NETWORK SCIENCE MINI PROJECT REPORT

Twitter Follow Graph

Group 2:

Rahul Maheshwari (MT19027)

P. Akshay Kumar (MT19094)

Piyush Pradeep Jain (MT19122)

Objective:

- To create a Twitter follow graph network and analyze its basic properties such as degree distribution, average path length, average clustering coefficient etc.
- To compare Twitter network with other social networks such as Facebook, LinkedIn, Flickr etc.
- To analyze if Twitter is a social network or an information network.

Datasets analyzed:

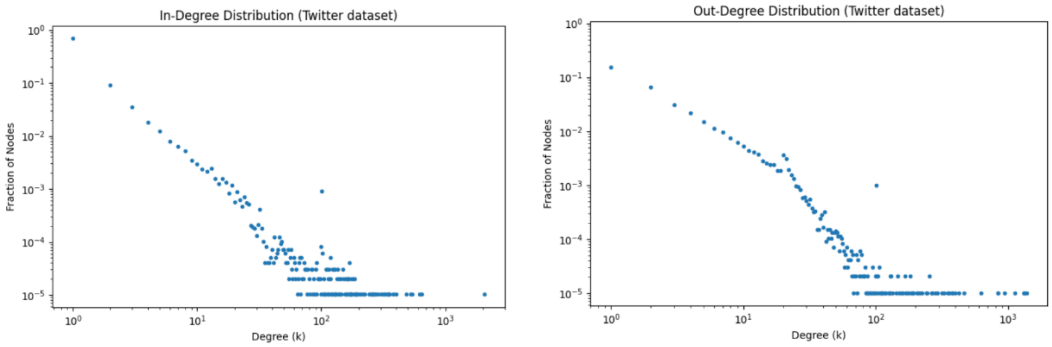
Dataset	Source
Twitter Follower Network, EU Email Network	SNAP repository
LinkedIn, Flickr	http://networkrepository.com
Twitch, Facebook (unlabeled)	https://github.com/benedekrozemberczki/datasets#twitch-ego-nets
Facebook (labelled)	https://www.kaggle.com/rozemberczki/musae-facebook-pagepage-network

Note: Twitter dataset size was 24GB. We sampled it in order to estimate properties like path length, diameter.

Additional work done:

- We developed a GUI to show the results in a succinct and presentable manner.
- Analyzed advanced properties on Twitter and Facebook networks such as community detection and page rank.
- Developed a Twitter-Ego dataset from Twitter dataset to analyze it's properties.

Twitter Degree Distribution plots:



Observations:

- From the above plots, it is evident that Twitter follow graph follows scale distribution due to presence of large number of hubs.
- An interesting takeaway from the plots is that outdegree distribution has a greater number of hubs with respect to indegree distribution. This can be attributed to the fact that there exists a decent number of accounts which follow huge number of other user accounts. These could be bots or news related accounts which follow huge number of accounts to get the latest updates.

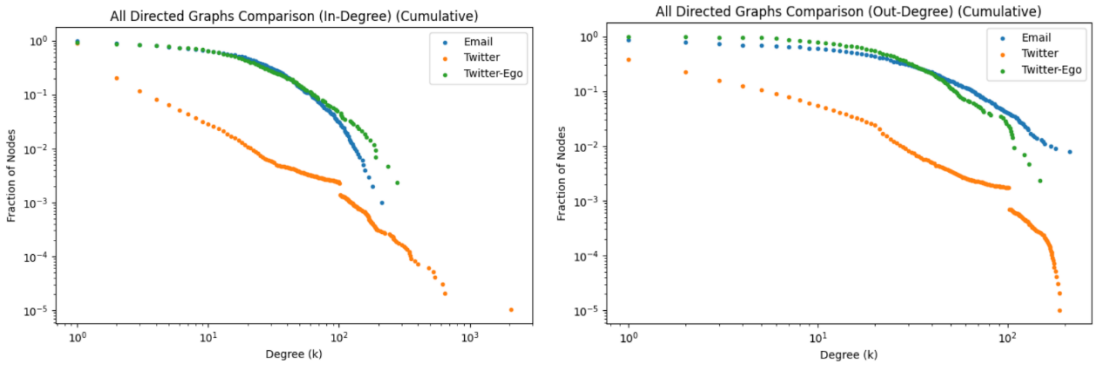
Twitter Network properties:

