

# Network Analysis

## GPlus social network

Samuele Crea

June 2024

### 1 The Graph

il grafo preso in analisi è quello che rappresenta le connessioni tra utenti sul social network google plus ormai caduto in disuso da qualche anno.  
il file che rappresenta il nostro grafo è di tipo .edges e contiene quindi solo i link fra nodi.

I nodi non contengono informazioni e sono rappresentati da id numerici.  
Non sono presenti in alcun modo i nomi degli utenti, ma solo numeri che li rappresentano in modo anonimo.

Per qualsiasi analisi all'interno di questa relazione è stata usata la libreria per python di networkX.

Volendo definire alcune caratteristiche base del dataset per un'analisi preliminare dei dati possiamo guardare la Table 1 rappresentata qui sotto:

Nodes	Edges	Density	Avarage degree
23628	39194	0.00014	3.317

Table 1: quick overview of the graph

Come possiamo vedere, si tratta di un grafo fortemente sparso considerando una densità veramente bassa e un avarage degree di solo 3.3.

Il grafo contiene circa 23600 nodi con poco meno del doppio di edges.  
La bassa densità suggerisce che la maggior parte dei nodi ha poche connessioni dirette e che esistono probabilmente alcune componenti connesse significative all'interno del grafo.

Volendo andare a fondo nell'analisi dell'avarage degree del grafo possiamo produrre un grafo come quello visualizzato in figure 1:

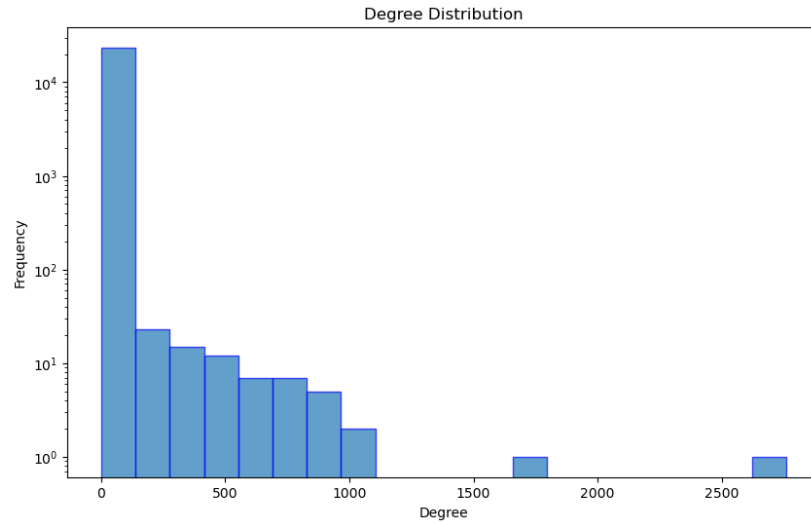


Figure 1: degree distribution

Come possiamo constatare dalla figura, la distribuzione dei gradi dei nodi mostra che la maggior parte dei nodi ha un grado relativamente basso, mentre pochi nodi hanno un grado molto alto. Questo è tipico nei grafi di social network, dove pochi nodi (individui) sono altamente connessi mentre la maggior parte ha poche connessioni.

Volendo approfondire l'analisi sulla distribuzione del grado dei nodi possiamo vedere quanto riportato in table 2:

Min Degree	Max Degree
1	2761

Table 2: quick overview of the degree distribution

Come ci si poteva aspettare, esistono nodi con grado uguale a 1 (guardando il grafico in Figure 1 si può notare come questi siano la stragrande maggioranza nel grafo) ed esistono nodi (anche se pochi) con davvero molte connessioni fino a 2761.

In particolare come possiamo vedere dalla Table 3 i nodi con un numero davvero elevato di connessioni sono principalmente due:

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

Table 3: Degree of Nodes

I due nodi con un grado davvero alto rispetto alla media sono 2300 e 8306. Con molta probabilità questi due nodi avranno un'importanza cruciale all'interno del grafo.

Attraverso la libreria NetworkX è stato poi possibile rappresentare in modo semplice il grafo, nonostante le sue dimensioni piuttosto elevate.

Per farlo ho utilizzato le funzioni di networkX `spring layout` e `draw_networkx` per ottenere un risultato chiaro e significativo.

