Uncovering Organizational Structures Through Facebook Friendship Networks

David Mateo Carpio R.

Final Individual Report
Complex Networks

2025-02-05

Contents

1 Introduction **1**

2 Methods **2**
2.1 General Network Metrics 2
2.2 Centrality Measures 3
2.3 Degree Distribution Analysis 5
2.4 Mathematical Definitions of Distributions 6
2.5 Comparison with other Models 7
2.6 Community Detection 7
2.7 Visualization 8
2.8 Computational Implementation 8

3 Results **8**
3.1 Network Metrics 8
3.2 Degree Distributions 10
3.2.1 L2 network 10
3.2.2 Erdős-Rényi model degree distributions 11
3.2.3 Barabasi-Albert degree distributions 12
3.3 Centrality Measures 13
3.4 Community detection 15
3.5 Visualization 16

4 Conclusions **16**

A Appendix **17**
A.1 Rank of Ranks 17
A.2 Visualization using Greedy Algorithm 18
A.3 Coding and Github Repository 19

References **19**

Abstract

Social networks facilitate interaction among peers within an organization; however, they inevitably expose personal and organizational information. In this work, using a comprehensive set of network science tools, we analyze the structural properties of a Facebook-based organizational network, corresponding to a high-tech company, denoted as $L2$. We find that the network exhibits a low-density and assortative structure with small-world properties, characterized by a short average path length and moderate clustering coefficients. The degree distribution follows a heavy-tailed structure, deviating from the Erdős-Rényi model, while statistical fitting suggests an exponential distribution, indicating deviations from the the power-law behavior of the Barabási-Albert model. Using centrality measures, we identify a set of top-ranked nodes likely to hold management roles within the organization. In addition, we detect distinct communities, suggesting different organizational roles and functional groups of the company. Finally, we provide a visualization of the network, highlighting structural properties, communities, and key nodes. These findings highlight the risk of exposing organizational structures through publicly available employee data on social networks.

1 Introduction

In recent years, online social networks have emerged as a key resource for organizations, facilitating global collaboration and interaction among employees. However, users of social networks often expose personal details about themselves and their connections through their profile pages, as well as sensitive business information, including details about their workplace [1]. In this aspect, the analysis of informal networks using network science theory within organizations has been widely studied as a means to identify communities and leadership roles. For instance, communication data, such as email logs, can be used to detect communities and leaders within complex organizational networks [2]. Additionally, social networks data from platforms like Facebook can also be analyzed to infer networks of informal social relationships among employees of a given organization [3]. By analyzing mined data, networks of informal social relationships among employees can be extracted for a given organization [4].

In this report, we analyze a network, denoted as $L2$ because of privacy protection, which is available at [5]. This network was built by considering 94219 informal connections among 5524 Facebook users who indicated on their Facebook profiles that they worked for the $L2$ organization [4]. Specifically, the studied network corresponds to a large technology corporation that provides hardware and software products, infrastructure, and other technology services to global customers. According to the company's official web page, there were more than 50000 employees in the year 2013, the time at the network was built. In the constructed network, nodes represents Facebook users who indicate that they work in the target organization, while links represent Facebook friendships between these users. Since Facebook friendships are reciprocal, the network is modeled as an undirected graph.

The aim of this work is to perform a comprehensive structural analysis of the $L2$ network, incorporating global network metrics, centrality-based measures, and community detection algorithms to understand the organization structure and identify influential individuals within the network using

only publicly available social connections. The report is structured as follows: first, we introduce the main concepts and network metrics analyzed. Then, we compute and evaluate these metrics while comparing them to two null models with the same number of nodes and connections. Finally, we detect communities inside the network, and provide a graph of the network highlighting its main features.

2 Methods

In this section, we provide the mathematical and computational tools employed to analyze the network. The methodologies include network structural metrics, centrality measures, statistical analysis, and community detection.

2.1 General Network Metrics

We begin by computing basic structural properties of the network. We analyze a network (or graph) $G = (V, E)$. V is the set of vertices whose elements are the nodes of the graph. E is the edge set, which elements are the links or connections of the network. Then, based on [6], we define the following metrics:

- **Number of Nodes**: Denoted by n, this quantify the network's size. This is simply the size of the vertices set, $n = |V|$.
- **Number Edges**: Denoted by m, this quantify the total number of links in the network. This can be seen as the size of the edge set, $m = |E|$.
- **Degree of a Node**: The degree of a node k_i represents the number of edges connected to it. In an undirected network, it is given by:

$$k_i = \sum_j A_{ij} \tag{1}$$

  where $\mathbf{A}$ is the adjacency matrix, with $A_{ij} = 1$ if there is a link between nodes i and j, and 0 otherwise. Then, the **average degree**, denoted by $\langle k \rangle$, represents the average number of connections per node in the network and σ_k its standard deviation.

- **Network Density**: Measures the proportion of edges present relative to the maximum possible, defined as:

$$\delta = \frac{2m}{n(n-1)}. \tag{2}$$

- **Average Watts-Strogatz clustering coefficient:** Considering the local clustering coefficients C_i, we can obtain the average clustering coefficient $\overline{C}$ by:

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i, \quad C_i = \frac{2t_i}{k_i(k_i-1)}. \tag{3}$$

  where t_i is the number of triangles involving node i and k_i is its corresponding degree [7].

- **Global clustering coefficient**: Also called Newman transitivity index, it is given by:

$$C = \frac{3|C_3|}{|P_2|} \tag{4}$$

where $|C_3|$ is the total number of triangles, and $|P_2|$ the total number of paths of length two.

- **Average Path Length**: We calculate the average shortest path length, $\bar{l}$, given by:

$$\bar{l} = \frac{1}{n(n-1)} \sum_{i \neq j} d(i,j). \tag{5}$$

where $d(i,j)$ is the shortest distance between the node i and the node j.

- **Diameter:** The network diameter, D, is the maximum shortest path between any pair of nodes:

$$D = \max_{i,j \in V} \{d(i,j)\}. \tag{6}$$

- **Degree Assortativity**: Measures the preference of nodes to connect with others of similar degree, expressed as a Pearson correlation coefficient:

$$r = \frac{\frac{1}{m}\sum_{(i,j)\in E} k_i k_j - \left[\frac{1}{2m}\sum_{(i,j)\in E}(k_i + k_j)\right]^2}{\frac{1}{2m}\sum_{(i,j)\in E}(k_i^2 + k_j^2) - \left[\frac{1}{2m}\sum_{(i,j)\in E}(k_i + k_j)\right]^2}. \tag{7}$$

- **Bipartivity**: The bipartivity index b_s measures how close a network is to being bipartite. It is defined as 1 for a purely bipartite network and decreases monotonically as the network deviates from bipartite structure. The bipartivity index is computed using the adjacency matrix $\mathbf{A}$ of the network:

$$b_s = \frac{\mathrm{Tr}\left[\cosh(A)\right] - \mathrm{Tr}\left[\sinh(A)\right]}{\mathrm{Tr}\left[\cosh(A)\right] + \mathrm{Tr}\left[\sinh(A)\right]} = \frac{\mathrm{Tr}\left[\exp(-A)\right]}{\mathrm{Tr}\left[\exp(A)\right]}. \tag{8}$$

2.2 Centrality Measures

Centrality metrics are used to identify influential nodes. Considering the definitions given in [6], we calculate:

- **Degree Centrality**: This measure quantifies the connectivity of a node by counting the number of links it has. For a node i, the degree centrality is defined as:

$$DC(i) = k_i, \tag{9}$$

where k_i is the degree of node i. Degree centrality is a local measure of influence.