Network	Number of Nodes	Number of Edges	Average Path Length	Average Clustering Coefficient	Degree Assortativity	Diameter
Twitter	98292	212314	2.612596	0.076286	-0.08114	18

- From the low average clustering coefficient value, we can interpret that very few of the neighbors of nodes are mutually connected. Also, the low average path length is indicative of information percolating through the network faster. Hence, from the sample of dataset used from the Twitter follow graph dataset, we can interpret that **Twitter is more of an information network than a social network.**

Degree Distribution comparison plots:

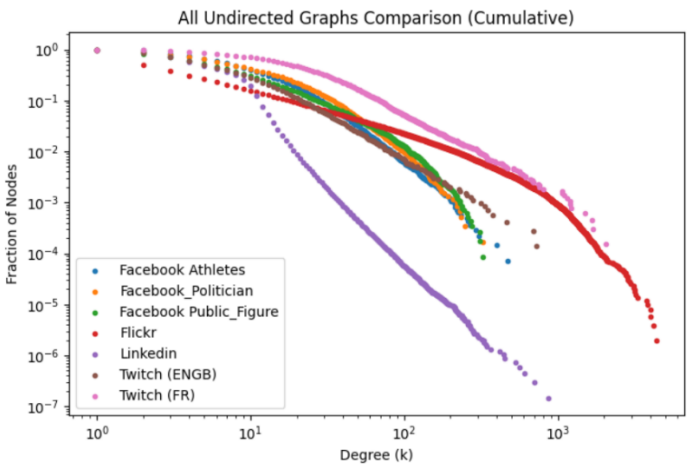
Directed Graphs – Twitter, Twitter-Ego and Email



Observations:

- All the plots follow power law distribution indicating presence of scale-free property.
- The behavior of Twitter-Ego and Email plots are similar due to the fact that Twitter-Ego is a network of hubs in Twitter dataset and Email network is composed of nodes which are well connected in terms of contacts which are essentially hubs as well.

Undirected Graphs – Facebook, Twitch, Flickr, LinkedIn



Observations:

- All undirected cumulative plots follow power law indicating scale-free nature of networks.
- An interesting takeaway from the above plot is that of LinkedIn network where the slope of the plot is significantly different (lesser) from the other plots. This is due to the fact that connections in LinkedIn are interconnected i.e., the network has relatively high clustering coefficient hence resulting in lower slope of the plot.

Network properties of remaining networks:

Network	Number of Nodes	Number of Edges	Average Path Length	Average Clustering Coefficient	Degree Assortativity	Diameter
Twitter-Ego	429	11416	2.505563	0.28431	-0.14497	8
EU Email	1005	25570	2.652826	0.372703	0.004335	7
Twitch_FR	6549	112666	2.680991	0.221706	-0.17815	7
Twitch_ENGB	7126	35324	3.677616	0.130928	-0.12191	10
Linkedin	2122488	4626421	-	0.267394	-0.0732	-
Flickr	513969	3190452	-	0.167599	0.157802	-
Facebook_Public_Figure	11565	67114	4.622979	0.179347	0.202162	15
Facebook_Politician	5908	41729	4.664107	0.385096	0.018244	14
Facebook_Athletes	13866	86858	4.274809	0.276188	-0.0269	11

Observations:

- Twitter-Ego in contrary with Twitter network show higher clustering coefficient indicating the presence of tightly knit communities between hub nodes of the Twitter follow graph network.
- Social networks such as Facebook, Email show high clustering coefficient.
- The path of networks is small indicating that information flows very fast on all these networks.
- Twitch network is disassortative due to the fact that the most popular gamers take part in many showdowns for the games they stream and end up following other not so popular streamers who are part of the showdowns.
- Facebook and Flickr network predominantly show assortative behavior since people are likely to follow their family and colleagues who are likely to have friends of the same order as them.

