LINK ANALYSIS STUDY

Of Facebook 'Ego' Network and Twitter Interactions Network

Pol Monné Parera, Pau Casé Barrera
polmonne7@gmail.com, paucase6@gmail.com

Table of Contents

Data Information

Source

For the development of the project two different datasets have been studied and compared.

The first one, and the main source studied, consists of a social 'ego' network from Facebook and the graph is undirected. The data was extracted from the Stanford Network Analysis Project (Leskovec, Social Circles: Facebook, s.f.).

The second one consists of the Twitter interactions network for the 117th United States Congress, both house of representatives and senate and is a directed graph. The data was also extracted from the Stanford Network Analysis Project (Leskovec, Twitter Interaction Network for the US Congress, 2023)

Description

The first dataset consists of an ego-network based on the 'friends lists' from Facebook. An ego-network being a specific kind of social network centered around a specific individual/'ego' (in this dataset 10 individuals), their direct connections ('alters') and any direct connections among these.

Data was collected from survey participants using the Facebook app and includes node features (profiles), circles, and ego networks.

It is important to consider that data had been anonymized by replacing the Facebook-internal ids for each user with a new value. Also, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured for data privacy reasons.

The 'ego' nodes studied are: 0, 107, 348, 414, 686, 698, 1684, 1912, 3437 and 3980.

The second dataset studied consists of the twitter interaction network for the 117th United States Congress, both house of representatives and congress. It is stated that the data was extracted using Twitter's API, and that later the empirical transmission probabilities were quantified according to a variety of metrics.

The focus of the research will be to analyze and extract information from the Facebook dataset, while the US Congress one will be a helpful tool of comparison and academic research on many parts of the assignment.

In both cases, the nodes are the social network users. In the Facebook dataset, edges are mutual connections (friends), while in the Twitter dataset, a directed and weighted approach is taken, where edges are heavier if more interactions happened from one node to another.

Data's Original Usage

The first dataset was used in a variety of papers for research and academical purposes. One of them, the previously cited paper from the author on learning to discover social circles in ego networks (Leskovec & McAuley, Learning to Discover Social Circles in Ego Networks, 2012). Nevertheless, it is important to note that the dataset used in the paper did not contain the data in the same format as provided for the general usage (e.g. interpretation of features was not obscured in the paper's dataset).

The second dataset has also had its original usage in academic and research environments. An example of this comes from the paper "A centrality measure for quantifying spread on weighted, directed networks" (Fink, et al., 2023).

Together, these datasets allow for deep research on different types of graphs (e.g. directed versus undirected), with a vast variety of academical possible usages comparing and exploring its data.

Size

The first dataset (Facebook social groups) consists of a total of 4,039 nodes, from those only ten are 'ego' nodes, and a total of 88,234 edges.

The second dataset (Twitter US Congress relationships) is a much smaller graph, consisting of only 475 nodes and 13,289 edges.

Global Properties

The following are the global properties of the Facebook social 'ego' network:

Property	Value	Interpretation/Discussion
Network diameter	8	Out of the complete network of Facebook users in the dataset the longest path from one to another is 8. This means that some of our ego networks are connected, because each ego network should have a diameter of 2.
Graph density	0.010819	The density of the graph tells us that there are many vertices compared to the edges in this undirected graph, interconnection is not very high, although we would need to compare to other similar datasets for comparison.
Average shortest path length	3.667	The average shortest path length between nodes is 3.667. This supports that the ego networks are connected between them, because otherwise the average shortest path length should not exceed 2. After seeing the diameter and average shortest path length values, this was practically a certainty. This was further studied by analyzing different edges and a connection was found between nodes 0 and 107, therefore connecting their networks. Many other connections were found both within alters and within egos.
Clustering coefficient	0.6055	This average clustering coefficient tells us that the network is well interconnected, with many edges forming triangles or other sizes of clusters.