Il grafo si può vedere nella figure 2 sottostante:

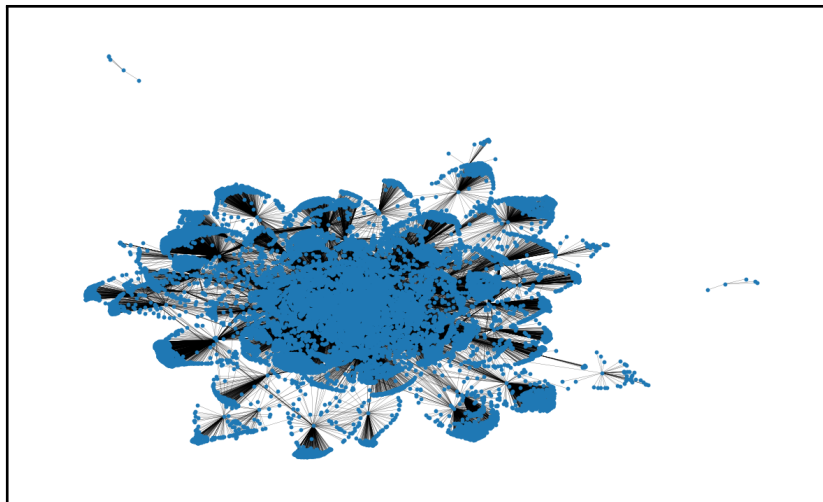


Figure 2: Distribution of degree

Come possiamo notare ad occhio nudo dalla rappresentazione del grafo, questo non è connesso. Controllando con l'apposita funzione di networkX questa supposizione è stata confermata.

Trovando però il giant component possiamo constatare che questo si discosta di poco per numero di nodi e archi rispetto al grafo completo.

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

Table 4: nodes in the giant component

Come possiamo vedere il grafo non è connesso per solo pochi nodi (quelli a destra e in alto a sinistra nell'immagine) e per questo grafo e giant component differiscono solo di 15 nodi.

Fatte queste premesse sul grafo possiamo passare ad uno studio più approfondito della centralità.

## 2 Centrality

In questa analisi ci concentreremo su tre misure principali:

1. **Degree Centrality**: è il numero totale di nodi vicini a distanza uno.
2. **Betweenness Centrality**: misura globale, è il numero di percorsi geodetici che passano attraverso un nodo.
3. **Closeness Centrality**: misura globale, è la media armonica delle distanze da un nodo agli altri nodi.

### 2.1 Degree Centrality

La Table 5 mostra i valori della degree centrality per alcuni dei nodi del grafo analizzato (quelli che la hanno più alta).

Analizzando i dati presentati, possiamo osservare che i nodi con la centralità più alta hanno valori significativamente superiori rispetto alla media, indicando che questi nodi sono connessi a un numero molto elevato di altri nodi.

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

Dalla tabella, si nota che il nodo 2300 ha la degree centrality più alta con un valore di 0.116858.

Questo ci conferma che il nodo 2300 ha un numero di connessioni dirette molto superiore alla media del grafo, rendendolo un nodo altamente influente nella rete.

Infatti, come già definito sopra, il nodo 2300 ha ben 2761 connessioni con altri nodi

Segue il nodo 8306 con il valore di 0.072079, ancora significativamente elevato ma notevolmente inferiore rispetto al nodo 2300.

Questi valori di degree centrality indicano che, nonostante la bassa densità complessiva del grafo (come discusso nella sezione precedente), esistono alcuni nodi che agiscono come nodi principali, facilitando la connettività all'interno della rete.

Questo è tipico nei grafi di social network.

La presenza di tali nodi è importante per la robustezza e l'efficienza della rete.

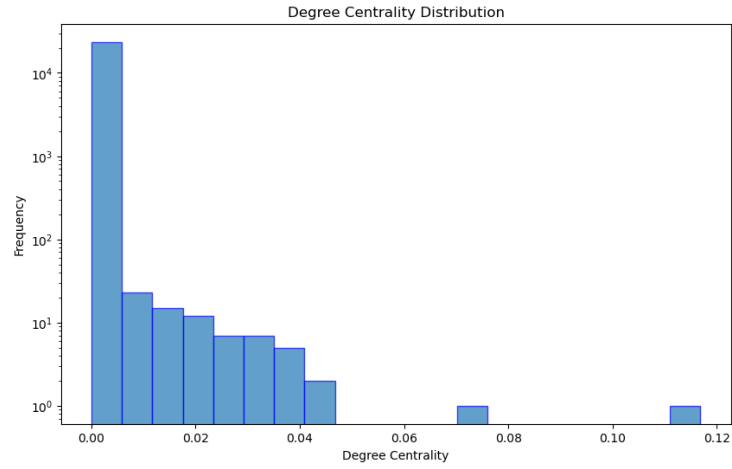


Figure 3: distribution of degree

Il grafico in Figure 3 rappresenta quello appena discusso parlando di degree centrality.