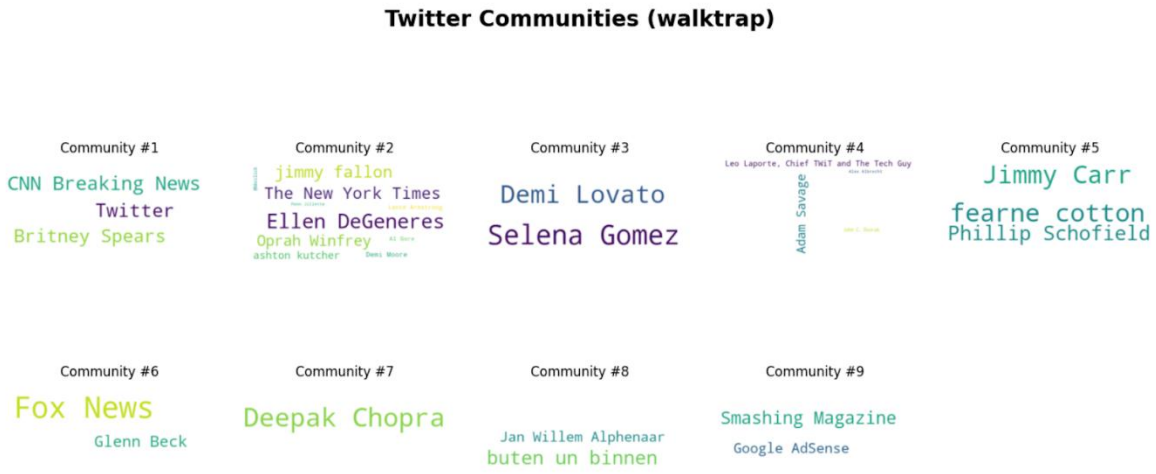
Note: The average path length and diameter fields are blank for LinkedIn and Flickr because of high density of the networks and the time taken to compute the properties was too high even with BFS flood fill algorithm.

Community detection algorithms:

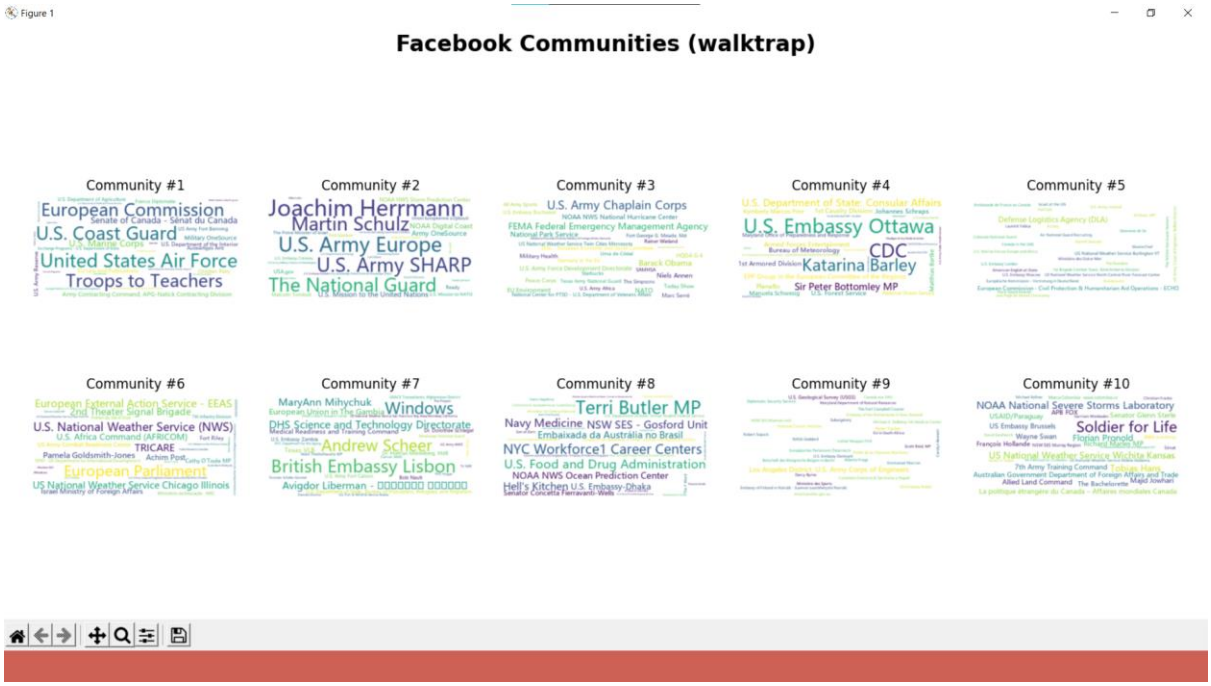
The following community detection algorithms were applied

