Network Analysis
GPlus social network

Samuele Crea

June 2024

1 The Graph

The graph taken into analysis is the one representing connections between users on the social network Google plus now fallen into disuse for a few years.
The file representing our graph is of type .edges and thus contains only the links between nodes.

Nodes contain no information and are represented by numeric ids.
There are no names of users in any way, only numbers representing them anonymously.

The networkX python library was used for any analysis within this report.

As we want to define some basic characteristics of the dataset for a preliminary analysis of the data, we can look at Table 1 represented below:

Nodes	Edges	Density	Avarage degree
23628	39194	0.00014	3.317

Table 1: quick overview of the graph

As we can see, this is a strongly sparse graph considering a really low density and an avarage degree of only 3.3.

The graph contains about 23600 nodes with just under twice as many edges.
The low density suggests that most nodes have few direct connections and that hubs probably exist in the graph.

Wanting to go deep into the avarage degree analysis of the graph, we can produce a graph like the one displayed in figure 1:

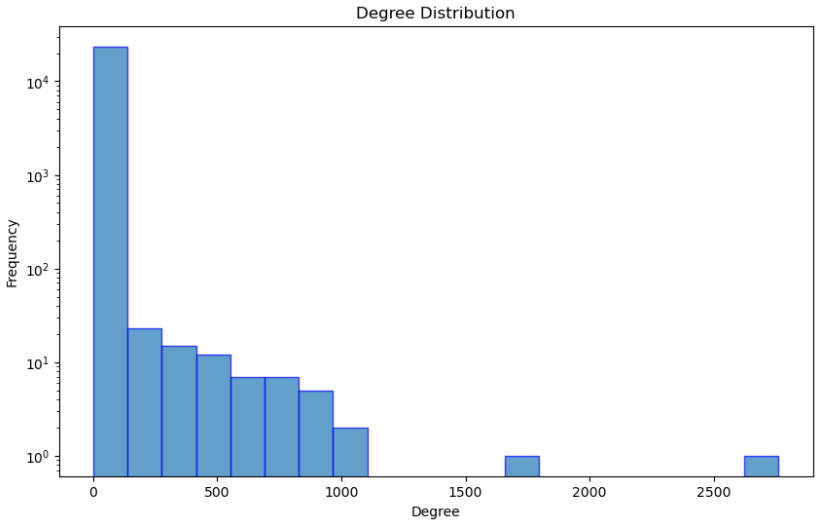

Figure 1: degree distribution

As we can see from the figure, the distribution of node degrees shows that most nodes have a relatively low degree, while a few nodes have a very high degree.
This is typical in social network graphs, where a few nodes (individuals) are highly connected while most have few connections.

Wanting to further analyze the nodes' degree distrubution, we can see what is shown in table 2:

Min Degree	Max Degree
1	2761

Table 2: quick overview of the degree distribution

As might be expected, there are nodes with degree equal to 1 (looking at the graph in Figure 1 we can see that these are the overwhelming majority in the graph) and there are nodes (though few) with really many connections up to 2761.

In particular as we can see from Table 3 the nodes with a really large number of connections are mainly two:

Node	Degree
2300	2761
8306	1703
2622	999
2376	986
11324	962
19205	937

Table 3: Degree of Nodes

The two nodes with a really high degree compared to the average are 2300 and 8306.
In all probability these two nodes will be of crucial importance within the graph.

Through the NetworkX library it was then possible to represent the graph in a simple way, despite its rather large size.

To do this, I used the NetworkX functions spring layout and draw networkx to obtain a clear and meaningful result.