Come possiamo vedere, dal grafico si notano le due barre dell'istogramma più a destra che corrispondono ai nodi 2300 e 8306.

## 2.2 Betweenness Centrality

La tabella 6 riporta, come nel caso precedente, i valori della betweenness centrality per alcuni dei nodi del grafo analizzato.

Un valore elevato di betweenness centrality indica che un nodo funge da ponte critico per il flusso di informazioni tra diverse parti del grafo.

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

Dalla tabella, si nota che il nodo 2300 ha il valore più alto di betweenness centrality con 0.221928.

Questo suggerisce che il nodo 2300 è un punto di passaggio chiave per molti percorsi nel grafo, rendendolo fondamentale per la connettività della rete. Il nodo 1876 segue con un valore di 0.160931, anch'esso indicativo di un ruolo significativo nella facilitazione del flusso di informazioni.

Valori elevati di betweenness centrality indicano che tali nodi sono essenziali per mantenere l'integrità del grafo, in quanto la loro rimozione potrebbe frammentare la rete e aumentare le distanze tra i rimanenti nodi. Ad esempio, in un social network, questi nodi possono rappresentare individui che collegano diverse comunità, facilitando l'interazione tra gruppi altrimenti disconnessi.

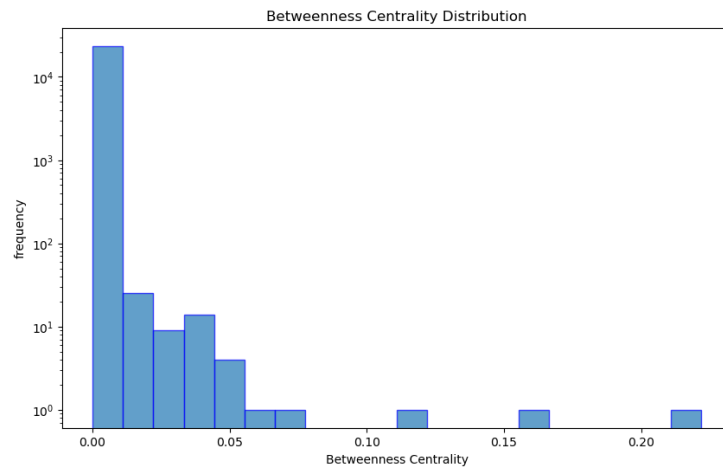


Figure 4: distribution of degree

Come possiamo notare dal grafico in Figure 4 ci sono pochi nodi con una betweenness centrality elevata, e per questo sono davvero importanti per il grafo, che senza di loro risulterebbe estremamente frammentato.

## 2.3 Closeness Centrality

La tabella 7 invece riporta i valori della closeness centrality per i nodi del grafo. Un valore alto di closeness centrality indica che un nodo è, in media, molto vicino a tutti gli altri nodi del grafo, il che suggerisce una posizione centrale ed elevata accessibilità all'interno della rete.

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

Dalla tabella, si nota che il nodo 2300 ha il valore più alto di closeness centrality con 0.408507. Questo indica che il nodo 2300 è mediamente più vicino a tutti gli altri nodi rispetto a qualsiasi altro nodo del grafo, confermando la sua posizione di nodo centrale nella rete. Seguono i nodi 1876 e 2622 con valori di 0.381952 e 0.371595 rispettivamente, che, sebbene inferiori a quelli del nodo 2300, indicano ancora una significativa centralità all'interno del grafo.

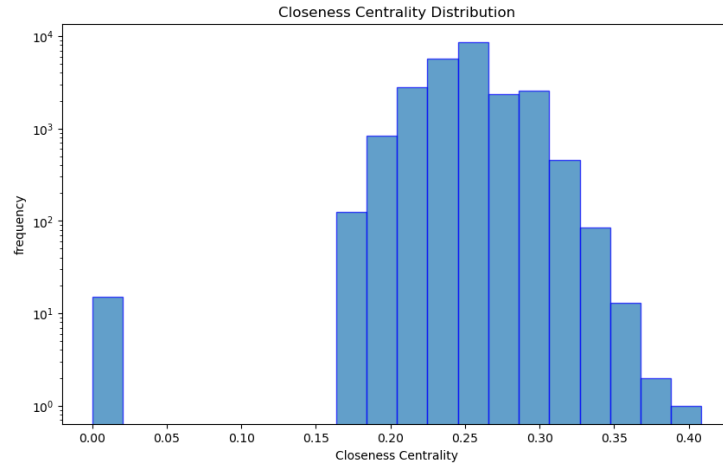


Figure 5: closeness centrality

## 2.4 Most important nodes

Come ormai è risultato chiaro dall'analisi del grafo esiste un nodo principale, decisamente più importante di tutti gli altri nodi, ovvero 2300.