1. Leiden
2. Surprise communities
3. Walktrap

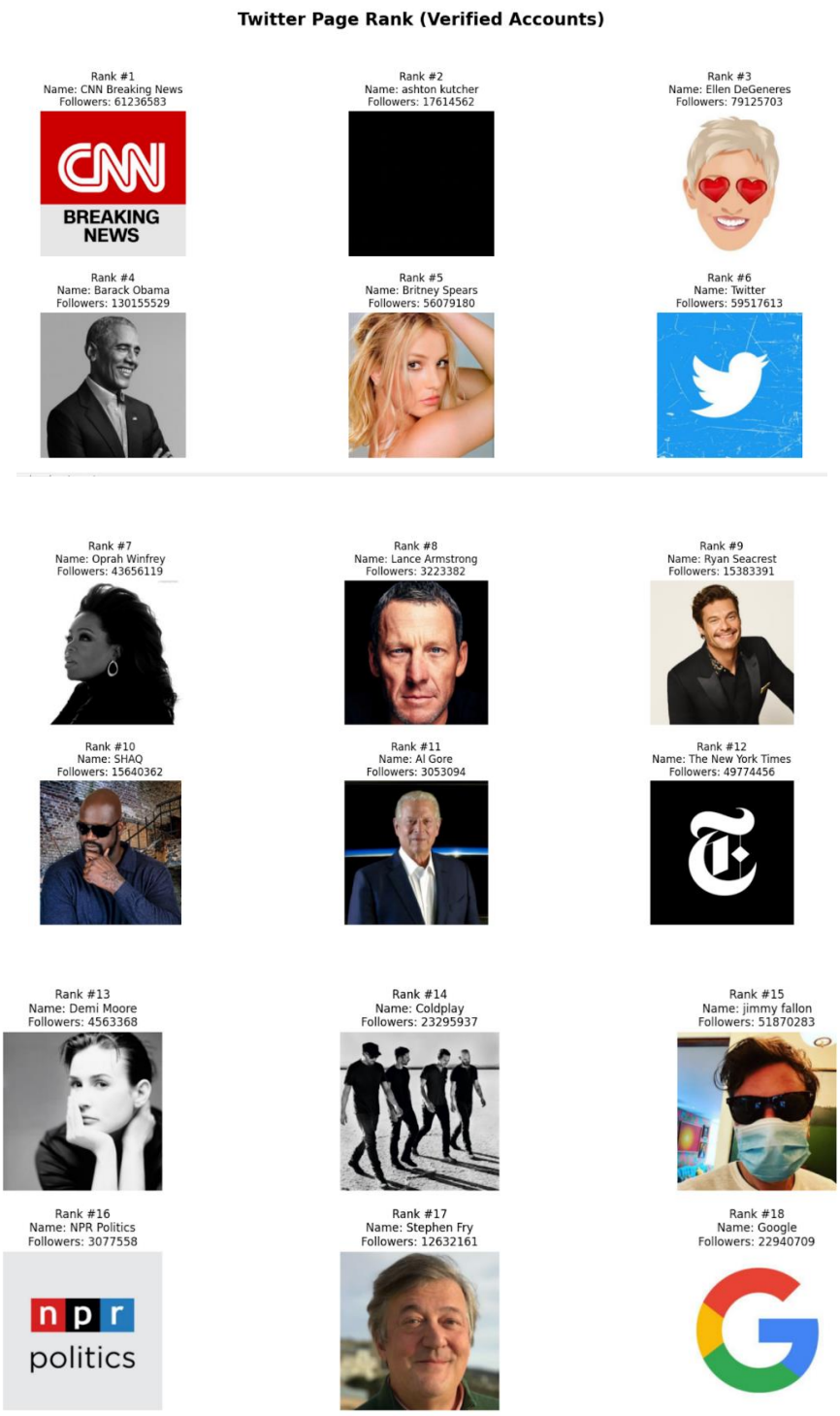
Walktrap community detection algorithm output on Twitter dataset