- **Closeness Centrality**: Closeness centrality evaluates how quickly a node can reach other nodes in the network. For a node i, it is defined as the reciprocal of the average shortest path length

from i to all other nodes:

$$CC(i) = \frac{n-1}{\sum_j d(i,j)}. \tag{10}$$

where $d(i,j)$ is the shortest path distance between nodes i and j.

- **Betweenness Centrality:** It is defined as the fraction of shortest paths between all pairs of nodes that pass through a given node. For a node k, its betweenness centrality is calculated as:

$$BC(k) = \sum_i \sum_j \frac{\rho(j,i,k)}{\rho(j,k)}, \quad i \neq j \neq k, \rho(i,j). \tag{11}$$

where $\rho(i,j)$ is the total number of shortest paths between nodes i and j, and $\rho(i,k,j)$ is the number of those shortest paths that pass through k.

A common characteristic of the following centrality measures is that they can be expressed in terms of spectral properties of the networks [8]. The ones analyzed in this work are the following:

- **Katz centrality:** The Katz centrality of a node i can be expressed in terms of the eigenvalues and eigenvectors of the adjacency matrix:

$$K(i) = \sum_q \sum_j \psi_j(i)\, \psi_j(q) \frac{1}{1-\alpha\lambda_j}, \tag{12}$$

where $\psi_\mathbf{j}$ is the eigenvector corresponding to the eigenvalue λ_j. The Katz centrality depends on the free parameter α, which must satisfy the condition $0 < \alpha < \lambda_1^{-1}$, where λ_1 is the greatest eigenvalue of the adjacency matrix $\mathbf{A}$.

- **Eigenvector Centrality**: Eigenvector centrality measures the influence of a node within a network by considering not only its direct connections but also the importance of the nodes to which it is connected. For a node i, the eigenvector centralities is:

$$\varphi_\mathbf{1}(i) = \frac{1}{\lambda_1} \left(\mathbf{A}\varphi_\mathbf{1}\right)_i, \tag{13}$$

where λ_1 is the largest eigenvalue of the adjacency matrix $\mathbf{A}$, and $\varphi_\mathbf{1}$ is the eigenvector corresponding to λ_1. The centrality of a node i is proportional to the sum of the centralities of its neighbors, capturing a recursive notion of importance within the network.

- **PageRank Centrality**: evaluates the importance of nodes by considering the structure of incoming links. For a directed graph, the PageRank of a node i is computed using the following matrix-based formulation (if the graph is undirected, it is first converted to a directed graph by treating each undirected edge as two directed edges). First, use the matrix $\mathbf{H}$ defined as:

$$H_{ij} = \begin{cases} \frac{1}{k_i^{\text{out}}} & \text{if there is a directed link from } i \text{ to } j, \\ 0 & \text{otherwise,} \end{cases}$$

where k_i^{out} is the out-degree of node i. Next, we transform $\mathbf{H}$ into a row-stochastic matrix $\mathbf{S}$ to handle nodes with $k_i^{\text{out}} = 0$:

$$\mathbf{S} = \mathbf{H} + \mathbf{a}\left(\frac{1}{n}\mathbf{1}^\top\right),$$

where $\mathbf{a}$ is a vector such that:

$$a_i = \begin{cases} 1 & \text{if } k_i^{\text{out}} = 0, \\ 0 & \text{otherwise,} \end{cases}$$

and n is the number of nodes. Finally, we $\mathbf{G}$ defined as:

$$\mathbf{G} = \alpha\mathbf{S} + \frac{1-\alpha}{n}\mathbf{1}\mathbf{1}^\top,$$

where $\alpha \in [0, 1]$ is the damping factor. The PageRank centrality vector $\mathbf{PR}$ is the principal left eigenvector of $\mathbf{G}$ corresponding to its largest eigenvalue, satisfying:

$$\mathbf{PR} = \mathbf{G}^\top\mathbf{PR}.$$

Finally, the PageRank centrality of a node i is given by:

$$PR(i) = (\mathbf{G}^\top\mathbf{PR})_i. \tag{14}$$

- **Subgraph Centrality**: The subgraph centrality of a node i is computed using the adjacency matrix $\mathbf{A}$ as:

$$EE(i) = \left(\sum_{l=0}^{\infty} c_l\mathbf{A}^l\right)_{ii}. \tag{15}$$

where coefficients c_l are expected to ensure that the series is convergent; they should give more weight to small powers of the adjacency matrix than to the larger ones; and they should produce positive numbers for all $i \in V$ [9].

2.3 Degree Distribution Analysis

The degree distribution is analyzed through the Probability Density Function (PDF) and Cumulative Distribution Function (CDF). We fit these distributions to theoretical models.

- **Probability Density Function (PDF)**: The PDF represents the probability that a randomly chosen node has a specific degree k. We computed it by:

$$p(k) = \frac{n_k}{n}, \tag{16}$$

where n_k is the number of nodes with degree k, and n the total number of nodes.

- **Cumulative Distribution Function**: This distribution function represents the probability of randomly selecting a node with a degree greater than or equal to k. It is computed as:

$$P(k) = \sum_{k'=k}^{\infty} p(k') \tag{17}$$

where $p(k')$ is the probability mass function of the degree distribution [10].

To further understand the degree distributions, we fit them to several theoretical models:

2.4 Mathematical Definitions of Distributions

- **Exponential Distribution**: Models exponential decays. For $k \geq 0$ and $\lambda \geq 0$, the PDF is given by:

$$f(k) = \lambda e^{-\lambda k}, \tag{18}$$

where λ is the rate parameter.

- **Power-law Distribution**: Describes scale-free networks with hubs. Its PDF is computed as:

$$f(k) = Ck^{-\alpha}, \quad k \geq k_{\min}, \tag{19}$$

where α is the power-law exponent, C is a normalization constant and k_{min} represents the minimum degree from which the power-law behavior holds in the network.

- **Log-normal Distribution**: Captures heavy-tailed degree distributions, the PDF is:

$$f(k) = \frac{1}{k\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(k)-\mu)^2}{2\sigma^2}}, \tag{20}$$

where μ and σ^2 are the mean and variance of the logarithm of k.

- **Weibull Distribution**: Flexible model for both short and long-tailed distributions, the PDF is given by:

$$f(k) = \frac{\beta}{\eta}\left(\frac{k}{\eta}\right)^{\beta-1} e^{-\left(\frac{k}{\eta}\right)^{\beta}}, \quad k \geq 0, \tag{21}$$

where β controls the shape, and η scales the distribution.

