ADVANCED LEARNING FOR TEXT AND GRAPH DATA
[M2, MVA]

Maha ELBAYAD
maha.elbayad@student.ecp.fr

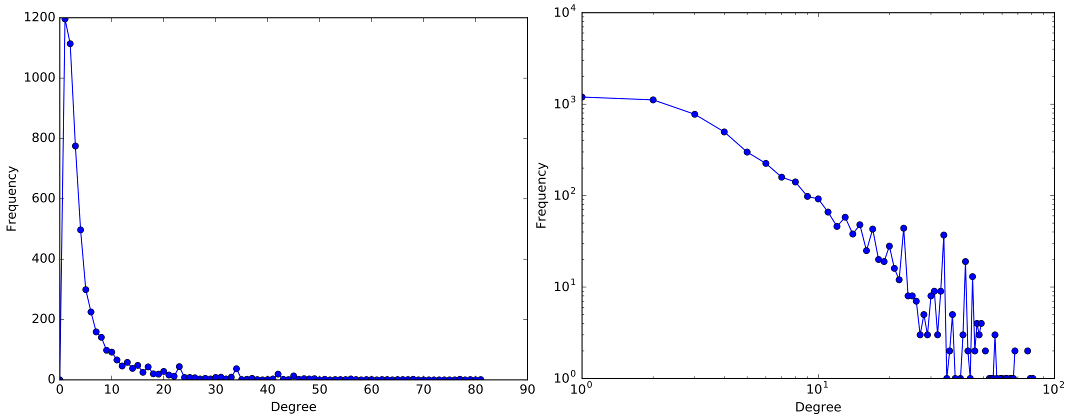
# Lab 2 and 3 : Graph Mining

## 1 Part 1: Analyzing a Real-World Graph

### 1:3. Network characteristics:

We note that, in fact, real world graphs have large connected compoenents, up to 92% of the total edges in the graph.
Moreover, the graph has a heavy-tailed degree distribution: few nodes with high degrees and the majority of nodes with small degrees (fig 1a). This is why a logarithmic scale is more apprpriate as it adapts to the wide range ([1,81]) and emphasises the more frequent small values.

Number of nodes: 5242
Number of edges: 14496
Number of connected components: 355
Fraction of nodes in GCC: 0.793
Fraction of edges in GCC: 0.926
Min degree: 1
Max degree: 81
Median degree: 3
Mean degree: 5.530



(a) Linear scale

(b) Logarithmic scale

Figure 1: Degrees distribution

### 4. Analysis of clustering structures in the graph.

We compute the total number of triangles in the graph using the `triangles` function of `NetworkX`, since every triangle is counted 3 times (for each vertex), we divide the output by 3. For the given graph, the total number of triangles equals **47779**

**Spectral triangles counting:**

We use the eigen-decomposition of the adjacency matrix $\mathbf{A}$ to approximate the triangles counting. The absolute values of the eigenvalues (fig 2a) are skewed, typically following a power law and their signs tend to alternate (symmetry about 0).
We've established that:

$$\Delta(G) = \frac{1}{6}\sum_{i=1}^{|V|}\lambda_i^3,\ sp(\mathbf{A}) = \{\lambda_1, ..., \lambda_{|V|} :\ |\lambda_1| \geq |\lambda_2|...|\lambda_{|V|}|\}$$

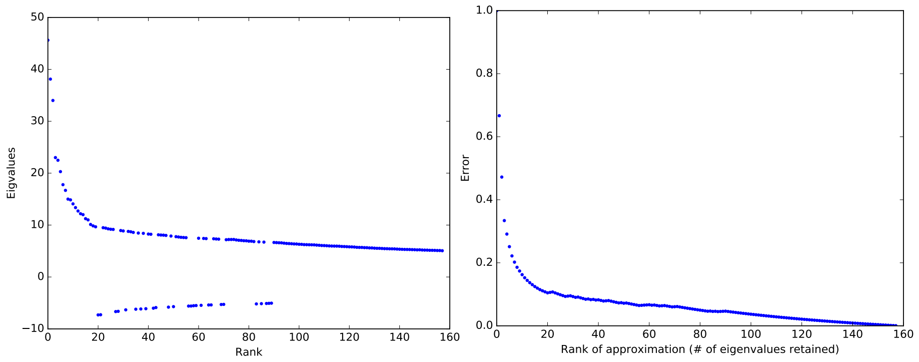
$$\text{Top-k approximation: } \tilde{\Delta}_k(G) = \frac{1}{6}\sum_{i=1}^{k}\lambda_i^3$$