The following are the global properties of the Twitter US Congress interactions network:

Property	Value	Interpretation/Discussion
Network diameter	4	Out of the complete network of Twitter congress users in the dataset the longest path from one to another is 4. This small value reflects perfectly the "small world" property typical of online social graphs.
Graph density	0.090802	The relatively sparse density still indicates a fairly cohesive network; many users share mutual connections but there are still distinct clusters (e.g. political parties). This value gives even more realism to the previous value of network diameter. It is not surprising that if on average each node is connected to approximately 9 other nodes, it only takes 4 connections at the most to connect to different nodes (take into consideration that 9^4 equals 6561). Compared to the Facebook dataset, this value is slightly lower, but the interconnection seems to be similar.
Average shortest path length	2.0639	The average shortest path length between nodes is 2.0639, this represents that on average any member can reach any other through just one intermediary. Once again this reinforces the "small world" nature. Even with a network with a relatively small graph density communication paths are very short. On a deeper level and in social terms, this finding implies high reachability and a potential for fast information flow (also across political divides).
Clustering coefficient	0.3014	This value indicates moderate clustering of the network, while it is globally well connected, users still form "local communities". This is likely to reflect party lines or regional affiliations.

Important Nodes

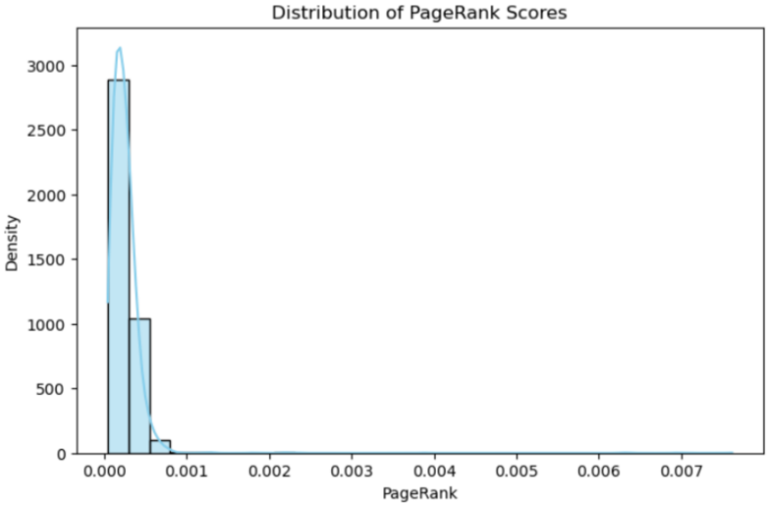
The following analysis derives from the Facebook dataset:

Top 5 Nodes with Highest PageRank

Node label or ID	PageRank	Betweenness	Hubness	Authority
3437	0.007615	0.18754	4.90E-09	4.90E-09
107	0.006936	0.497257	1.07E-05	1.07E-05
1684	0.006367	0.345927	3.93E-07	3.93E-07
0	0.00629	0.157352	2.12E-06	2.12E-06
1912	0.003877	0.235988	6.12E-03	6.12E-03

The Top 5 PageRank classification displays that there clearly are 4 nodes that stand out. The difference on the value of PageRank between node 1912 and node 0 is very large.

We can further explore this difference by plotting the distribution of the PageRank value on the whole network.



This plot shows that the social network analyzed has many non-important nodes, with PageRank scores between 0.000 and 0.001 including 99.7% of the nodes. Therefore, while the table shows there is a clear Top 4, we see from the distribution that any page in this top 0.3% has a very large value of links from other important nodes compared to most other nodes.

