministic outputs in our setup. Bold values are the best according to the corresponding quality criterion (high is best or low is best).

	Louvain	**labelprop**	**betweenness**
W1 best cluster (J)	L2 (0.50)	LP7 (0.33)	B3 (0.56)
W2 best cluster (J)	L13 (1.00)	LP48 (0.39)	B13 (0.96)
W3 best cluster (J)	L6 (1.00)	LP17 (0.26)	B6 (0.95)
W4 best cluster (J)	L8 (0.77)	LP51 (0.25)	B8 (0.74)
W5 best cluster (J)	L7 (0.81)	LP23 (0.38)	B7 (0.78)
W6 best cluster (J)	L10 (0.93)	LP30 (0.39)	B10 (0.88)
W7 best cluster (J)	L5 (1.00)	LP16 (0.62)	B5 (0.76)
W8 best cluster (J)	L11 (0.94)	LP43 (0.31)	B11 (0.91)
W9 best cluster (J)	L4 (0.99)	LP10 (0.82)	B2 (0.98)
W10 best cluster (J)	L9 (0.89)	LP28 (0.39)	B9 (0.88)
W11 best cluster (J)	L1 (0.99)	LP2 (0.33)	B1 (0.98)
W12 best cluster (J)	L12 (0.89)	LP36 (0.40)	B12 (0.91)
Weighted mean (J)	**0.858**	0.393	0.844
Min best-match (J)	0.505	0.252	**0.557**
Harmonic mean (J)	**0.813**	0.359	**0.813**

Table 8: Immuno: best cluster matches using Walktrap as reference partition (pseudo–ground truth). For each Walktrap community (rows), the table reports the detected community with maximum Jaccard similarity (value in parentheses). The last rows report global summaries (weighted mean with weights proportional to reference community sizes, minimum, and harmonic mean); values are rounded to 3 significant digits and bold indicates the highest score per summary (ties after rounding are broken using the unrounded values).

4 Discussion

4.1 Zachary's Karate Club Network

For the Karate Club network, the ground truth is known, so the significance score metrics are not used to select the reference clustering. It is worth noting that, in some cases, the best metric based on the quality criterion, that is, the highest or lowest value, is not always the one that most closely matches the ground truth. This highlights the difference between selecting metrics for reference clustering and evaluating against known ground truth.

In table 1 we see that Label Propagation gets significance metrics closest to the ground truth ones. This suggests this algorithm as the one generating the more accurate clustering of the network.

Now we want to compare the clustering algorithms using the similarity measure that we defined using Jaccard Similarity Index. In table 2 we can see that Label propagation gets the highest similarity scores, matching the previous result using other similarity measures. With the min-best match we also get an insight on how bad is the error on the most poorly recovered cluster. In this case we see that Label Propagation also gets the best result.

4.2 Barabasi-Albert Network

For the Barabási–Albert (BA) synthetic network the ground truth partition is known too, hence significance scores are used to describe how "community-like" the partitions returned by each algorithm are. In particular, according to the quality criteria discussed in class, clustering coefficient and modularity should be maximized, while expansion and conductance should be minimized.

From Table 3 we observe that Louvain, Walktrap, and Edge Betweenness yield values that are very close to the ground-truth scores across all four metrics. Conversely, Label Propagation deviates markedly (e.g., much lower modularity and clustering coefficient), indicating that the partition it produces is structurally different from the planted one, even if some external-connectivity scores can look "good" in isolation. This illustrates that optimizing a single quality criterion can lead to partitions that are plausible from a structural standpoint but still misaligned with the generating (ground-truth) communities.

To directly assess the quality of clusters, we compare each detected partition with the ground truth using the Jaccard-based best-match similarity.

Table 4 shows that Louvain achieves the most consistent agreement with the ground truth: each planted community admits a very high best-match Jaccard score, and the corresponding global aggregations (e.g., weighted and harmonic means) are also the highest among the tested methods. Walktrap and Edge Betweenness provide comparably strong results, but with a slightly worse "weakest" recovered cluster (as highlighted by the minimum best-match), suggesting that at least one community is split or contaminated more than in the Louvain solution.