In the previous summation the alternating signs cancel each other out and when considering only the top $k$ eigenvalues we would be simply ignoring the low energy eigenvalues. The error obviously decreases with the number of retained eigenvalues and we get an excellent approximaion with the top 20 strongest eigenvalues.
We can also approximate the number of triangles $\Delta^i(G)$, $\forall i \in V$, i.e., in a per node basis. In fact the i-th diagonal element of $\mathbf{A}^3 = U\Lambda^3U^T$ counts the number of triangles that node $i$ participates to (each counted twice):

$$\Delta^i(G) = \frac{1}{2}\mathbf{A}^3_{i,i} = \frac{1}{2}\sum_{j=1}^{|V|}\lambda_j^3 u_{i,j}^2$$

The top-k approximation count is:

$$\tilde{\Delta}^i_k(G) = \frac{1}{2}\sum_{j=1}^{k}\lambda_j^3 u_{i,j}^2$$



(a) Adjacency matrix eigenvalues (real part)

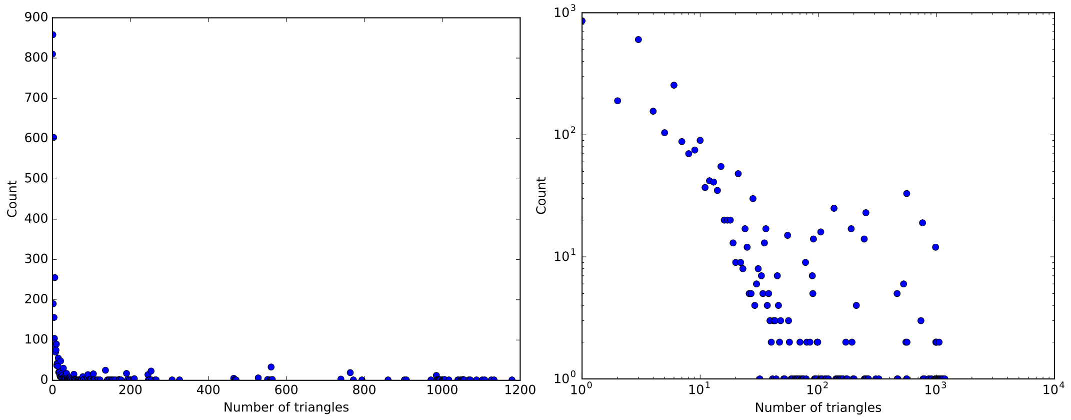
(b) Approximation error

Figure 2

**Triangle participation:**

```
t=nx.triangles(GCC)
t_values = sorted(set(t.values()))
t_hist = [t.values().count(x) for x in t_values]
```

`t_values` represents the unique values of the triangles count per node and `t_hist` the occurences count of each value. The distrbution is heavy-tailed / following a power law.



(a) Linear scale

(b) Logarithmic scale

Figure 3: the triangle participation distribution

**Clustering Coefficient:**

- Average clustering coefficient (GCC) 0.5568
- Average clustering coefficient (G) 0.52963

## 5. Node centrality:

We compute two of these centrality measures: degree and eigenvector and examine their correlation via Pearson coefficient (1 is total positive correlation, 0 is no correlation, and 1 is total negative correlation):

$$(\text{Pearson correlation coefficient, p-value}) = (0.5956, 0)$$

The p-value for testing non-correlation indicates that strong linear correlations do exist between the two centrality measures as can be seen in the two separate regions of fig 4.

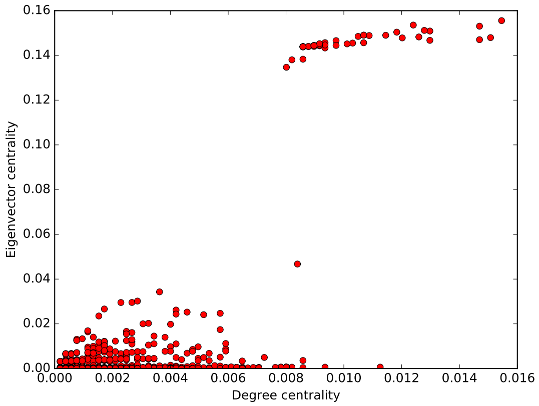


Figure 4: Centrality measures correlation

## 6. Random graph model:

The degree distribution of ER random graph model is not heavy-tailed which makes it unrealistic for real-world graphs. The same goes for the triangles distribution. Moreover, on a random graph the clustering coefficient is too low.

