Network Analysis Project

Social Data Management

Mahmoud KOBBI

Paris-Saclay University, Paris, France

February 9, 2018

1 Study case

The *ego-Facebook* dataset of the Stanford Network Analysis Project was used for the purpose of the analysis. It consists of an unidirected, unweighted graph describing a subset of the social circles on the social network.

2 Environment

The study was led on a **Python 2.7** conda environment enriched with network analysis-centered packages **NetworkX**, and **python-louvain**.

- **NetworkX**: NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. Most of the needed algorithms are present in this packages; namely degree and clustering coefficient calculation, all pairs shortest path determination, as well as centrality computation.
- **python-louvain** This module implements community detection.

  It uses the Louvain method described in [**?**]

The full list of dependencies is specified in the project's requirements.txt file .

3 Minimum Requirements

3.1 Number of nodes and edges

For graph $\mathcal{G}$:

$$L = 88234, N = 4039$$

3.2 Graph plot

Graph $\mathcal{G}$ is shown in Fig. 1

3.3 Degrees

3.3.1 Real network

Degree distribution seems to follow a power law, with an average degree equal to

$$\langle k \rangle = 43.6910126269$$

It is shown in Fig. 2.

3.3.2 Random network

Degrees follow the distribution K:

$$K \sim \mathcal{B}(N-1, p)$$

where

$$p = \frac{\langle k \rangle}{N-1}$$

The average degree is equal to the real network's.
Degree histogram is shown in Fig. 3

3.4 Clustering coefficient

3.4.1 Real network

Clustering coefficient distribution is described in Fig. 4.
Roughly, three zones can be observed. The higher the coefficient, the more numerous is the count. One may correlate degree with clustering coefficient as a pattern is observed

3.4.2 Random network

Clustering coefficient distribution is described in Fig. 5. Only one zone is seen. Clustering coefficient is homogenous. Hence, degree and clustering coefficient are independent.

3.5 Distance

3.5.1 Real network

Distance histogram of $\mathcal{G}$ is shown in Fig. 6. Range of possible distances is limited, as

$$\mathcal{D}_{\mathcal{G}} = [\![0, 8]\!]$$

Besides

$$\langle d \rangle = 3.6925068497$$

$$d_{max} = 8$$

$$\frac{\langle d \rangle}{\frac{\ln N}{\ln \ln N}} = 0.9412560239806701$$

Hence, $\mathcal{G}$ is in critical region between small and ultra-small worlds.

3.5.2 Random network

Distance histogram of $\mathcal{R}$ is shown in Fig. 7. Range of possible distances is the same, as

$$\mathcal{D}_{\mathcal{R}} = [\![0, 8]\!]$$

Besides

$$\langle d \rangle = 3.691592636562027$$

$$d_{max} = 8$$

$$\frac{\langle d \rangle}{\frac{\ln N}{\ln \ln N}} = 0.9412560239806701$$

] So, $\mathcal{R}$ is in the critical region too.

4 Extra requirements

4.1 Communities

Louvain algorithm partitions the graph into 16 communities, as shown by figure 8. Some of the communities are interwined. One may ask whether people are members of multiple communities or communities can be fully included in larger ones.

4.2 Triangles

Let $T_{\mathcal{X}}, X \in \{\mathcal{R}, \mathcal{G}\}$. We have:

$$card(\mathcal{G}) = 4836030$$

$$card(\mathcal{R}) = 41127$$

As we can see,

$$card(\mathcal{G}) \gg card(\mathcal{R})$$

, This means that $\mathcal{G}$ is way more tightly knit than $\mathcal{R}$.

4.3 Other centrality measures

4.3.1 Betweenness centrality

Centrality histograms are displayed in Fig 9. In $\mathcal{R}$, we observe that:

$$\langle c_{\mathcal{R}} \rangle \sim \min_{i \in V_{\mathcal{R}}} c_i \sim \max_{i \in V_{\mathcal{R}}} c_i$$

Whereas, in $\mathcal{G}$

$$\langle c_{\mathcal{G}} \rangle \ll \max_{i \in V_{\mathcal{G}}} c_i$$

Hence, $i_{\mathcal{G}}^{center} = \arg\max_{i \in V_{\mathcal{G}}} c_i$ represents the central node of that graph. In contrast, $\mathcal{R}$ doesn't have any.