The graph can be seen in Figure 2 below:

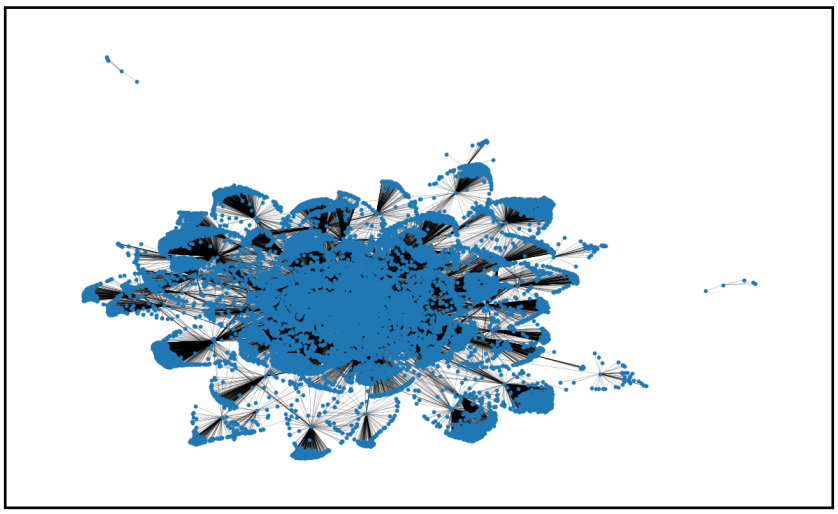

Figure 2: Graph representation

As we can see by eye from the representation of the graph, it is not connected. Checking with the appropriate function of networkX this assumption has been confirmed.
Finding the giant component, however, I was able to see that this deviates only slightly in number of nodes and arcs from the complete graph.

Nodes - Graph	Edges - Graph	Nodes - Giant comp	Edges - Giant comp
23628	39194	23613	39182

Table 4: nodes in the giant component

As we can see, the graph is not connected for only a few nodes (the ones on the right and top left in the image) which are only 15.

Once these assumptions about the graph have been made, we can move on to a more in-depth study of centrality.

2 Centrality

In this analysis we will focus on three main measures:

1. **Degree Centrality**:Is the total number of neighboring nodes at distance one.
2. **Betweenness Centrality**: is the number of geodesic paths passing through a node.
3. **Closeness Centrality**: is the harmonic mean of the distances from one node to the other nodes.

2.1 Degree Centrality

Table 5 shows the degree centrality values for some of the nodes in the analyzed graph (those with the highest degree centrality).

Analyzing the data presented, we can see that the nodes with the highest degree centrality have significantly higher values than the average, indicating that these nodes are connected to a very large number of other nodes.

Node	Degree Centrality
2300	0.116858
8306	0.072079
2622	0.042282
2376	0.041732
11324	0.040716
19205	0.039658
1876	0.039362
15599	0.039023
8892	0.035426
7101	0.033225

Table 5: Degree Centrality of Nodes

From the table, we see that node 2300 has the highest degree centrality with a value of 0.116858.

This confirms that node 2300 has a much higher number of direct connections than the graph average, making it a highly influential node in the network.

In fact, as defined above, node 2300 has as many as 2761 connections with other nodes.

It is followed by node 8306 with the value of 0.072079, still significantly high but significantly lower than node 2300.

These degree centrality values indicate that, despite the low overall density of the graph (as discussed in the previous section), there are some nodes that act as main nodes(hub), facilitating connectivity within the network.

This is typical in social network graphs.
The presence of such nodes is important for the robustness and efficiency of the network.

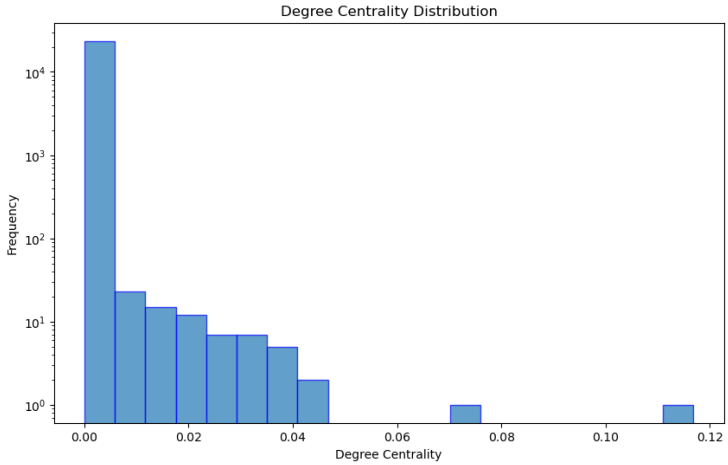

Figure 3: distribution of degree

The graph in Figure 3 represents that just discussed when talking about degree centrality.
As we can see, from the graph we notice the two rightmost histogram bars that correspond to nodes 2300 and 8306.