Number of nodes: 200

Number of edges: 2015

Number of connected components: 1

Fraction of nodes in GCC: 1.0

Fraction of edges in GCC: 1.0

Min degree 11

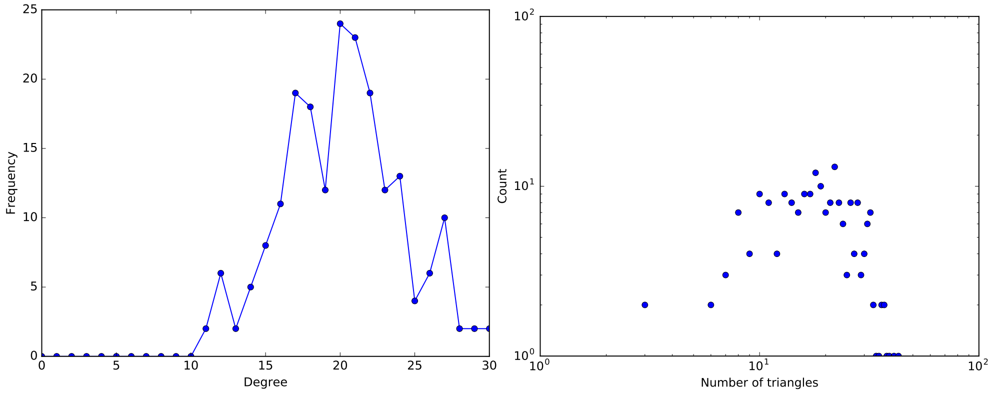
Max degree 30

Median degree 20

Mean degree 20.15

Total number of triangles 1330

Average clustering coefficient 0.100



(a) Degrees distribution

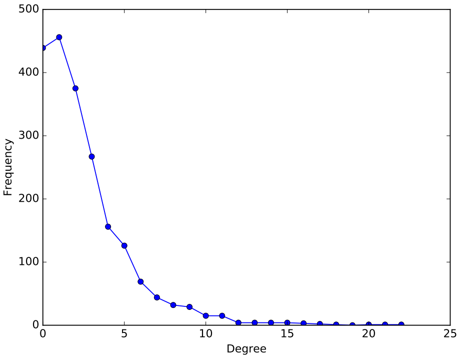
(b) Triangles distribution

Figure 5: Random graph

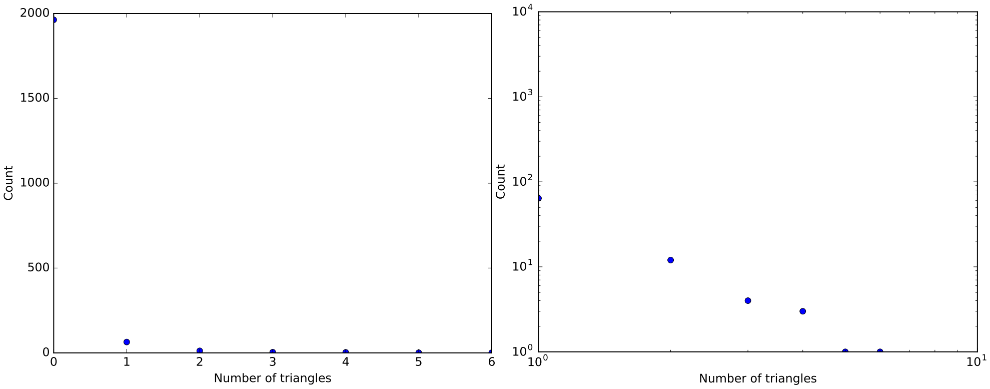
### 7. Kronecker graph model:

With the Kronecker (stochastic) graph model we're able to mimic some propreties of the real-world graphs, namely, the power law degree distribution and the heavy-tailed triangle participation.

