Finding and assessing community structure

Complex and Social Networks (Fall 2024-2025)

Adrià Casanova, Dmitriy Chukhray

27th of November, 2024

1 Introduction

In this assignment, we are tasked with comparing different community finding algorithms through different metrics. In our case, we have selected four networks (karate, a synthetic one, ENRON and dolphins) and tested the Louvain, Label Propagation, Walktrap and Edge Betweenness clustering algorithms. Then, we have compared them with modularity and coverage scores. Moreover, we have implemented the local and global Jaccard indexes between clusters and used them to find the best clustering for each of our networks. The dolphins network can be found in the following GitHub repository: https://github.com/balajisriraj/Network-Analysis-Key-Players-Community—Detection—Dolphins.

2 The Jaccard index

The Jaccard index between two clusters is defined as

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

The closer to 1 is $J(A,B)$, the more similar that A and B are.

To compare two clusterings numerically, we need a global Jaccard indexes that combines the local Jaccard indexes between pairs of clusters from different clusterings.

The weighted mean of the local Jaccard idexes (weights given by fraction of number of nodes in each cluster) is, then, a reasonable clusterings similarity, since it gives more importance to large clusters than to small ones. Compared to the mean value, it accounts for the size of each cluster instead of considering all clusters equally relevant. This reflects the idea that larger clusters have a greater impact on the overall structure of the network.

Another way of combining the vector of Jaccard indices to quantify clusterings similarity is taking the harmonic mean, which mitigates the impact of large outliers and aggravates the impact of small ones. Hence, very similar clusters are not so important when using the harmonic mean, while very distinct clusters penalize a lot in the global Jaccard index. Basically, harmonic mean of is useful when we want to ensure that all clusters, regardless of size, are equally considered, and when it's important to identify and emphasize any poor matches between clusters.

3 Results

The function *evaluate_significance* returns a table with many different clusters' metrics, but we had to decide which ones to use as a reference during the ground truth proposition for networks with no ground truth. Modularity is a measure of the strength of the division of a network into communities. The closer the value of modularity is to 1, the better, as it indicates well-defined communities (clusters) where most edges are within clusters and few are between clusters. Coverage measures the fraction of all edges in the graph that are contained within the communities (clusters). The closer the value of coverage is to 1, the better, as it means that the communities (clusters) capture most of the network's structure.

According to the information for this assignment, ground truths for karate and synthetically created networks can be directly extracted from them. For ENRON and dolphins networks we had to apply 4 different clustering algorithms and then choose a ground truth based on selected metrics scores. Tables 1 and 2 below show the scores for all 4 clustering algorithms for ENRON and dolphins networks.

Metric	**Louvain**	**LabelPropagation**	**Walktrap**	**EdgeBetweenness**
Modularity	0.2607	0.1988	0.1529	0.0338
Coverage	0.9778	0.9631	0.9191	0.9945

Table 1: Modularity and Coverage for ENRON Network

Metric	**Louvain**	**LabelPropagation**	**Walktrap**	**EdgeBetweenness**
Modularity	0.5233	0.4981	0.4888	0.5194
Coverage	0.7610	0.8113	0.8239	0.7987

Table 2: Modularity and Coverage for Dolphins Network

For the ENRON network, we have to go with Louvain-produced clusters as the ground truth because its best modularity score (0.2607) is produced by the Louvain algorithm. The coverage metric is the best using the Edge Betweenness algorithm (0.9945), but it is not much higher than under the Louvain algorithm (0.9777). If we compare the Edge Betweenness score in modularity it is approximately 8.5 times lower than that of Louvain. Therefore, we can assume clusters produced by the Louvain algorithm are the closest to the ground truth for the ENRON network.

For the Dolphins network, we have to go with the Label Propagation produced clusters as the ground truth. In both modularity and coverage scores, there is no clear "leader" among clustering algorithms. Therefore, to choose the clustering algorithm, we decided to sum up both metrics and divide them by 2. This way we identify Label Propagation clusters as the best with a score of 0.66.

The important thing to mention is that we applied all 4 clustering algorithms to 2 networks only once, even though there are algorithms that involve randomness. Whenever we tried to apply clustering algorithms on networks more than once they always returned the same results (with no specific *set.seed* function, with alternating *set.seed* functions for every iteration, and with specific *set.seed* function). Thus, we expect everyone to obtain the same metrics scores with *set.seed* function being set to 42.

After setting ground truths for ENRON and dolphins networks we can evaluate its metrics in addition to local and global Jaccard indices. The ground truth of karate network has 2 clusters, whereas the Louvain algorithm produced 4 clusters, the Label Propagation algorithm produced 3 clusters, the Walktrap algorithm produced 4 clusters, and the Edge Betweenness algorithm produced 6 clusters. The highest global Jaccard index for karate network is given by the clusters of the Label Propagation algorithm with the value of 0.856. If we take a look at the extended table of *evaluate_significance* function, and not just modularity and coverage, the values of the Label Propagation clusters are the closest to the ground truth. An interesting observation is that the modularity score of the Label Propagation clusters is not the highest nor the closest to the ground truth of karate network. However, the difference is practically negligible, and if we take a look at all the other metric scores the Label Propagation clusters are outperforming all the other clusters produced by other algorithms. Local Jaccard index table of the Label Propagation clusters also strongly indicates that it's the closest clustering to the ground truth. It has the closest number of clusters to the ground truth (3 created against 2 from ground truth), and clusters' similarities (Label Propagation clusters 1 and 2 compose 81.25 percent and 18.75 percent of the ground truth cluster 1 and Label Propagation cluster 3 is the same as cluster 2 of the ground truth).

In the case of synthetic network, the best algorithm based on global Jaccard indices is Edge Betweenness with a score of 0.933. For this network, all algorithms produced 4 clusters except for the

Label Propagation which produced only 1 cluster (hence it has a low global Jaccard index of 0.285). Looking at the global Jaccard indices of the Louvain and the Walktrap algorithm, they are considerably close to that of the Edge Betweenness, with values of 0.853 and 0.877 respectively. Just like in the case of karate network, the best global Jaccard index means the best and/or the closest values of metrics to the ground truth. This time, the best clustering algorithm (Edge Betweenness) produces the exact same value of coverage (0.7175) and practically the same value of modularity (0.4594 against 0.4598 of the ground truth).

In the case of the last 2 networks, it is where it gets problematic. We don't know what the real ground truth is, therefore we have to assume that our approach of selecting clusterings that are the closest to the ground truth is somewhat right. We were doing that by using the highest values of modularity and coverage as the reference of the ground truths. Since global Jaccrad indices were agreeing with relevant metrics (modularity and coverage) on what the best clusterings are for networks with ground truth known apriori, we can assume the same for networks with unknown ground truths. In both ENRON and dolphins networks, the highest global Jaccard indices were given to the clustering algorithms that were used as ground truths. However, in the case of dolphins network the differences between the Label Propagation, the Walktrap, and the Edge Betweenees were small.

4 Conclusions

Having a ground truth is an essential element of performing good and meaningful network analysis. It is possible to do it without ground truth but many assumptions would have to be made and there would be less certainty in the results. The global Jaccard index has proven itself to be an effective metric in identifying similarities between clusters.