4.3.2 PageRank

PageRank histograms are displayed in Fig 10.

$$\langle c_{\mathcal{R}} \rangle \sim \min_{i \in V_{\mathcal{R}}} PR_i \sim \max_{i \in V_{\mathcal{R}}} c_i$$

Whereas, in $\mathcal{G}$

$$\langle c_{\mathcal{G}} \rangle \ll \max_{i \in V_{\mathcal{G}}} PR_i$$

Hence, $i_{\mathcal{G}}^{PageRank} = \arg\max_{i \in V_{\mathcal{G}}} PR_i$ must be a celebrity. In contrast, $\mathcal{R}$ doesn't have any star node.

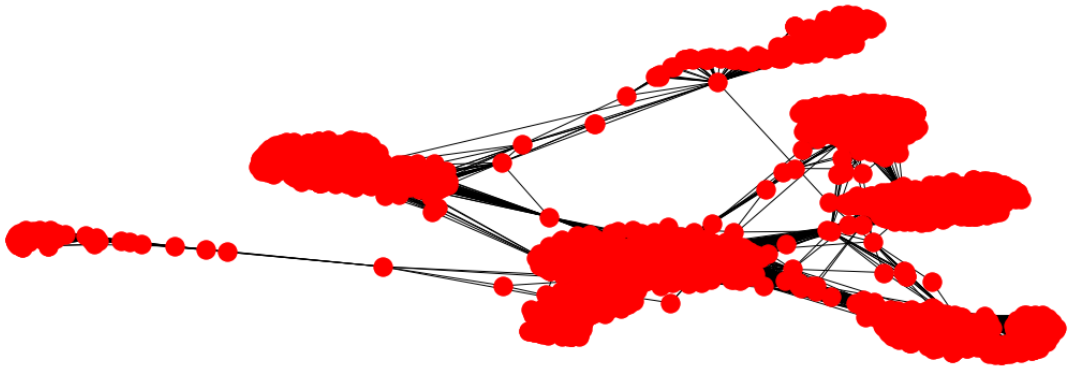

Figure 1: ego-Facebook network

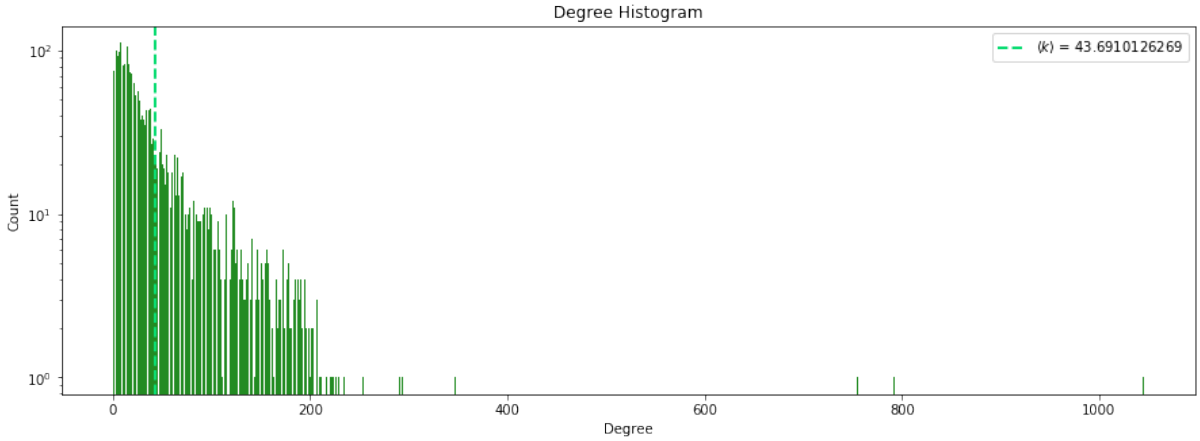

Figure 2: Degree histogram of $\mathcal{G}$

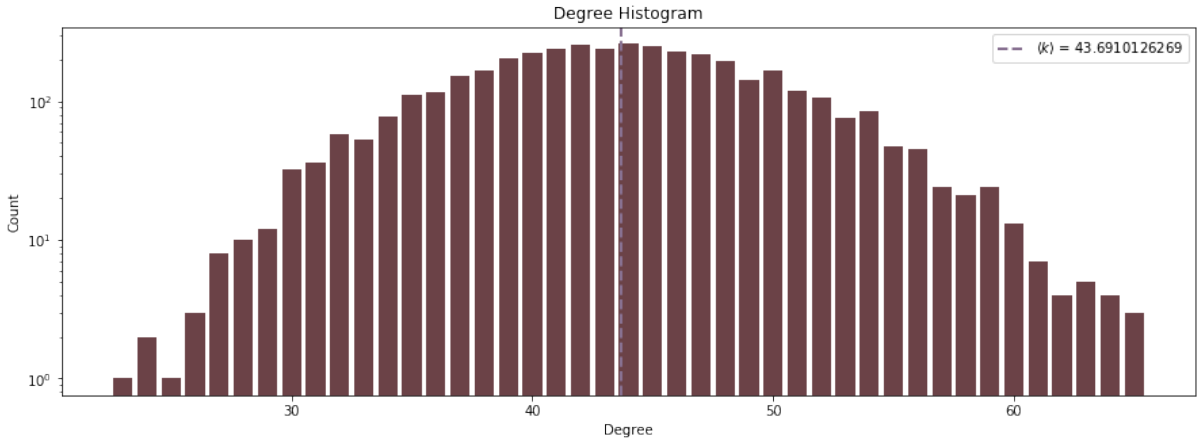