Number of nodes: 2048  
Number of edges: 2597  
Number of connected components: 507  
Fraction of nodes in GCC: 0.714  
Fraction of edges in GCC: 0.964  
Min degree 0  
Max degree 22  
Median degree 2  
Mean degree 2.536  
Total number of triangles 41  
Average clustering coefficient 0.003



Degree distribution



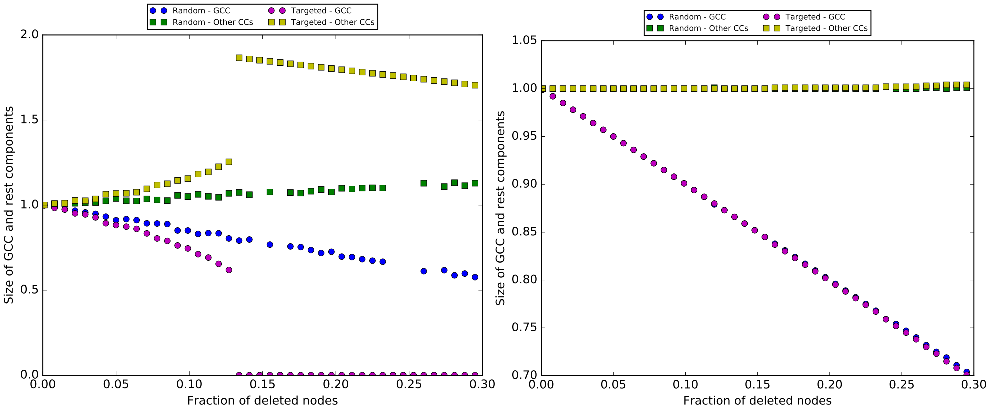
(a) Linear

(b) Logarithmic

Figure 6: Triangles distribution

## 2 Part 2: Robustness of the Network

We note that the real-world networks are robust against random deletion as their degree distribution is fat-tailed i.e just the tail ($\approx 1\%$) determines the topology of the net and randomly deleting vertices is less likely to change the network structure. On the other hand, targeting the tail nodes (highest degrees) of the network affects its structure strongly (up from $p = 0.13$).  
Tested on the random graph (ER) the two strategies have similar effects seeing the degrees distribution (fig 5a) where the nodes with the highest degrees are frequent i.e the most likely deleted with the random strategy.



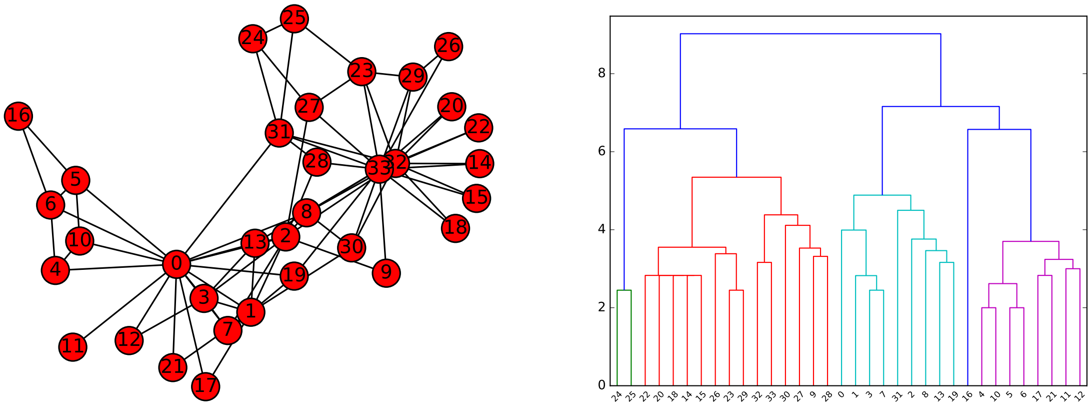
(a) CA-GrQc

(b) Erdõs-Rényi

Figure 7: Triangles distribution

# 3 Part 3: Community Detection

We perform hierarchical clustering using the pairwise distances between the graph nodes. On the dendogram we can see the two main clusters as two branchs occuring at about the same distance ($\approx 6.5$). The first cluster [24,...,28] and the second one [0,..,12].