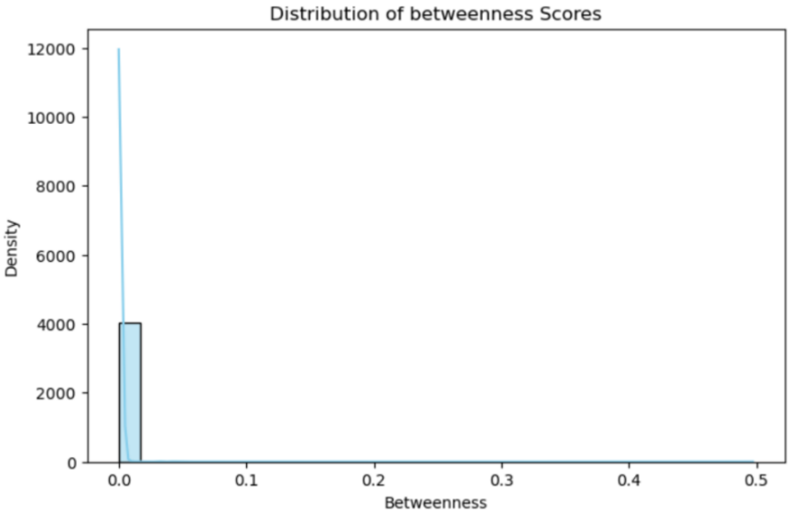
This dataset is a study of 10 specific people (egos) whose diverse anonymous data is also stored in separate files. The top 5 nodes shown in the list are all part of these 10 ego nodes studied in the dataset, which is coherent.

If we increased the size of the dataset by connecting all the friends of the friends of these 10 people, the dataset would grow enormously. With this change the PageRank scores of the friends of the 10 people would probably increase dramatically, but the new friends of friends would get a very low score again, probably keeping the same distribution.

Top 5 Nodes with Highest Betweenness

Node label or ID	PageRank	Betweenness	Hubness	Authority
107	0.006936	0.497257	1.072491e-05	1.072491e-05
1684	0.006367	0.345927	3.926877e-07	3.926877e-07
1912	0.003877	0.235988	6.117290e-03	6.117290e-03
3437	0.007615	0.187540	4.902888e-09	4.902888e-09
0	0.006290	0.157352	2.124473e-06	2.124473e-06

Analyzing the betweenness values of the top 5 nodes, we run into the same nodes again, although in a different order. This is again understandable when considering the type of dataset studied. This indicates that these nodes connect different communities together more than any of their friends, if we only consider their friends in the network.



The betweenness graph seems even less distributed, with mostly all nodes having practically a value of 0. Considering most nodes belong to closed networks, this could be an expected result.

Top 5 Nodes with Highest Hubness

Node label or ID	PageRank	Betweenness	Hubness	Authority
1912	0.003877	0.235988	6.117290e-03	6.117290e-03
2266	0.000455	0.001764	5.577249e-03	5.577249e-03
2206	0.000380	0.000005	5.517567e-03	5.517567e-03
2233	0.000421	0.000039	5.461203e-03	5.461203e-03
2464	0.000364	0.000004	5.403845e-03	5.403845e-03

The hubness scores give us a different view of the dataset. While the Top 1 hub is 1912, one of the ego nodes, the following four are alter nodes.

These nodes are all found in the ego network of node 1912. This points to this network being denser than the others, meaning that the friends of node 1912 are friends with a lot of friends of node 1912, with the four nodes displayed having the most connections within this network. Their low betweenness values support this theory, since they are simply part of a very connected network, they do not connect two communities but instead are part of one.

Given this surprise, it was analyzed how many connections these nodes had among them, as well as the average links that their neighbors had.

Node	Num_Neighbors	Avg_Links_of_Neighbors
1912 (ego node, top 1 hubness)	755	140.367881
2266 (top 2 hubness)	234	164.132968
2206 (top 3 hubness)	210	165.770947
2233 (top 4 hubness)	222	164.123606
2464 (top 5 hubness)	202	165.902276
2003 (random node in 1912 ego network)	48	106.008362
3980 (ego node)	59	18.830986
3437 (ego node)	547	58.644053

As demonstrated by the hub score, these mentioned nodes have neighbors with more neighbors than the ego nodes (except 1912). From this we know that 1912's ego network has a lot of interconnections within it. To support this is, a random node in the 1912 network was selected, but a broader analysis was also done in the following table.

Ego	Num_Alters	Avg_Links_of_Alters
0	347	54.55966
107	1045	111.613749
348	229	61.909651
414	159	57.000602
686	170	40.359175
3437	547	58.644053
3980	59	18.830986
1912	755	140.367881
1684	792	79.129933

This table allows us to visualize all ego networks with very simple metrics. It supports the fact that the 1912 ego network has far more interconnections than the other ego networks. Then, the 107-ego network is also very large and has a high number of interconnections within it too, but far from the ego network of node 1912.