Figure 3: Degree histogram of $\mathcal{R}$

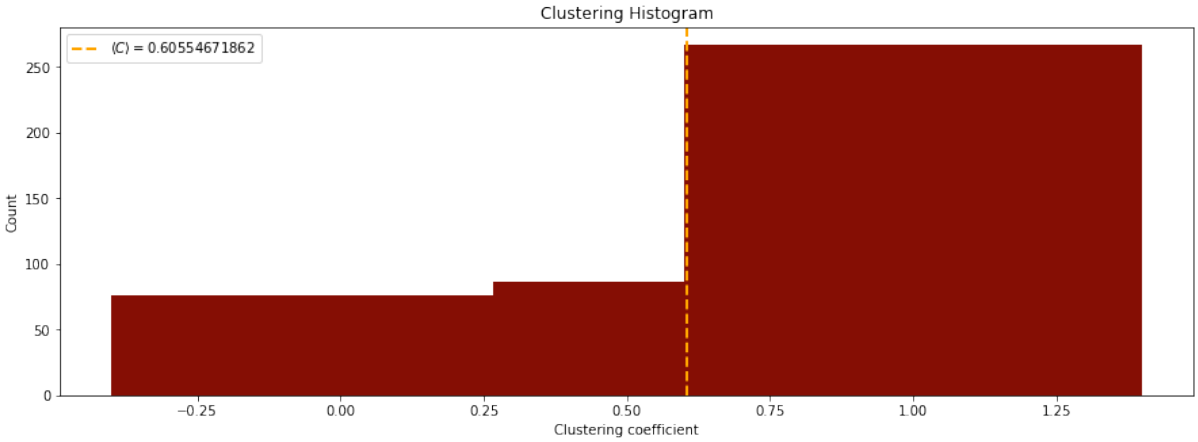

Figure 4: Clustering histogram of $\mathcal{G}$

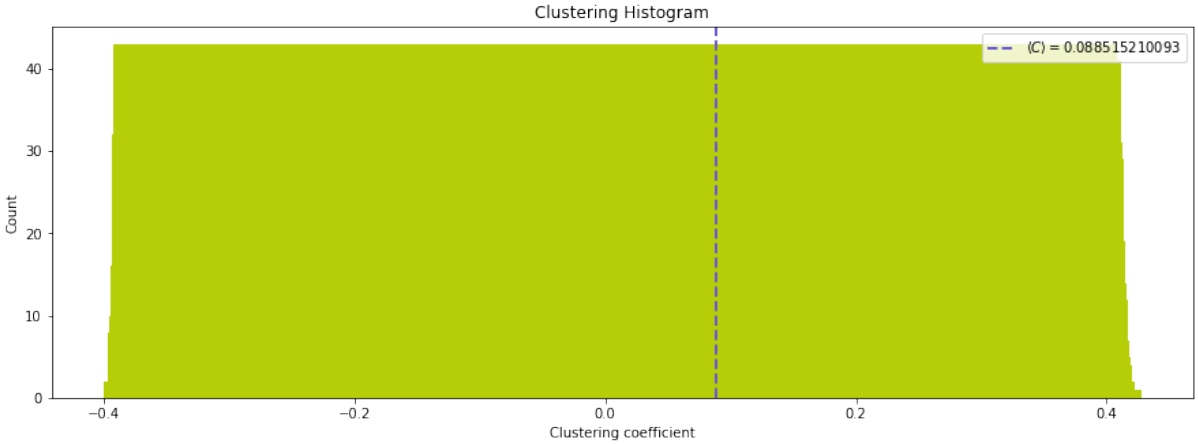

Figure 5: Clustering histogram of $\mathcal{R}$

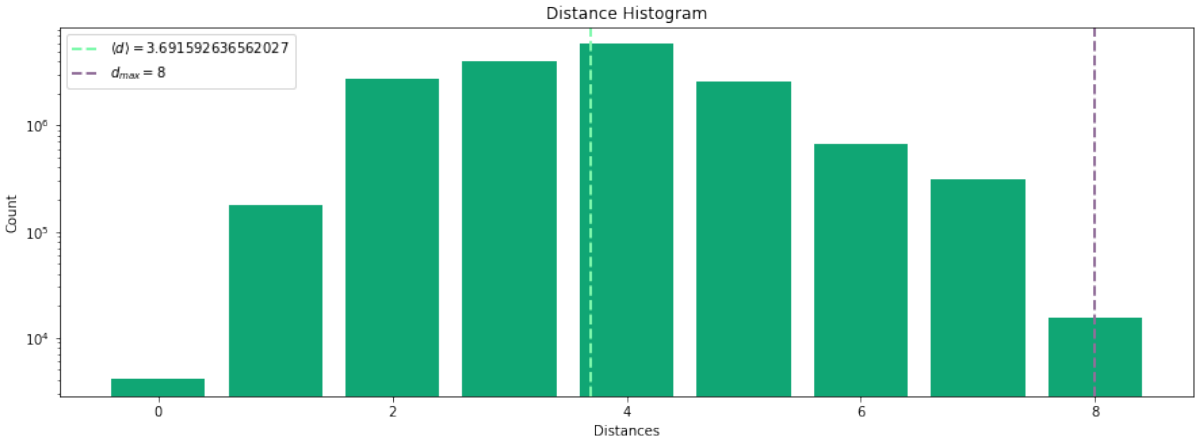

Figure 6: Distance histogram of $\mathcal{G}$