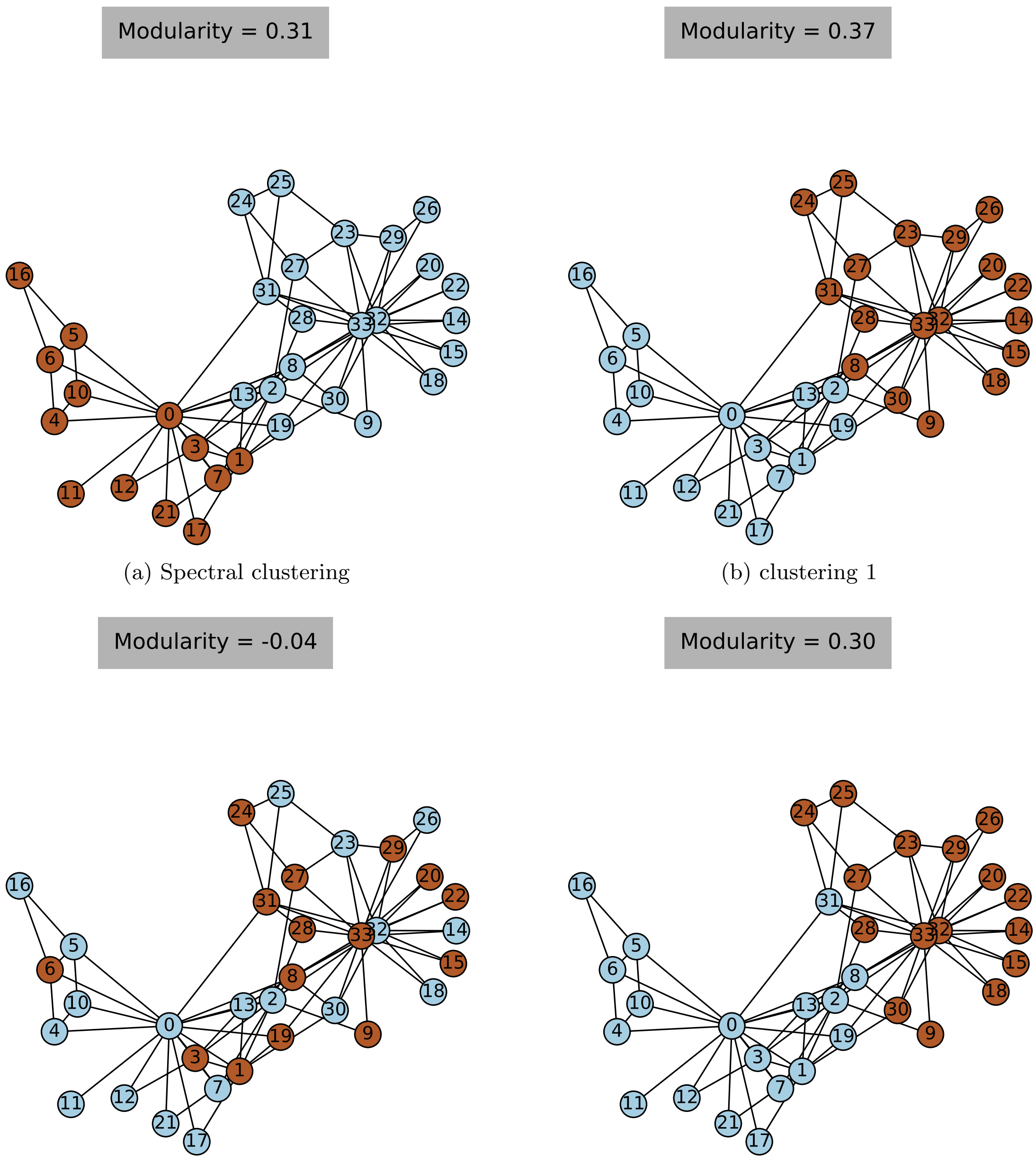


(a) Karate graph

(b) Hierarchical clustering

Figure 8: Zachary' Karate club network

The four clusterings below have different qualities reflected in their modularity measures. Clustering 1 being the best of the 4 (highest modularity and graphical quality) and clustering 2 the worst with a negative modularity and intertwined clusters. The spectral clustering scores slightly better than the hierarchical clustering (+.01) although it differs from the best clustering with 3 nodes $\{2, 13, 19\}$ while the hierarchical is only two nodes away $\{8, 31\}$.



(a) Spectral clustering

(b) clustering 1

(c) clustering 2

(d) Hierarchical clustering

Figure 9: Karate club clusterings