- **Exponential Modified Gaussian (EMG) Distribution**: Combines a Gaussian core with an exponential decay component. The PDF is given by:

$$f(k) = \frac{\lambda}{2} e^{\frac{\lambda}{2}\left(2\mu+\lambda\sigma^2-2k\right)} \operatorname{erfc}\left(\frac{\mu+\lambda\sigma^2-k}{\sqrt{2}\sigma}\right), \tag{22}$$

where: μ is the location parameter, σ is the scale parameter of the Gaussian component, λ controls the exponential decay, and $\operatorname{erfc}(x)$ is the complementary error function. This distribution captures asymmetric heavy-tailed distributions and it is useful for modeling networks with constrained high-degree nodes.

The parameters for each model were computed using least squares. Then, the models were compared with the observed data to assess the network's degree distribution. Specifically, we calculate the mean square error for the PDF, while we calculate the p-value for the CFD.

2.5 Comparison with other Models

To validate the observed network properties, we compare them with network models with the same number of nodes and edges. We perform 10 realizations of each model and calculate the mean and standard deviation of their metrics. The models that we analyze are:

- **Erdős-Rényi Model**: Generates random graphs by connecting n nodes with m edges, where each edge is chosen uniformly at random. This results in networks with a homogeneous degree distribution and minimal clustering. The probability of an edge (e) between two nodes is:

$$p_e = \frac{m}{\binom{n}{2}}. \tag{23}$$

  The degree distribution, $p(k)$ follows a binomial distribution as function of p_e:

$$p(k) = \binom{n-1}{k} p_e^k (1-p_e)^{n-1-k}, \tag{24}$$

  where k is the degree of a node. For large n, the distribution becomes a Poisson distribution. To ensure that the Erdős-Rényi network maintains the same number of nodes n and the same number of edges m as the analyzed network, we determine the edge probability given by equation 23 based on the specific n and m values of the $L2$ network.

- **Barabási-Albert Model**: Constructs scale-free networks using preferential attachment, where new nodes are more likely to connect to existing nodes with higher degrees. This model is useful for identifying whether the observed network exhibits hub-dominated structures. The probability of a new node connecting to an existing node i is proportional to its degree k_i:

$$p(i) = \frac{k_i}{\sum_j k_j}.$$

  This process leads to a power-law degree distribution:

$$p(k) \sim k^{-\gamma}, \quad \gamma \approx 3. \tag{25}$$

  To ensure that the Barabási-Albert network maintains the same number of nodes n and the same number of edges m as the real network, we set the parameter of the attachment as $m_{\text{BA}} = m/n$.

2.6 Community Detection

We employ algorithms to identify communities within the network:

- **Greedy Modularity Optimization**: Iteratively merges nodes into communities to maximize modularity [11]. This method efficiently identifies community structures by optimizing the modularity score, making it suitable for large networks due to its significantly reduced computational complexity compared to earlier algorithms [12].

- **Louvain Algorithm**: Optimizes modularity through hierarchical clustering [13]. This method iteratively refines the community structure, efficiently uncovering hierarchical relationships by aggregating communities at multiple levels.

To assess the significance of detected communities, we evaluate partitions using modularity (Q), performance and coverage.

- **Modularity:** Measures the quality of a given partition of a graph into communities by comparing its structure to a random baseline, ensuring that detected communities are not merely random clusters [14]. It is defined as:

$$Q = \sum_{r=1}^{n_C} \left[\frac{m_r}{m} - \left(\frac{\sum_{j \in V_r} k_j}{2m} \right)^2 \right], \tag{26}$$

  where m_r is the number of edges within community r, m is the total number of edges in the network, and k_j is the degree of node j in community r.

- **Coverage:** Measures the fraction of intra-community edges relative to the total number of edges in the network, providing an intuitive assessment of how well the partition captures internal connectivity.

- **Performance:** Evaluates the fraction of correctly classified node pairs, considering both intra-community and inter-community connections, allowing for a broader assessment of community detection effectiveness.

2.7 Visualization

The networks were visualized using NetworkX library. Node sizes and colors were mapped to degree and community memberships, respectively. We highlight the top ranked nodes in the centrality measures by changing their shape and we distinguish the edges corresponding to the diameter of the network.

2.8 Computational Implementation

All analyses were conducted using Python, primarily leveraging the NetworkX library. Custom algorithms were implemented for specific metrics not available in standard libraries. The source code is included in a repository which can be found in the Appendix A.3.

3 Results

3.1 Network Metrics

The analysis of the $L2$ network reveals several key structural properties, as summarized in Table 1. First, we corroborate that the network consists of $n = 5524$ nodes and $m = 94219$ edges, with a relatively low density of $\delta = 6.17 \times 10^{-3}$. This is typical for large real-world networks, where most nodes are not directly connected but remain accessible through a small number of intermediaries [15]. Also, we compute that the average degree of the networks is $\langle k \rangle = 34.1$, which implies that on average, each node has 34 connections. However, the standard deviation of the degrees, $\sigma_k = 31.8$,

suggests a high variability in connectivity leading to heterogeneous degree distribution. Then, it is expected that few nodes (hubs) have a high number of connections, while most nodes have significantly fewer. As reference, the average Facebook user in 2012 had around 190 friends on the platform [16].

Metric	**L2**
Number of nodes (n)	5524
Number of edges (m)	94219
Network density (δ)	6.17×10^{-3}
Average degree $\langle k \rangle$	34.1
Standard deviation of the degrees (σ_k)	31.8

Table 1: Fundamental information of the $L2$ network.

Table 2 presents a summary of the structural metrics computed for the $L2$ network and compared with those of the null models. The first feature that we notice is the high clustering coefficient of the real world network. The average clustering coefficient (Watts-Strogatz) of the $L2$ networks is $\overline{C} = 0.36$, while the global clustering coefficient is $C = 0.22$, indicating that the network exhibits a moderate level of local connectivity. These values are significantly larger than those of the null models. For the ER model, both clustering coefficients are close to 6×10^{-3}, while for the BA model, these are approximately 2×10^{-2}. This confirms that the real network exhibits a tendency for local connectivity, meaning that individual in the organization tends to form local groups. This property is typical in social networks, where clustering is naturally enhanced by common interests, shared work environments, or social interactions [17].

Another features of the $L2$ networks are its diameter, $D = 9$ and its average path length, $\bar{l} = 3.50$. These values support the idea that the network has small-world characteristics, where most nodes can be reached within few steps. This path length is also shorter than the average distance between pairs of users in the complete Facebook network, which in 2011 was reported as 4.7 [18]. On the other hand, both the ER and BA models exhibit shorter path lengths ($\bar{l} \approx 2.8$ for ER and $\bar{l} \approx 2.7$ for BA) and a smaller diameter, $D = 4$, for both. These differences could be due to the way the $L2$ network was built. Since it only includes publicly available Facebook friendships and requires that users explicitly list the $L2$ corporation as their employer, many connections may be missing, leading to longer paths compared to the more densely connected random models.