2.2 Betweenness Centrality

Table 6 shows, as in the previous case, the values of betweenness centrality for some of the nodes in the analyzed graph.

A high value of betweenness centrality indicates that a node serves as a critical bridge for the flow of information between different parts of the graph.

Node	Betweenness Centrality
2300	0.221928
1876	0.160931
8306	0.112780
15599	0.067009
2622	0.059116
18440	0.054284
19205	0.054052
4965	0.048714
7101	0.045732
5958	0.042765

Table 6: Betweenness Centrality of Nodes

From the table, we see that node 2300 has the highest value of betweenness centrality with 0.221928.
This suggests that node 2300 is a key crossing point for many paths in the graph, making it critical for network connectivity.
Node 1876 follows with a value of 0.160931, also indicative of a significant role in facilitating information flow.

High values of betweenness centrality indicate that such nodes are essential for maintaining the integrity of the graph, as their removal could fragment the network and increase the distances between the remaining nodes.
For example, in a social network, these nodes may represent individuals connecting different communities, facilitating interaction between otherwise disconnected groups.

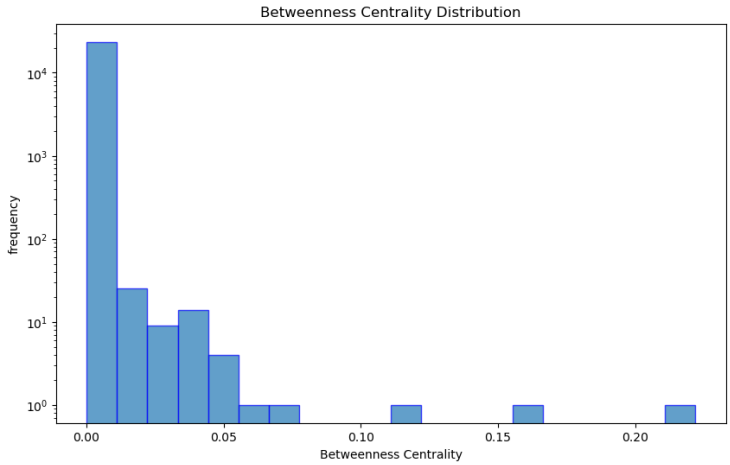

Figure 4: distribution of degree

As we can see from the graph in Figure 4 there are few nodes with high betweenness centrality, and that is why they are really important for the graph, which without them would be extremely fragmented.

2.3 Closeness Centrality

Table 7 shows the values of closeness centrality for the graph nodes.
A high value of closeness centrality indicates that a node is, on average, very close to all other nodes in the graph, suggesting a central location and high accessibility within the network.

Node	Closeness Centrality
2300	0.408507
1876	0.381952
2622	0.371595
6306	0.366169
629	0.358797
3234	0.358443
3331	0.357363
3414	0.355891
1074	0.352689
2255	0.352657

Table 7: Closeness Centrality of Nodes

From the table, we see that node 2300 has the highest value of closeness centrality with 0.408507.
This indicates that node 2300 is on average closer to all other nodes than any other node in the graph, confirming its position as a central node in the network. This is followed by nodes 1876 and 2622 with values of 0.381952 and 0.371595 respectively, which, although lower than those of node 2300, still indicate significant centrality within the graph.