Since the Facebook graph studied lacks directionality in the edges, the hubness and authority scores are identical.

Top 5 Nodes with Highest Authority

Node label or ID	PageRank	Betweenness	Hubness	Authority
1912	0.003877	0.235988	6.12E-03	6.12E-03
107	0.006936	0.497257	1.07E-05	1.07E-05
1684	0.006367	0.345927	3.93E-07	3.93E-07
0	0.006290	0.157352	2.124473E-06	2.124473E-06
3437	0.007615	0.187540	4.902888E-09	4.902888E-09

Community Detection

Method used: Clauset–Newman–Moore (CNM), Greedy approach

This method takes all nodes into one community and starts joining them, like the Agglomerative method for clustering. It uses the increase of modularity of each community as a metric to merge pairs of communities. The metric measures how dense the community would stay within and how far other communities would be.

Parameters: resolution=1, cutoff=1

Community #	Nodes included	Label/Description
0	#1684 + 982 nodes	Community around ego node 1684
1	#107, #348, #414 + 812 nodes	Community around ego node 107
2	#3437 + 547 nodes	Community around ego node 3437
3	#1912 + 542 nodes	Community around ego node 1912
4	#0 + 371 nodes	Community around ego node 0
5	219 nodes	Community around ego node 1684 not included
6	208 nodes	Community around ego node 1912 not included
7	#686, #698 + 204 nodes	Community around ego nodes 686 and 698
8	#3980 + 58 nodes	Community around ego node 3980
9	37	Community around ego node 107 not included
10	25	Community around ego node 414 not included
11	18	Community around ego node 1684 not included
12	6	Community around ego node 1684 not included

Communities matched to Ego networks

Community_ID	Num_Nodes	%net0	%net 107	%net 348	%net 414	%net 686	%net 698	%net 3437	%net 3980	%net 1912	%net 1684
0	983	0	46.49	0	0	0	0	0	0	0	54.63
1	815	0.12	66.01	26.26	16.07	0	0	0.12	0.12	0.37	0.12
2	548	0	0.18	0	0	0	0.91	99.45	0	0	0
3	543	0	0.55	0	0	0	0	0	0	99.82	0
4	372	93.01	2.42	4.03	0.81	0	0	0	0	0.54	2.69
5	219	0	0	0	0	0	0	0	0	0	100
6	208	0	0	0	0	0	0	0	0	100	0
7	206	0	0	0	0	82.52	30.58	0.49	0	0	0.49
8	59	0	0	0	0	0	0	0	98.31	0	0
9	37	0	100	0	0	0	0	0	0	0	0
10	25	0	0	0	100	0	0	0	0	0	0
11	18	0	0	0	0	0	0	0	0	0	100
12	6	0	0	0	0	0	0	0	0	0	100

From these two tables, we can see some interesting results from the community-finding method. Most ego nodes are found in this greedy approach, although some seem more split up than others.

Ego nodes 107, 384 and 414 are determined to be in the same community, and the same happens for 686 and 698. The other nodes are in a community with no other ego nodes, but with alters from other ego nodes. The agglomerative way of bringing communities together from the greedy algorithm might have caused this. Despite this, it is still a sign of closeness between the different ego networks. Additionally, some other communities that do not include any ego node were found, all belonging to the same ego network in each case.

From this table, especially on the community ID 7, it was found that some ego networks were overlapping, alters had connections to alters from other ego nodes. This was plotted on a table.