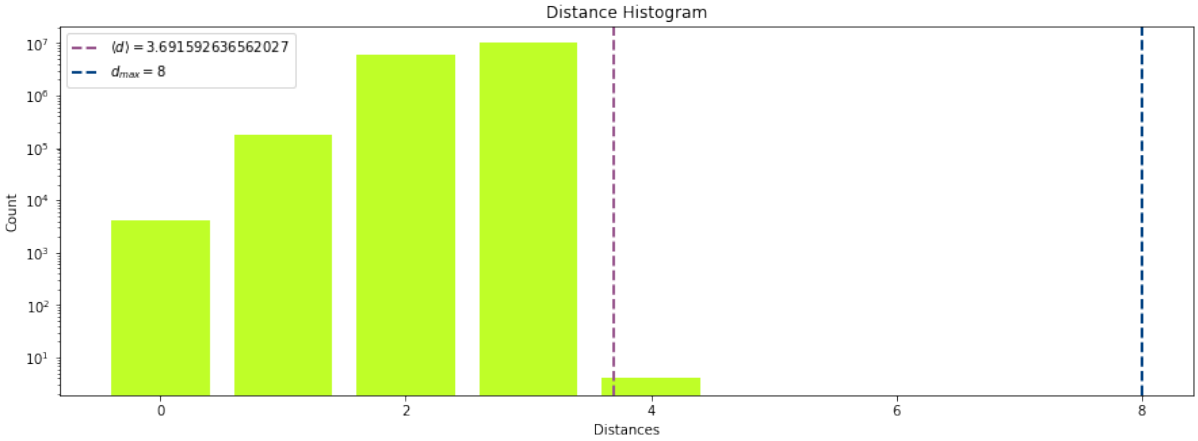

Figure 7: Distance histogram of $\mathcal{R}$

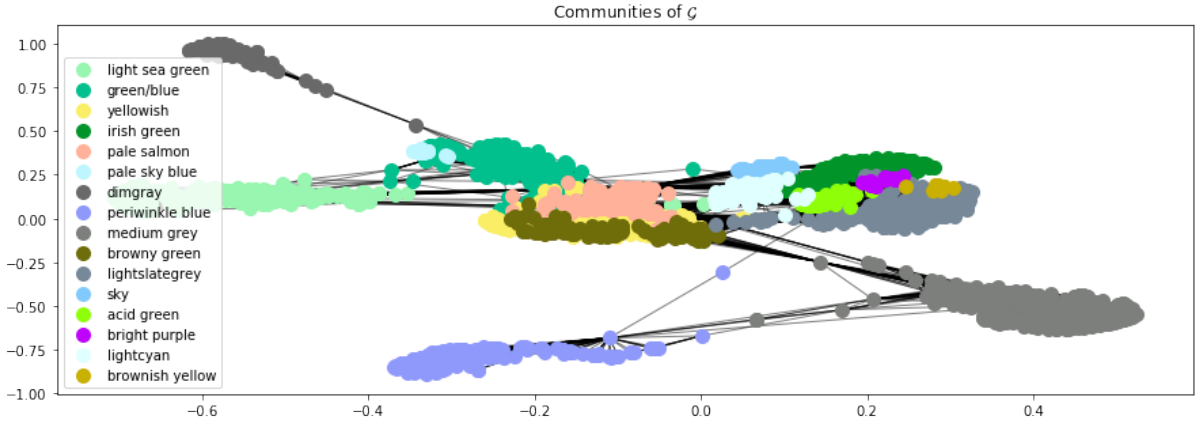

Figure 8: Communities of $\mathcal{G}$

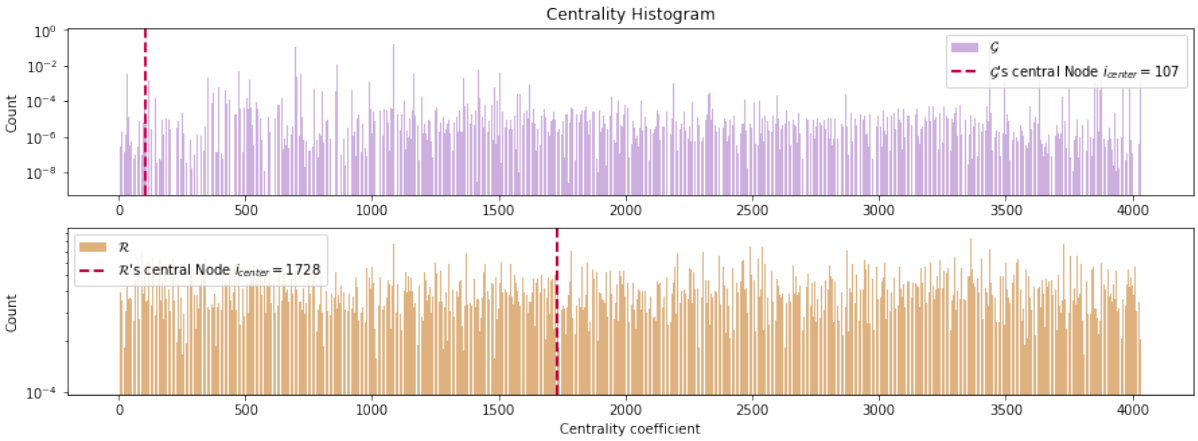

Figure 9: Centrality Histograms

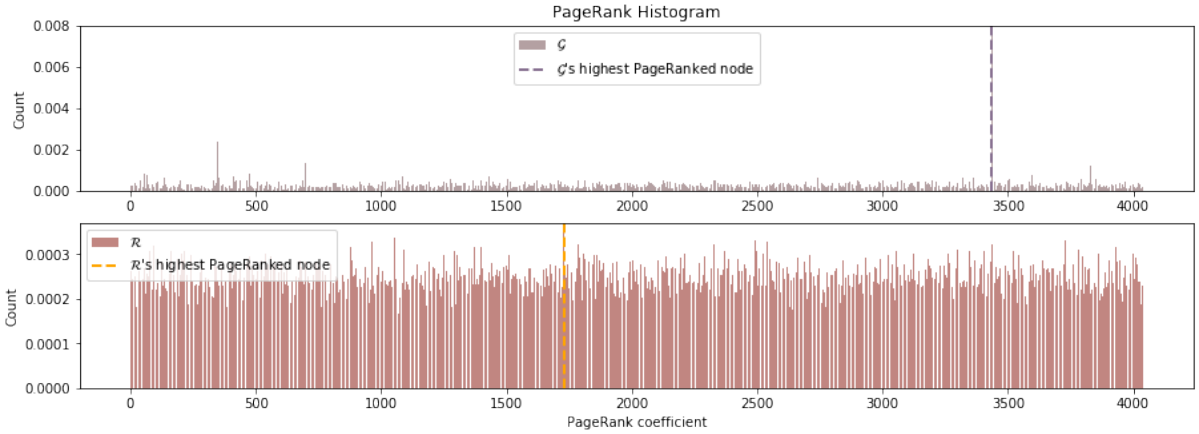

Figure 10: PageRank Histograms