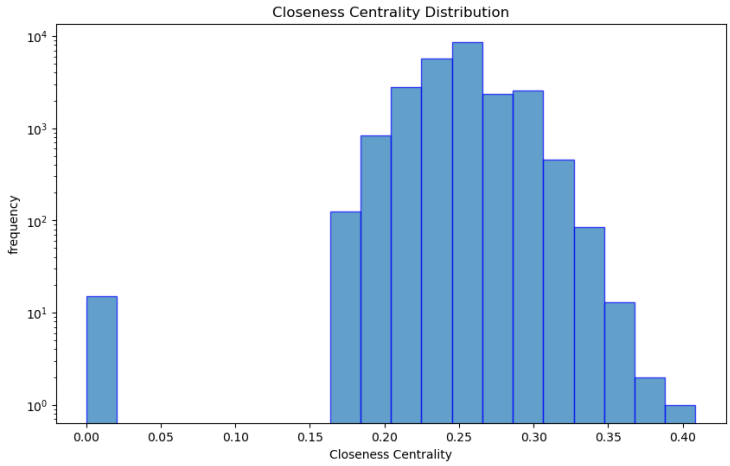

Figure 5: closeness centrality

2.4 Most important nodes

As has become clear by now from the analysis of the graph there is one main node, definitely more important than all the other nodes, which is 2300.
This node was found to be the one with the largest number of connections and at the same time the one with the highest centrality values.

metrics	node 2300
Degree	2761
Degree centrality	0.1168
Betweenness centrality	0.2219
Closeness centrality	0.4085

Table 8: Metrics of node 2300

This node is both the center of the graph and at the same time a very important bridge for communication between the other nodes.
It is clear that if this node were eliminated there would be great fragmentation within the graph.

Other important nodes are 1876 which has high Closeness centrality and Betweenness centrality and node 8306 which has high Degree centrality.

3 Assortativity

Assortativity is a measure that indicates the tendency of nodes in a graph to connect with other nodes having a similar degree.
A positive value of assortativity suggests that nodes tend to connect with other nodes with a similar degree (positive assortativity), while a negative value indicates that nodes tend to connect with nodes with very different degrees (disassortativity).

For the GPlus graph analyzed, the degree assortativity coefficient is -0.3885157446925048.
This negative value indicates strong disassortativity in the graph.
In other words, nodes with a high degree tend to connect predominantly with nodes with a low degree, rather than with other high-degree nodes.

This type of structure is common in social networks, where some individuals have a very large number of connections (e.g., influencers), while most users have few connections.

This structure may suggest greater dependence on central nodes, making the network vulnerable to targeted removal of these key nodes.
As we can recall from the Centrality analysis, these essential nodes that connect with nodes with much smaller degree do exist (a prime example is 2300) and their removal could prove to be a big problem for the graph.

4 Clustering

The clustering coefficient of a graph measures the degree to which nodes in a graph tend to join together in clusters.

This is calculated by taking the average of the clustering coefficients of all nodes. The clustering coefficient of a node is the ratio of the number of triangles that pass through the node to the maximum number of triangles that could pass through the node.

In our case, this is 0.17412604214483396.
This value indicates that there is a moderate tendency for nodes in the graph to form clusters.

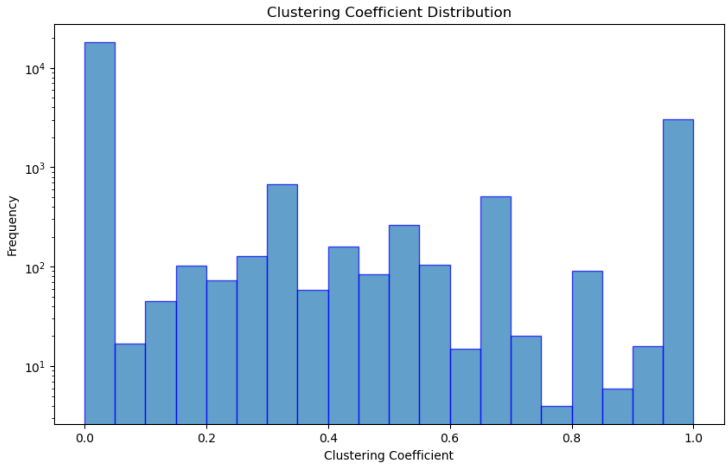

Figure 6: distribution of clustering

As can be seen from the figure, the clustering coefficient is distributed rather uniformly with a peak on the extremes.

The uniform distribution between these two extremes indicates structural diversity within the network.
Some nodes belong to subgroups with high connection density (clusters), while others are more isolated.

This combination of dense clusters and peripheral nodes is typical of social networks, where strongly interconnected communities exist interspersed with nodes with few connections.

The transitivity of the graph, on the other hand, is 0.0037087945918890955.