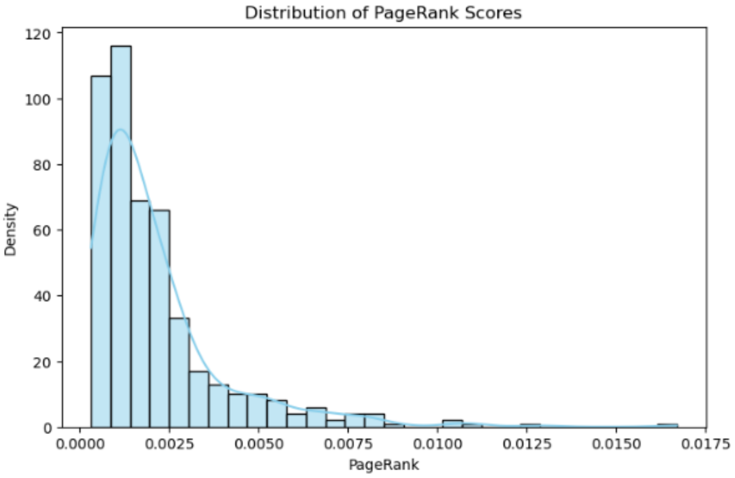
Overlap of alters in ego networks

Comm ID	107	348	414	686	698	3437	3980	1912	1684	0
107	NaN	17	17	0	0	1	0	6	14	2
348	17	NaN	45	0	0	1	0	2	1	4
414	17	45	NaN	0	0	1	1	2	1	3
686	0	0	0	NaN	27	1	0	0	0	0
698	0	0	0	27	NaN	2	0	0	1	0
3437	1	1	1	1	2	NaN	0	0	0	0
3980	0	0	1	0	0	0	NaN	0	0	0
1912	6	2	2	0	0	0	0	NaN	1	2
1684	14	1	1	0	1	0	0	1	NaN	3
0	2	4	3	0	0	0	0	2	3	NaN

Comparison

For comparison, another network was investigated. This time it was a weighted directed network representing Twitter interactions between US Congress representatives. (Leskovec, Twitter Interaction Network for the US Congress, 2023).

The same analysis was conducted, and different results were found. This dataset's edges had weights representing the number of interactions of one congressperson to another, in a directed manner.

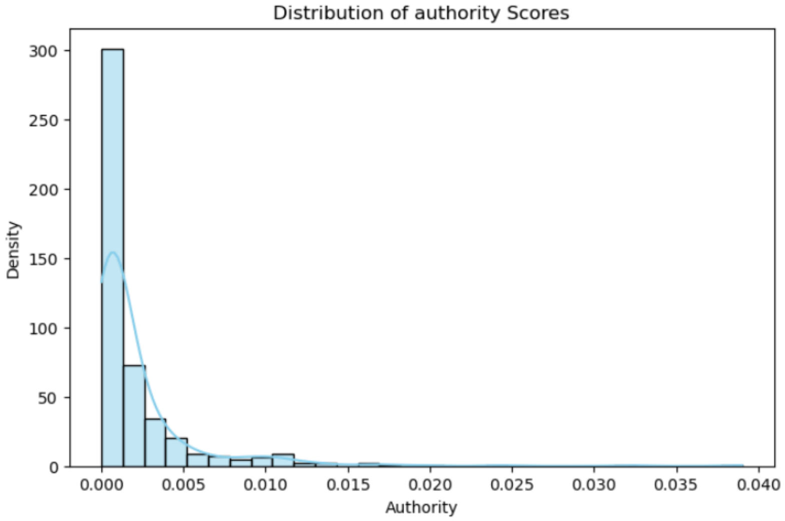


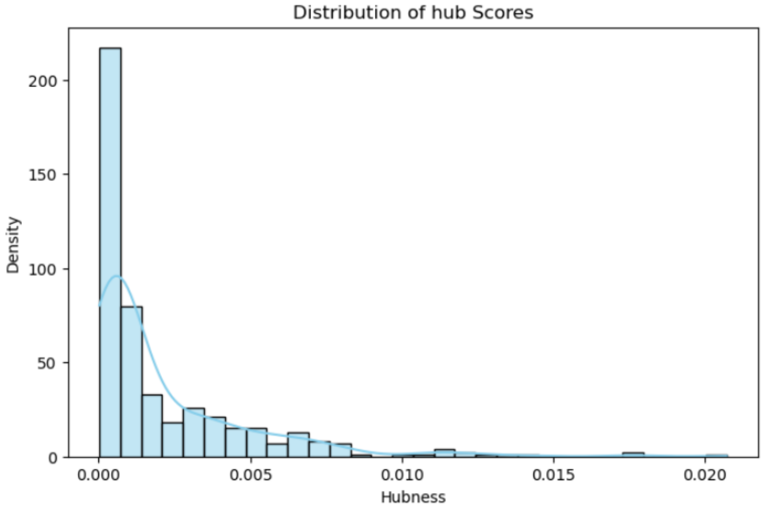
This time, the distribution of the PageRank scores had a greater variety, reflecting the smaller, more interconnected network. This contrasts with the results in the Facebook network, which was centered around ego nodes, unlike this Twitter network.