Furthermore, the degree assortativity coefficient for the $L2$ network is $r = 0.10$, indicating that nodes with similar degrees are more likely to be connected. This behavior is characteristic of social networks, where individuals with similar numbers of connections tend to interact more frequently, a pattern also observed in Facebook networks [19]. In contrast, the ER model shows near-zero assortativity ($r \approx -1.6 \times 10^{-3}$), indicating that connections are formed randomly. The BA model exhibits a negative assortativity of $r \approx -0.01$, which is expected due to is the preferential attachment mechanism. In the model, new nodes are more likely to connect to highly connected hubs rather than to other low-degree nodes, leading to disassortative.

Lastly, the bipartivity index is $b_s = 1.20 \times 10^{-25}$. This extremely low result suggests that

the network structure does not naturally separate into two distinct sets of nodes. Similarly, ER ($b_s \approx 7.76 \times 10^{-9}$) and BA ($b_s \approx 3.17 \times 10^{-20}$) models show low values of bipartivity. In general, neither the ER nor BA models fully capture the structural properties of the analyzed real-world network.

Metric	L2 (Real)	ER Model	BA Model
Average Clustering Coefficient $(\overline{C})$	3.61×10^{-1}	$(6.13 \pm 0.11) \times 10^{-3}$	$(2.48 \pm 0.07) \times 10^{-2}$
Global Clustering Coefficient (C)	2.22×10^{-1}	$(6.14 \pm 0.12) \times 10^{-3}$	$(2.34 \pm 0.02) \times 10^{-2}$
Average Shortest Path Length $(\overline{l})$	3.503	2.802 ± 0.001	2.706 ± 0.003
Diameter (D)	9	4 ± 0	4 ± 0
Degree Assortativity (r)	1.04×10^{-1}	$(-1.62 \pm 3.39) \times 10^{-3}$	$(-1.16 \pm 0.33) \times 10^{-2}$
Bipartivity Index (b_s)	1.20×10^{-25}	$(7.76 \pm 0.72) \times 10^{-9}$	$(3.17 \pm 0.86) \times 10^{-20}$

Table 2: Comparison of graph metrics for the real network ($L2$) and the Erdős-Rényi (ER) and Barabási-Albert (BA) models. Values for the random models are reported as mean $\pm$ standard deviation from 10 random realizations.

3.2 Degree Distributions

3.2.1 L2 network

A further understanding can be done by considering the degree distribution. Figure 1 shows the PDF for the degree distribution of the $L2$ network, along with several fitted distributions. The main plot displays the histogram of degree values, while the inset graph provides a plot in semi-log scale for each degree value. Table 3 lists the estimated parameters for each distribution and their corresponding mean squared error (MSE) (calculated for the inset graph). A key observation is that, unlike many real-world social networks [20], the $L2$ network does not exhibit a power-law degree distribution, as shown in Figure 1. This aligns with previous findings indicating that Facebook's global network does not exhibit strict scale-free behavior [16]. Instead, the degree distribution is better characterized by distributions with an exponential decay, such as the exponential distribution or the exponential modified Gaussian (EMG). Also, log-normal and Weibull distributions appear to capture the behavior of the distribution, while we can discard the power-law distribution. We do not decide the best fit at this point as the tail's data is noisy.

Distribution	Parameters	MSE
Exponential	$\lambda = 0.029$	2.96×10^{-6}
Log-normal	$\mu = 3.14$, $\sigma = 0.88$	2.38×10^{-6}
Exponential Modified Gaussian	$\mu = 5.15$, $\sigma = 1.32$, $\lambda = 0.037$	1.09×10^{-6}
Weibull	$\eta = 29.21$, $\beta = 1.37$	6.30×10^{-6}
Power-law	$\gamma = 1.10$	6.57×10^{-5}

Table 3: Estimated parameters using least squares and the corresponding mean squared errors (MSE) considering different fitted distributions for the PDF of the $L2$ degree distribution.

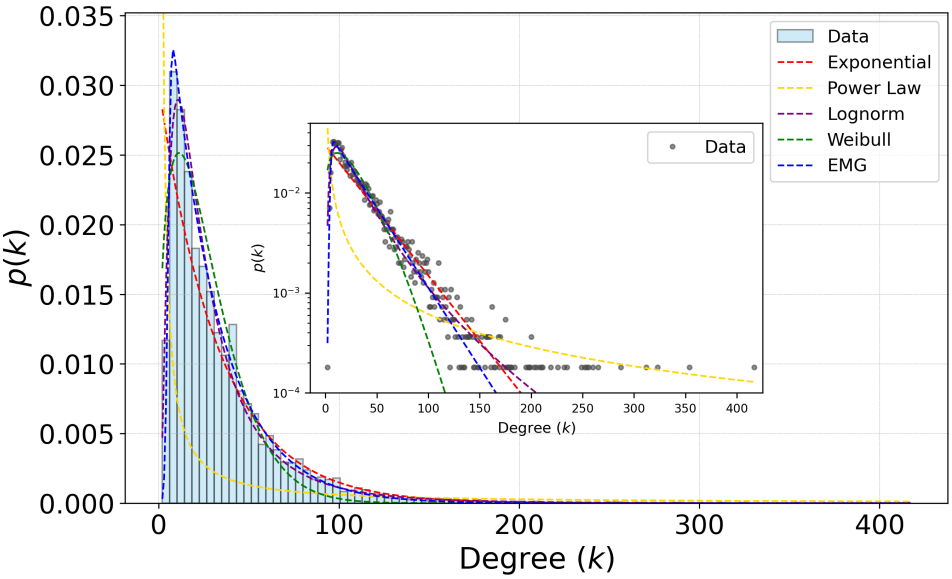


Figure 1: PDF of the $L2$ degree distribution in linear scale, alongside fitted distributions: Exponential (red), Power Law (yellow), Lognormal (purple), Weibull (green), and Exponentially Modified Gaussian EMG (blue). Inset: Semi-log plot of the degree distribution. We discard the power-law distribution.

To reduce noise, specially in the tail of the probability distribution, we computed the CDF using equation 17 [10]. Figure 2 shows the CDF for the degree distribution $P(k)$ of the $L2$ network, alongside with various fitted distributions. We corroborate that the distribution exhibits an exponential decay, as evident from the straight-line behavior in the semi-log plot. This results suggests that while a few nodes have significantly higher degrees, the majority of nodes maintain a relatively small number of connections. However, high-degree nodes are significantly less frequent than in scale-free networks. The inset graph, which plots the CDF on a semi-log scale, confirms the lack of a clear power-law regime as the distribution posses a curvature in this scale. One explanation could arise since Facebook enforces constraints (such as a maximum of $5,000$ friends per user), which prevents the formation of big hubs.

Table 4 presents the parameters used to fit various distribution functions to the CDF. These parameters closely correspond to those obtained from the PDF fitting. Additionally, we assess the quality of the fits using the p-value test. For this criteria, we find that the exponential distribution provides the best fit, followed by the log-normal distribution. The exponentially modified Gaussian (EMG) distribution also shows statistical significance, whereas the Weibull distribution fails to adequately capture the behavior of the degree distribution.