This very low value indicates that, globally, the probability that two neighbors of a node are also neighbors of each other is extremely low.

This suggests that, although local clusters may exist within the graph, the overall structure of the graph is not conducive to the formation of triangles on a global scale.

5 Communities

The analysis of community structure in the Google Plus graph revealed the presence of 34 distinct communities.
Communities are subsets of nodes that are more densely connected to each other than to the rest of the graph.
To find the communities within the graph, I used the networkX function greedy modularity communities.

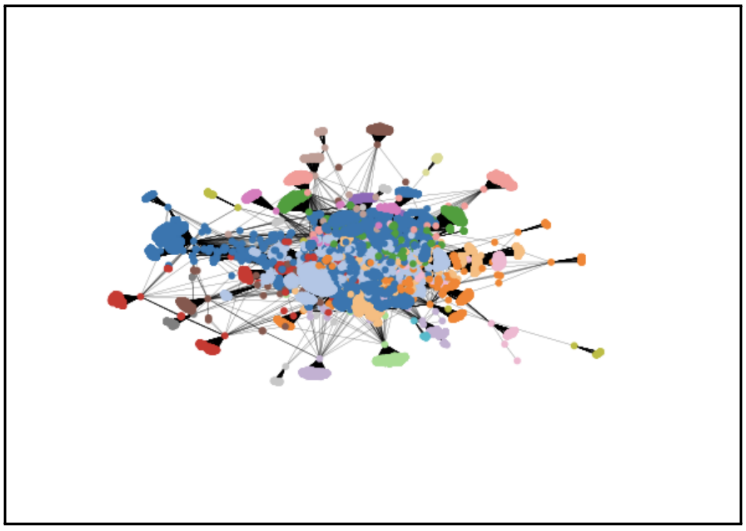

Figure 7: A better visualization of the graph

In Figure 7, a visualization of the structure of communities within the graph can be observed.
Each color represents a distinct community, highlighting how the nodes within them are closely interconnected.

6 Visualization with Graphia

A representation of the graph through the use of the Graphia tool, which allows graphs to be drawn quickly and more representatively than networkX and its libraries, is shown below.

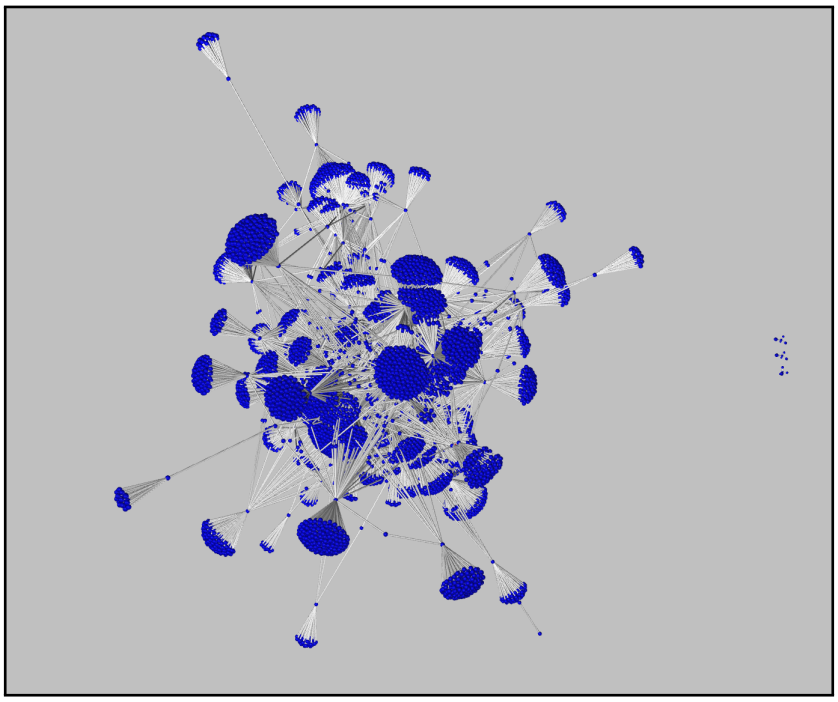

Figure 8: A better visualization of the graph