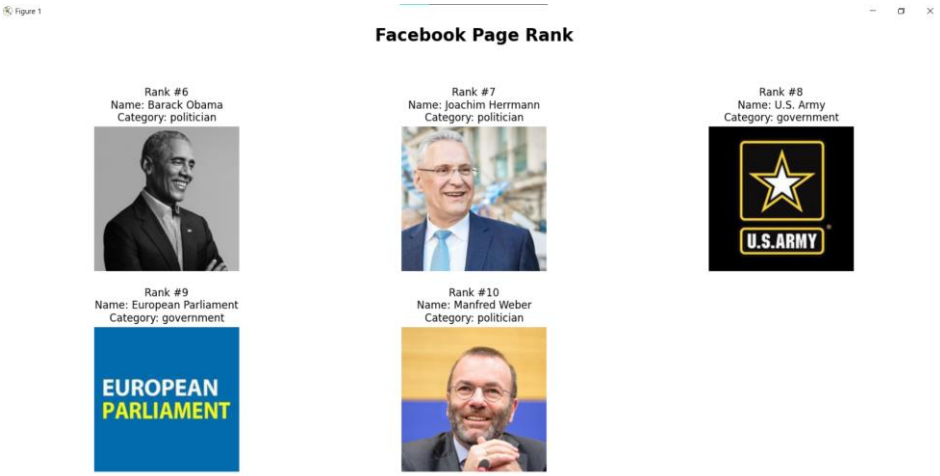
Walktrap community detection algorithm output on Facebook dataset



Page rank output on Twitter Dataset [Top Results]:



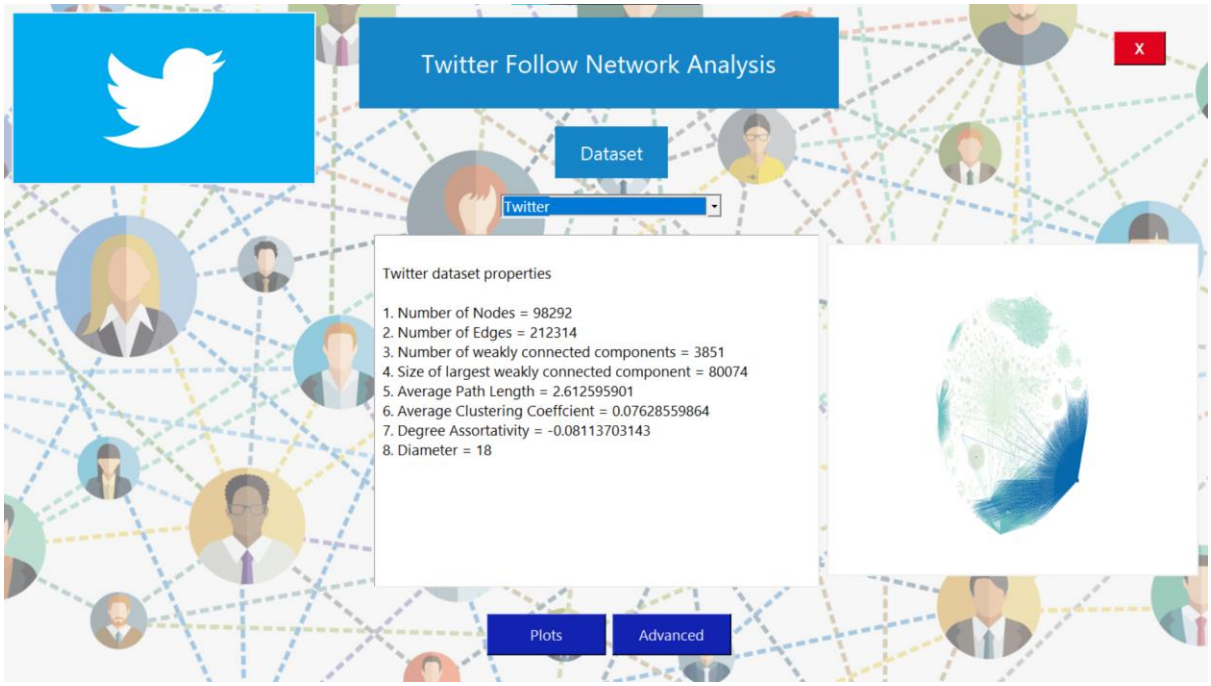
Page rank output on Facebook Dataset [Top Results]:



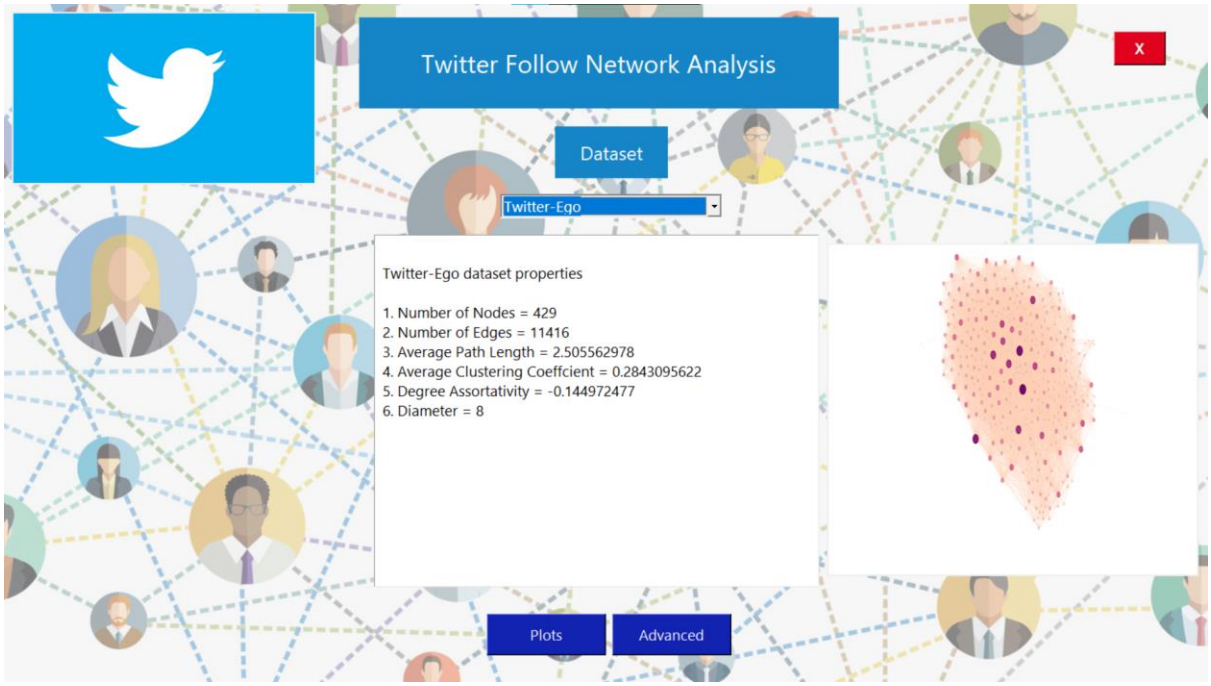
GUI Screenshots

Main Screen

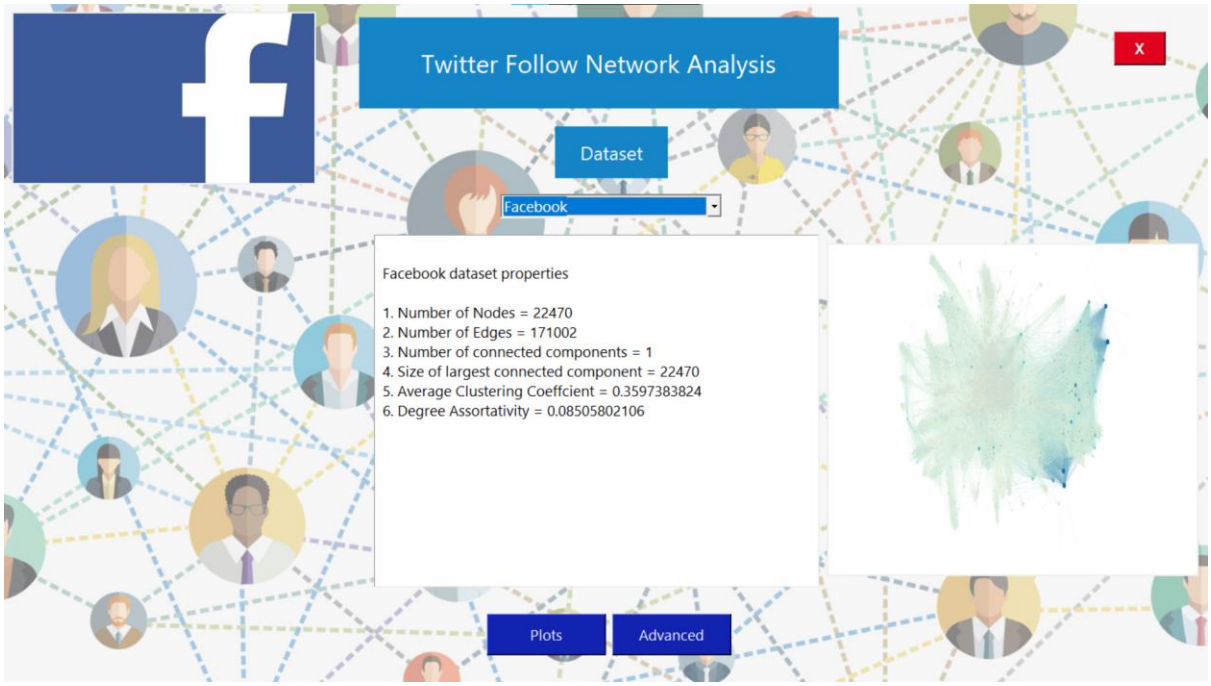