3.2.2 Erdős-Rényi model degree distributions

We can now compare the degree distributions for the $L2$ network with other models. Figure 3 compares the degree distributions of the ER model with the real-world $L2$ network. Figure 3 (a) shows that the PDF of the ER model exhibits an approximately Poisson-like distribution, which is expected given the high number of edges. The distribution peaks at $k = 34$, close to the average degree of the $L2$ network. We also notice that the nodes' degrees lie in the range $k \in [12, 59]$, lacking high-degree nodes found in the real data.

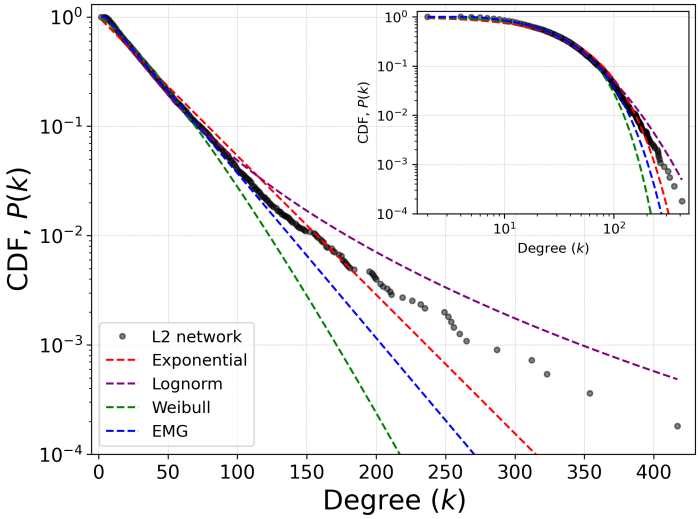


Figure 2: CDF of the $L2$ degree distribution compared against fitted distributions: exponential (red), lognormal (purple), Weibull (green), and exponentially modified Gaussian EMG (blue). The linear trend in the semi-log scale suggests an exponential distribution, while the log-log inset reveals curvature, indicating deviation from a power-law behavior.

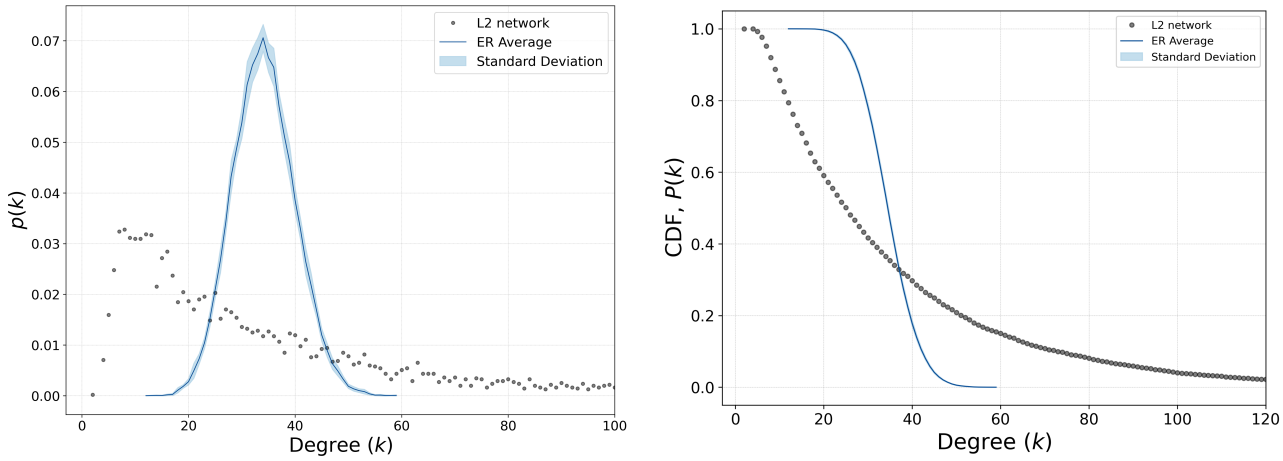
Distribution	**Parameters**	**p-value**
Exponential	$\lambda = 0.030$	0.91
Log-normal	$\mu = 3.17, \sigma = 0.84$	0.44
Exponential Modified Gaussian	$\mu = 4.17, \sigma = 0.56, \lambda = 0.035$	0.37
Weibull	$\eta = 35.20, \beta = 1.22$	0.00

Table 4: Estimated parameters using least squares for different fitted distributions for the CDF of the $L2$ degree distribution. The p-value for each fit shows the maximum value for the exponential distribution.

Figure 3 (b) shows the CDF alongside the real network data. The ER model exhibits a much faster decay, contrasting with the heavy-tailed nature of the $L2$ network. The discrepancy arises because the ER model generates links randomly, resulting in a more homogeneous degree distribution that fails to capture the presence of highly connected hubs characteristic of social networks. Thus, we can conclude that ER model does not describe the $L2$ network.

3.2.3 Barabasi-Albert degree distributions

Figure 4 (a) shows the PDF, highlighting that the BA model exhibits a heavy-tailed distribution, with the presence of high-degree nodes (hubs). However, the model does not fully reproduce the real degree distribution, particularly for intermediate-degree, where the real network follows an exponential decay rather than a power-law distribution. Additionally, the BA model fails to accurately represent the low-degree nodes. This discrepancy arises because, in order to maintain the same number of edges and nodes as the real network, each newly introduced node in the BA model was required to attach to 17 existing nodes, preventing the emergence of nodes with very low degrees.



(a) (b)

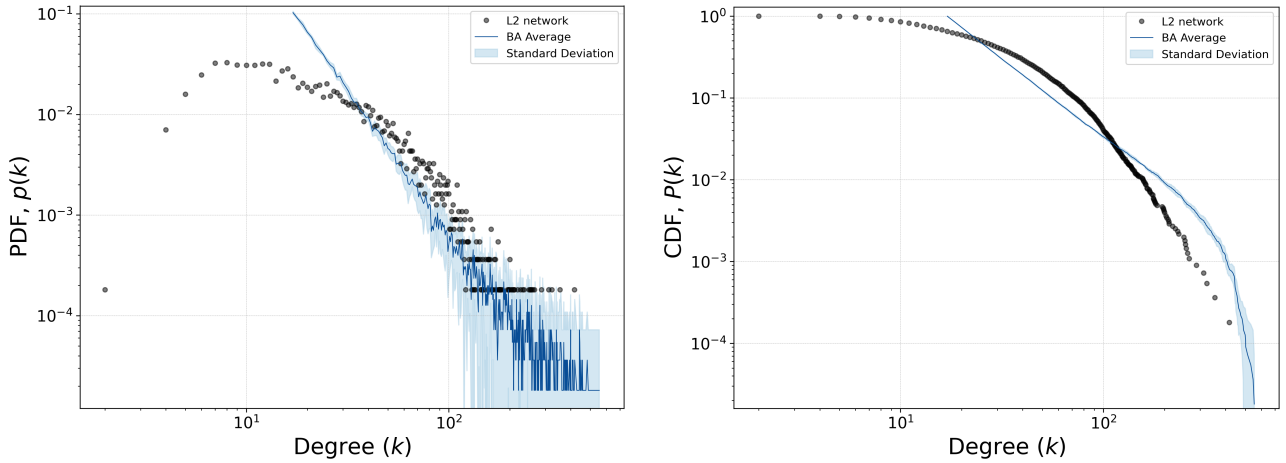
Figure 3: Comparison of the degree distributions for the $L2$ network (black dots) and the Erdős-Rényi (ER) model (a) PDF and (b) CDF. The results of the null model are obtained from the average (blue line) of 10 random realizations of an ER network with the same number of nodes and edges than the real network. The shaded region represents the standard deviation.