Il nodo in questione è risultato quello col maggior numero di connessioni e allo stesso tempo quello con i valori di centralità più alti.



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

Table 8: Metrics of node 2300

Questo nodo è sia il centro del grafo e allo stesso tempo un bridge molto importante per la comunicazione tra gli altri nodi.

E' chiaro che se questo nodo venisse eliminato ci sarebbero grandi frammentazioni all'interno del grafo.

Altri nodi importanti sono 1876 che ha alta Closeness centrality e Betweenness centrality e il nodo 8306 che ha alta Degree centrality.

### 3 Assortativity

L'assortatività è una misura che indica la tendenza dei nodi di un grafo a connettersi con altri nodi aventi un grado simile.

Un valore positivo di assortativity suggerisce che i nodi tendono a connettersi con altri nodi con un grado simile (assortatività positiva), mentre un valore negativo indica che i nodi tendono a connettersi con nodi con gradi molto diversi (disassortatività).

Per il grafo di GPlus analizzato, il coefficiente di assortatività di grado è -0.3885157446925048.

Questo valore negativo indica una forte disassortatività nel grafo.

In altre parole, i nodi con un alto grado tendono a connettersi prevalentemente con nodi con un basso grado, piuttosto che con altri nodi ad alto grado.

Questo tipo di struttura è comune nei social network, dove alcuni individui hanno un numero molto elevato di connessioni (ad esempio, influencer), mentre la maggior parte degli utenti ha poche connessioni.

Questa struttura può suggerire una dipendenza maggiore dai nodi centrali, rendendo la rete vulnerabile alla rimozione mirata di questi nodi chiave.

Come possiamo ricordare dall'analisi della Centrality, questi nodi essenziali che si collegano con nodi con degree molto più piccolo esistono eccome (un esempio lampante è 2300) e la loro rimozione potrebbe rivelarsi un grosso problema per il grafo.

## 4 Clustering

Il coefficiente di clustering di un grafo misura il grado con cui i nodi di un grafo tendono a unirsi in cluster.

Questo viene calcolato prendendo la media dei coefficienti di clustering di tutti i nodi.

Il coefficiente di clustering di un nodo è il rapporto tra il numero di triangoli che passano attraverso il nodo e il numero massimo di triangoli che potrebbero passare attraverso il nodo.

Nel nostro caso, questo è 0.17412604214483396.

Questo valore indica che c'è una tendenza moderata per i nodi del grafo a formare cluster.

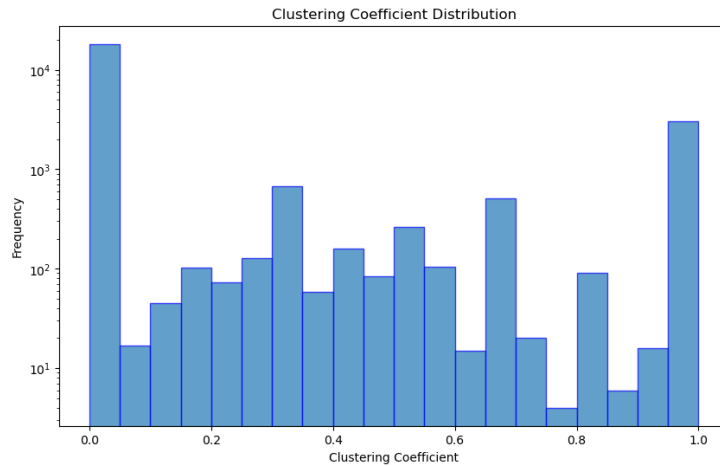


Figure 6: distribution of clustering

Come si nota dalla figura il coefficiente di clustering è distribuito piuttosto uniformemente con un picco sugli estremi.

La distribuzione uniforme tra questi due estremi indica una diversità strutturale all'interno della rete.

Alcuni nodi appartengono a sottogruppi con alta densità di connessioni (cluster), mentre altri sono più isolati.

Questa combinazione di cluster densi e nodi periferici è tipica dei social network,

dove esistono comunità fortemente interconnesse intervallate da nodi con poche connessioni.