username	pagerank	betweenness	hub	authority
GOPLeader	0.016696	0.070677	0.011312	0.039029
RepCasten	0.012845	0.049959	0.000374	0.002795
RepChipRoy	0.011028	0.037404	0.004206	0.032205
RepMikeJohnson	0.01066	0.009661	0.005424	0.024478
RepChuyGarcia	0.01058	0.037622	0.000601	0.002055

Among the top 5 nodes with highest PageRank, we find some of the most influential figures of the main parties in congress, meaning they also got the most interactions on their Tweets. We can see that GOPLeader got far more interactions from important nodes than any other member of congress and has a high hub score compared to the other members with the top 5 PageRank.

Distributions of hubness and authority are different in this network than the Facebook one, as displayed in the following figures, although they share a similar distribution. Also, since this is a directed graph, different tables for maximum authorities compared to the maximum hub scores.





The distribution of hub and authority scores demonstrates the variation between the Facebook net in two ways. First, they are different plots, because this is a directed graph, different from the undirected Facebook graph. Second, it is not based on an ego network but on a known group of people that commonly interact with each other, so the measures spread out more evenly along the maximum and minimum hubness values.

The top 5 members of congress on authority and hub values are entirely different.

Top 5 Authority Scores

Username	Pagerank	Betweenness	Hub	Authority
GOPLeader	0.016696	0.070677	0.011312	0.0390
RepChipRoy	0.011028	0.037404	0.004206	0.0322
RepMikeJohnson	0.01066	0.009661	0.005424	0.0245
RepAndyBiggsAZ	0.008026	0.039888	0.008104	0.0196
CongressmanHice	0.008157	0.005239	0.006279	0.0174

Top 5 Hub Scores

Username	PageRank	Betweenness	Hub	Authority
RepBobGood	0.000975	0	0.020715	0.000926
RepCloudTX	0.001049	0	0.017802	0.003543
SteveScalise	0.00493	0.000616	0.017448	0.010878
RepJamesComer	0.00111	0	0.014486	0.00145
VernBuchanan	0.001635	0	0.013216	0.004252

These tables reflect how HITS portrays two different results when calculating authority and hubness in a directed network. In this case, we can see authority correlates closely with PageRank, since these important nodes are found on both algorithms, signifying a high importance on the network.

However, hubness shows a completely different Top 5. This Top 5 may be the members of congress who interact the most with the most nodes with the highest authority in the network. Interestingly, betweenness of this top 5 is extremely low, potentially showing how they are not acting as bridges among different communities, but instead interacting a lot with the nodes with the most authority.

Bibliography

Fink, C. G., Fullin, K., Gutierrez, G., Omodt, N., Zinnecker, S., Sprint, G., & McCulloch, S. (2023, March 16). *A centrality measure for quantifying spread on weighted, directed networks.* Retrieved from Cornell University Arxiv: https://arxiv.org/abs/2303.09684

Leskovec, J. (2023). *Twitter Interaction Network for the US Congress.* Retrieved from Stanford Network Analysis Project: https://snap.stanford.edu/data/congress-twitter.html

Leskovec, J. (n.d.). *Social Circles: Facebook*. Retrieved from Stanford Network Analysis Project: https://snap.stanford.edu/data/ego-Facebook.html

Leskovec, J., & McAuley, J. (2012). *Learning to Discover Social Circles in Ego Networks.* Retrieved from http://i.stanford.edu/~julian/pdfs/nips2012.pdf