These differences are also evident in the CDF. Figure 4 (b) shows the CDF for the BA model degree distribution alongside the real data. We corroborate that the BA model fail in reproducing the behavior of the real data. In the log-log scale, the real data exhibits curvature, while the BA model follows the expected straight-line trend. Although the BA model captures the slower decay of the degree distribution more accurately than the ER model, the $L2$ network still deviates from the theoretical BA prediction. As we mentioned in Section 3.2.1, the $L2$ is better represented by an exponential degree distribution.

3.3 Centrality Measures

Table 5 highlights the top 25 nodes ranked by different centrality measures, where nodes are numerically labeled to anonymize Facebook identities while preserving the structural analysis. This ranks provides insights into the roles and influence of specific employees within the network. For instance, degree centrality highlights highly connected individuals who may act as communication hubs, while closeness centrality identifies employees best positioned to efficiently reach others. Also, betweenness centrality reveals key intermediaries who bridge different groups, facilitating cross-team interactions. Meanwhile, Eigenvector and Katz centralities emphasize employees connected to other influential individuals, likely representing high-ranking executives. PageRank centrality highlights individuals strategically positioned within influential subgroups, and subgraph centrality detects employees deeply integrated into specialized teams.

The rankings reveal that node 260 plays a fundamental role in the network, as it consistently ranks highest across multiple measures, including degree, closeness, eigenvector, Katz, and subgraph centralities. This suggests that node 260 functions as a key hub with extensive connectivity and influence within the organization. On the other hand, betweenness centrality highlights nodes such as 231,



(a) (b)

Figure 4: Comparison of the degree distributions for the real-world $L2$ network (black dots) and the Barabasi-Albert (BA) model. (a) PDF and (b) CDF. The results for the BA model correspond to the average over 10 random realizations (blue line), with the shaded region representing the standard deviation. The BA model does not accurately describe the $L2$ network, which is better characterized by an exponential degree distribution.

549, and 1760, identifying them as key intermediaries that facilitate connections between different parts of the network. Interestingly, PageRank ranks node 76 at the top, suggesting that although this node is not necessarily the most connected one, it is well-positioned within an influential neighborhood.

According to previous studies on this network [4], management positions (e.g., team leaders, project managers, vice presidents) were identified with 100% accuracy among nodes ranked in the top 20 for degree, closeness, betweenness, eigenvector, and PageRank centralities, except for degree centrality, which exhibited a 90% precision. These conclusions were drawn using publicly available data sources, such as Facebook, LinkedIn and Google search engines. However, in our case, we were unable to directly verify management roles, as the dataset used for this study does not contain personal information about the nodes.

Another interesting feature is the strong similarity between Eigenvector, Katz, and Subgraph centralities rankings. Specifically, Katz and Eigenvector centralities share 24 out of the top 25 ranked nodes, while Eigenvector and Subgraph centralities share all top 25 ranked in the same positions. This similarity aligns with their spectral properties, which emphasize influence through highly connected neighbors. In contrast, these three measures differ from betweeness centrality, sharing only 7 common nodes.

To gain a broader perspective, in Table 7 (Appendix A.1), we extend the analysis considering a rank of ranks based on the position of nodes across all centrality measures in 5. This ranking includes a total of 51 nodes, each appearing in at least one of the top 25 lists of Table 5, identifying them as the most influential nodes within the network. Among these, four nodes appear in all the centrality measures ranks, and clearly outstand in the metrics, these are the nodes: 260, 312, 524 and 249. Additionally, node 571 stands out, appearing in six of the rankings, further reinforcing its relevance

Ranking	Degree	Closeness	Betweenness	Eigenvector	Katz	PageRank	Subgraph
1	260	260	231	260	260	76	260
2	76	586	549	571	571	260	571
3	312	524	951	312	312	951	312
4	571	571	260	524	524	142	524
5	524	249	1760	249	249	64	249
6	586	202	852	1522	315	312	1522
7	249	563	64	315	1540	852	315
8	1540	247	586	1540	1522	571	1540
9	372	231	76	481	481	524	481
10	315	169	202	1763	372	399	1763
11	951	262	563	372	1763	586	372
12	64	312	1262	860	860	249	860
13	142	1262	247	822	586	372	822
14	852	570	439	509	822	315	509
15	1971	356	312	586	509	1540	586
16	356	1102	142	508	1102	311	508
17	570	2507	169	1568	508	202	1568
18	1102	1139	282	1102	1568	2507	1102
19	481	1077	249	570	570	1736	570
20	2507	852	855	2149	2149	570	2149
21	202	372	372	1986	1986	356	1986
22	1705	64	262	793	356	1102	793
23	860	257	524	1705	1971	2510	1705
24	1262	1987	1139	1971	1705	2510	1971
25	311	481	570	356	1262	169	356

Table 5: Top 25 nodes ranked by centrality measures. Each column represents the ranking of nodes based on a specific centrality metric.

in multiple aspects of network connectivity and influence.

3.4 Community detection

With the aim of getting insights about the internal structure of the organization, we apply the Greedy Optimization and Louvain algorithms to detect community structures within the network. Results are summarized in Table 6. Using the Greedy algorithm, we identify 4 communities, which are in agreement with the reported results [4]. For this method, we also obtain a modularity value of $Q = 0.51$ indicating a moderately strong community structure, with high coverage (0.88) and performance (0.67). In contrast, the Louvain algorithm detects 11 communities, achieving a higher modularity value of $Q = 0.64$ along with high performance (0.88) but lower coverage (0.73) values.

These results highlight that while the Greedy algorithm finds fewer, broader communities, the Louvain method detects finer divisions in the organization. The choice of algorithm depends on the desired level of resolution for community detection. Different communities may correspond to various departments within the organization, where higher or lower resolution can reveal different

levels of organizational structure [21]. We prefer visualize the network with the Louvain Algorithm as it exhibits greater modularity.

Algorithm	Number of communities	Modularity	Coverage	Performance
Greedy Optimization	4	0.51	0.88	0.67
Louvain	11	0.61	0.73	0.88

Table 6: Comparison of community detection algorithms applied to the $L2$, showing the number of communities, modularity, coverage, and performance for each method.