La transitivity del grafo, invece, è 0.0037087945918890955.

Questo valore molto basso indica che, a livello globale, la probabilità che due vicini di un nodo siano anche vicini tra loro è estremamente bassa.

Ciò suggerisce che, sebbene possano esistere cluster locali all'interno del grafo, la struttura complessiva del grafo non favorisce la formazione di triangoli su scala globale.

## 5 Communities

L'analisi della struttura delle comunità nel grafo di Google Plus ha rivelato la presenza di 34 comunità distinte.

Le comunità sono sottoinsiemi di nodi che sono più densamente connessi tra loro rispetto al resto del grafo.

Per trovare le community all'interno del grafo ho utilizzato la funzione di networkX `greedy modularity communities`.

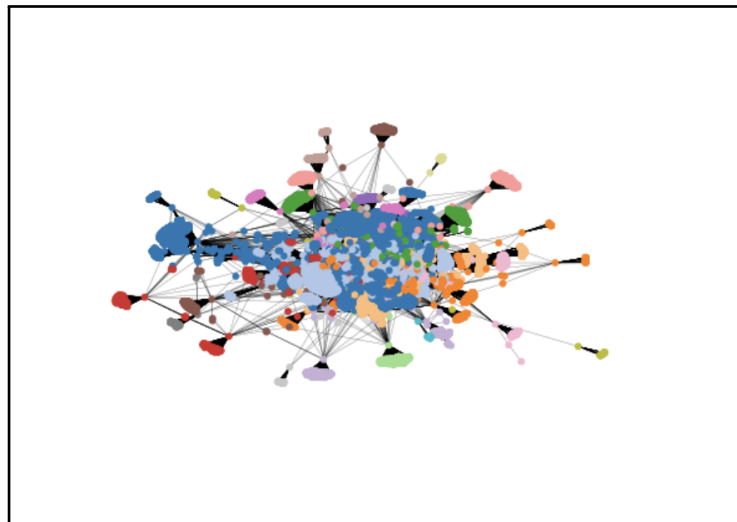


Figure 7: A better visualization of the graph

In Figure 7, è possibile osservare una visualizzazione della struttura delle comunità all'interno del grafo.

Ogni colore rappresenta una comunità distinta, evidenziando come i nodi all'interno delle stesse siano strettamente interconnessi.

## 6 Visualization with Graphia

Qui sotto è riportata una rappresentazione del grafo attraverso l'utilizzo del tool Graphia, che permette di disegnare grafi in modo rapido e più rappresentativo rispetto a networkX e le sue librerie.

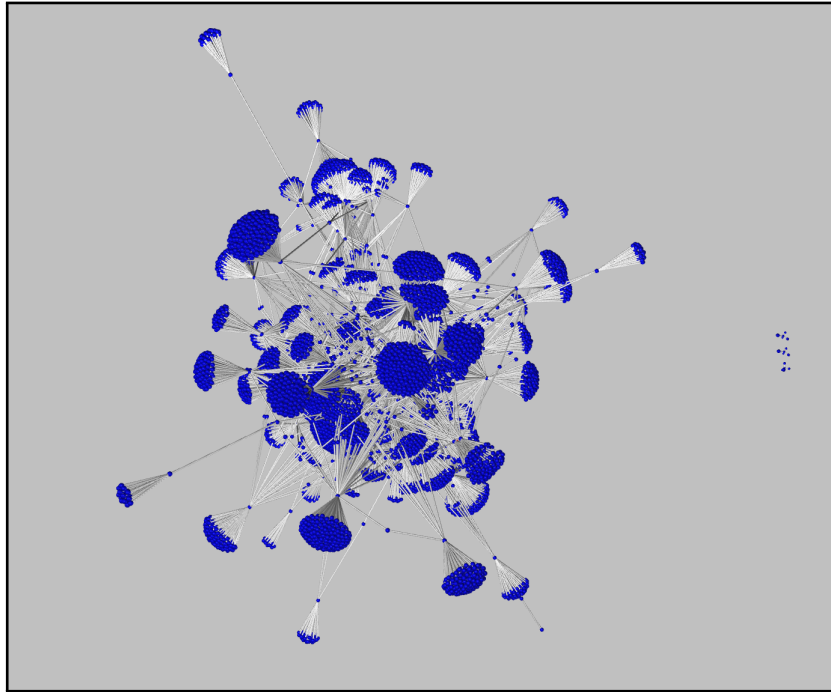


Figure 8: A better visualization of the graph