4.3 ENRON Network

For the Enron network, the ground truth is unknown, so we must rely on significance score metrics to select a reference clustering. In table 5 we observe that Louvain and Walktrap produce similar results, both outperforming the other algorithms. We choose Walktrap as the reference clustering, as it is deterministic, unlike Louvain, which may yield different results in each run. This ensures that our experiments are more reproducible. Then the closest clustering to the reference is given by Louvain or Label Propagation, while Edge Betweenness gets significantly lower scores.

Looking at the Jaccard similarity measure in table 6, we see that Label Propagation achieves the highest scores, closely followed by Louvain, which shows similar results in the weighted mean of Jaccard indices. However, we observe that Louvain's minimum best-match Jaccard is significantly lower than that of Label Propagation. This suggests that, while both algorithms perform similarly on average, Louvain performs worse on certain clusters, whereas Label Propagation is more consistent across all clusters.

Looking back to the similarity scores from table 5 we notice that Louvain was getting the higher scores in expansion, which is based on external connectivity, and conductance, that also takes external connectivity into account, while getting lower score in clustering coefficient, based on internal connectivity, than Label Propagation. This may explain why Louvain performs well on average in terms of Jaccard similarity, while showing weaker performance in Local Jaccard for certain clusters.

We also see that Edge Betweenness performs significantly worse in all similarity measures, including the Jaccard indices, compared to the other algorithms.

4.4 Immuno network

For the Immuno network the ground truth partition is unknown, so we need a reference clustering to make a fair comparison across algorithms. We select the reference using the significance (quality) metrics, where clustering coefficient and modularity are typically interpreted as *high is best*, while expansion and conductance are interpreted as *low is best*. In Table 7, Walktrap achieves the highest number of best values according to these rules, so we choose Walktrap as the reference clustering; moreover, Walktrap is deterministic, which makes the experiment reproducible.

After fixing Walktrap as reference, Table 8 compares all other partitions to it using the Jaccard best-match similarity. Louvain and Edge Betweenness are the closest to the Walktrap reference overall, while Label Propagation is clearly less similar. More in detail, Louvain achieves the best average agreement (highest weighted mean Jaccard), whereas Edge Betweenness achieves a better worst-case behaviour (higher minimum best-match). The harmonic mean is the same for the two algorithms.

A practical implication is that, under the chosen reference, Louvain is the best option when the goal is to match the reference well *on average*, while Edge Betweenness is preferable when it is important to avoid strongly mis-reconstructing at least one community. Finally, the lower similarity of Label Propagation suggests that it tends to produce a qualitatively different partition on this network (e.g., by merging or splitting some groups differently than Walktrap), so its communities should be interpreted with more caution.

5 Conclusion

Across all experiments, we noticed that evaluating community detection depends strongly on whether a reliable ground truth partition is available. When ground truth exists (Karate Club and the synthetic Barabási–Albert graph), similarity measures such as the Jaccard best-match provide the most direct notion of correctness, while significance/quality metrics mainly help describe the structural properties of the obtained partitions.

When ground truth is unknown (Enron and Immuno), selecting a reference clustering becomes necessary, and significance metrics provide a principled way to choose it by ranking algorithms according to a mix of criteria where some are "high is best" (e.g., clustering coefficient, modularity) and others are "low is best" (e.g., expansion, conductance). In this setting, determinism is also important for reproducibility, hence preferring deterministic references (e.g., Walktrap in our setup) makes repeated comparisons more stable and interpretable.

From a methodological standpoint, global similarity scores (e.g., weighted mean Jaccard) and local/worst-case indicators (e.g., minimum best-match Jaccard) capture different aspects of performance. In particular, two algorithms can look similar on average while still having very different behaviour on the hardest-to-recover community, which is important if downstream analyses depend on all communities being reasonably well recovered.

Finally, results across networks confirm that there is no universally best algorithm: Label Propagation performed best on Karate, Louvain was the most faithful on the BA ground truth, and for real networks without ground truth (Enron and Immuno) different methods trade off average agreement versus robustness on the weakest cluster.