3.5 Visualization

Figure 5 presents a visualization of the $L2$ organization network, reconstructed from Facebook friendships. Nodes represent Facebook users affiliated with the corporation, while links indicate friendship connections between them. The color of each node corresponds to its assigned community, as detected by the Louvain algorithm, illustrating the network's modular structure. Node sizes are proportional to their degree centrality, highlighting the most connected individuals within the network. Additionally, nodes ranked in centrality measures are plotted with a square shape and labeled, allowing a clearer identification of key nodes in the network. Finally, edges highlighting red are those corresponding to the diameter of the network.

This visualization provides a representation of the overall structure within the organization, showcasing how employees form distinct interconnected groups. A way to determine the roles associated with each community is by searching for professional information about its members. In this case, we were unable since the nodes are anonymized. However, another visualization is provided in Appendix A.2, where the communities are identified by the Greedy algorithm and categorized according to [4].

4 Conclusions

In this work, we realized an exhaustive analysis of a network built using publicly available data mined from Facebook users who indicated their association with a large technology organization. By applying network science tools, we uncovered valuable insights into the organization. For instance, centrality measures identify leadership roles withing the organization, while community detection algorithms revealed information about the internal organization structure. Additionally, we extracted information that the organization itself might not be aware of, such as distances between employees within the network or the degree of assortativity in their connections. These findings reveal the vulnerability of internal organizational information due to the exposure of employees' personal data. Future work could explore multilayer networks by incorporating data from other social platforms, such as LinkedIn, to gain a more comprehensive understanding of organizational structures and professional interactions.

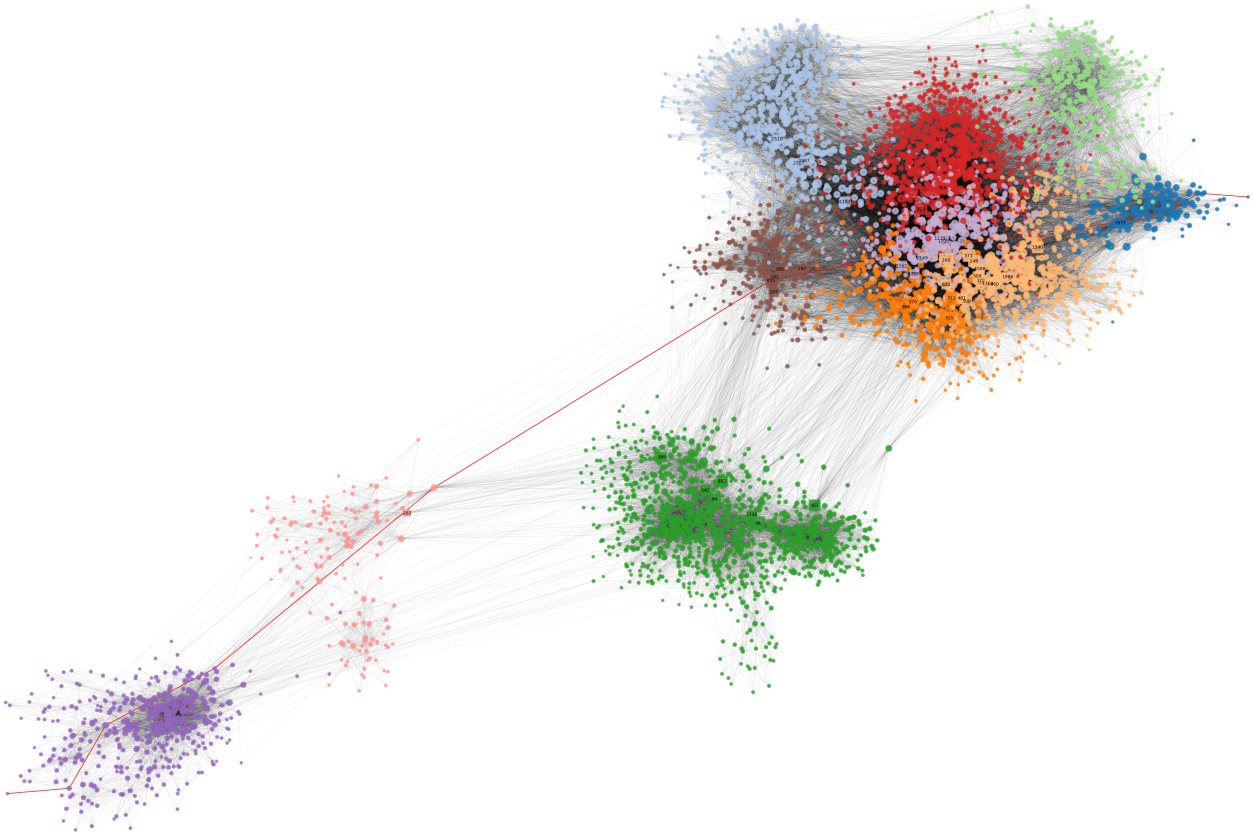


Figure 5: $L2$ company networks visualization. Network reconstructed from Facebook friendships. Nodes represent Facebook users that belong to the corporation and links Facebook friendships. The color of the nodes represent their communities using Louvain algorithm. Size of the nodes ar proportional to degree centrality. Also, the shape of the nodes that are ranked in the centrality measures have a square shape and are labeled. Edges in red are those corresponding to the diameter of the network.

A Appendix

A.1 Rank of Ranks

To provide a comprehensive ranking of node importance, we aggregate scores across multiple centrality measures. Each node's ranking score is determined by assigning 25 points to the top-ranked node, 24 points to the second, and so on, down to 1 point for the 25th ranked node in each centrality measure.

The Total Score column in Table 7 represents the sum of points each node accumulated across all centrality measures, allowing us to identify the most influential nodes overall. The Appearances column indicates the number of times each node appeared in the top 25 rankings across different centrality metrics. Nodes that appear in all seven rankings have a broader structural influence, while those appearing in fewer rankings may be more specialized in their importance.

Position	Node	Total Score	Appearances	**Position**	Node	Total Score	Appearances
1	260	171	7	27	247	31	2
2	312	137	7	28	1262	30	4
3	571	134	6	29	508	29	3
4	524	130	7	30	169	26	3
5	249	124	7	31	1568	26	3
6	586	112	7	32	549	24	1
7	372	86	7	33	2507	23	3
8	315	86	5	34	1760	21	1
9	1540	84	5	35	262	19	2
10	76	66	3	36	2149	18	3
11	951	61	3	37	1971	18	4
12	481	59	5	38	399	16	1
13	64	58	4	39	1986	15	3
14	1522	58	3	40	439	12	1
15	852	57	4	41	1705	12	4
16	202	50	4	42	311	11	2
17	570	49	7	43	1139	10	2
18	1102	48	6	44	282	8	1
19	1763	47	3	45	793	8	2
20	860	45	4	46	1736	7	1
21	142	45	3	47	1077	7	1
22	231	42	2	48	855	6	1
23	822	38	3	49	2510	5	2
24	509	35	3	50	257	3	1
25	563	34	2	51	1987	2	1
26	356	32	6				

Table 7: Rank of Ranks: Nodes ranked by total score across all centrality measures. A node is assigned 25 points if a node appears first in one centrality measure, while it is assigned one point if it appears 25^{th} The 'Appearances' column indicates the number of centrality measures in which each node was ranked.

A.2 Visualization using Greedy Algorithm

An alternative way to visualize Figure 5 is by considering the communities detected using the Greedy optimization algorithm represented by the nodes' colors. These communities align with those identified in the main reference [4].

Following previous findings, the detected communities can be categorized as follows:

- Green: International Senior management (Senior management, Senior researchers)
- Blue: East Asian Headquarter (management and consultants)
- Orange: East Asian Headquarter (R&Ds and consultants)
- Red: The company's amateur sports team.

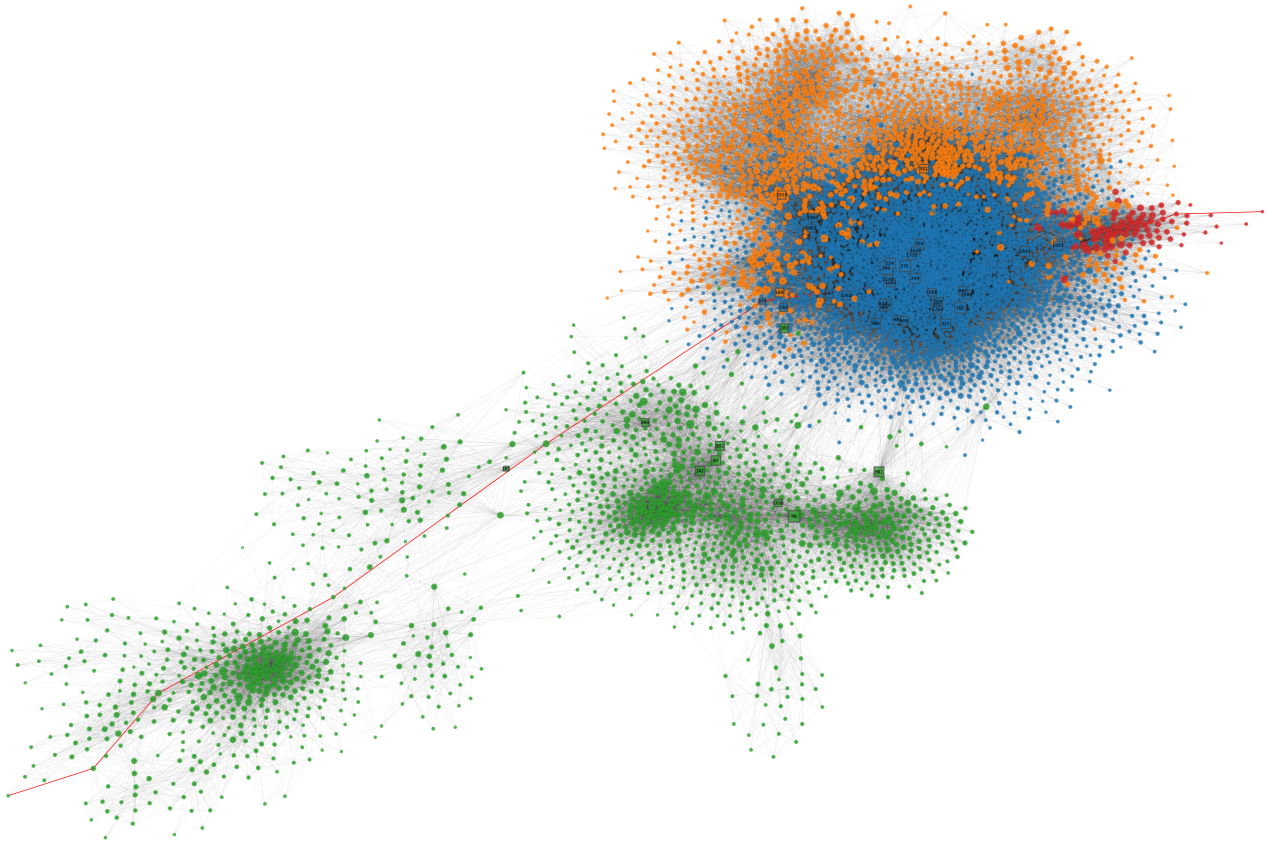


Figure 6: $L2$ company networks visualization. Network reconstructed from Facebook friendships. Nodes represent Facebook users who belong to the corporation, and links represent Facebook friendships. The color of the nodes represent their communities using greedy modularity optimization. Size of the nodes ar proportional to degree centrality. Also, the shape of the nodes that are ranked in the centrality measures have a square shape and are labeled. Edges in red are those corresponding to the diameter of the network.

A.3 Coding and Github Repository

All the computations were performed in Python, mainly using the library NetworkX [22]. For statistical analysis we used mainly SciPy library [23]. We implement some functions when needed. For a complete and detailed version of the code, including all analyses, it is available in the GitHub repository:

https://github.com/mateocarpio/Organization-Structures-Facebook.

References

[1] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th annual computer security applications conference*, pp. 93–102, 2011.

[2] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," in *Communities and Technologies: Proceedings of the First International Conference on Communities and Technologies; C&T 2003*, pp. 81–96, Springer, 2003.

[3] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the facebook," in *International workshop on privacy enhancing technologies*, pp. 36–58, Springer, 2006.

[4] M. Fire and R. Puzis, "Organization mining using online social networks," *Networks and Spatial Economics*, vol. 16, no. 2, pp. 545–578, 2016.

[5] Netzschleuder, "Netzschleuder: A collection of networks," 2013. Accessed: February 5, 2025.

[6] E. Estrada, *The structure of complex networks: theory and applications*. American Chemical Society, 2012.

[7] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology*, vol. 105, no. 2, pp. 493–527, 1999.

[8] E. Estrada and P. A. Knight, *A first course in network theory*. Oxford University Press, USA, 2015.

[9] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review EâStatistical, Nonlinear, and Soft Matter Physics*, vol. 71, no. 5, p. 056103, 2005.

[10] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[11] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review EâStatistical, Nonlinear, and Soft Matter Physics*, vol. 70, no. 6, p. 066111, 2004.

[12] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review EâStatistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 6, p. 066133, 2004.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[14] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[15] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the national academy of sciences*, vol. 97, no. 21, pp. 11149–11152, 2000.

[16] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[17] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Crawling facebook for social network analysis purposes," in *Proceedings of the international conference on web intelligence, mining and semantics*, pp. 1–8, 2011.

[18] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th annual ACM Web science conference*, pp. 33–42, 2012.

[19] F. Buccafurri, G. Lax, and A. Nocera, "A new form of assortativity in online social networks," *International Journal of Human-Computer Studies*, vol. 80, pp. 56–65, 2015.

[20] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? the structure of the twitter follow graph," in *Proceedings of the 23rd international conference on world wide web*, pp. 493–498, 2014.

[21] E. Ferrara, "A large-scale community structure analysis in facebook," *EPJ Data Science*, vol. 1, pp. 1–30, 2012.

[22] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11 – 15, 2008.

[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.