Twitter main screen



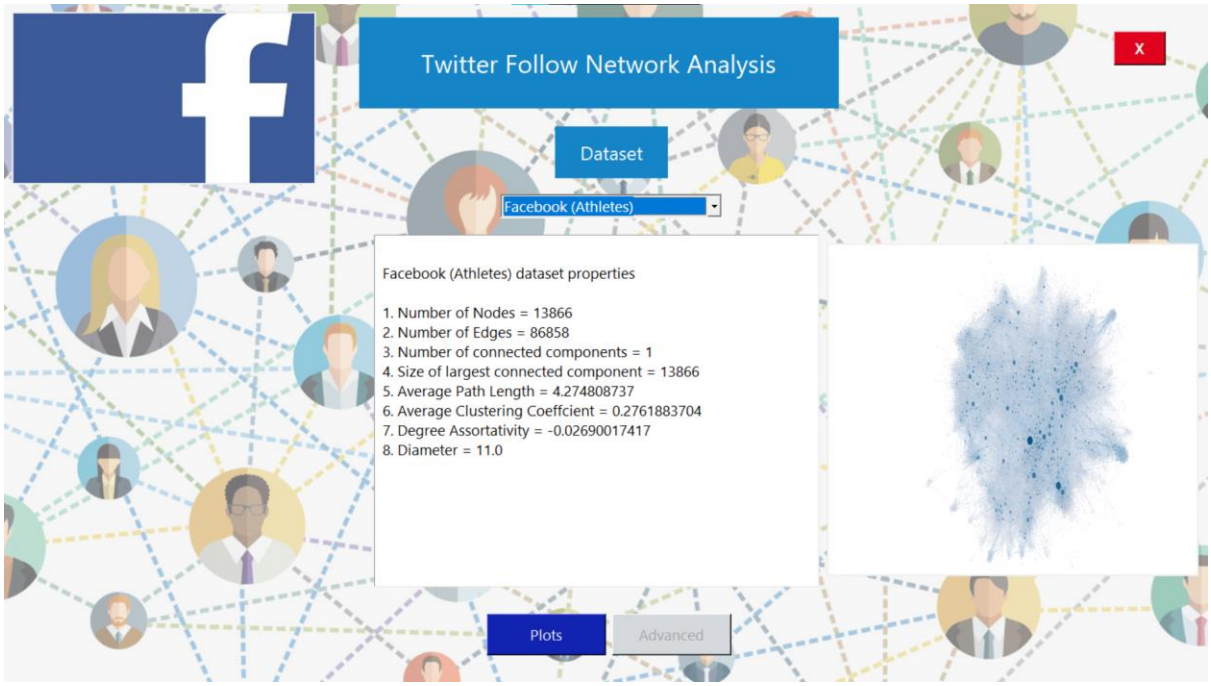
Twitter ego main screen



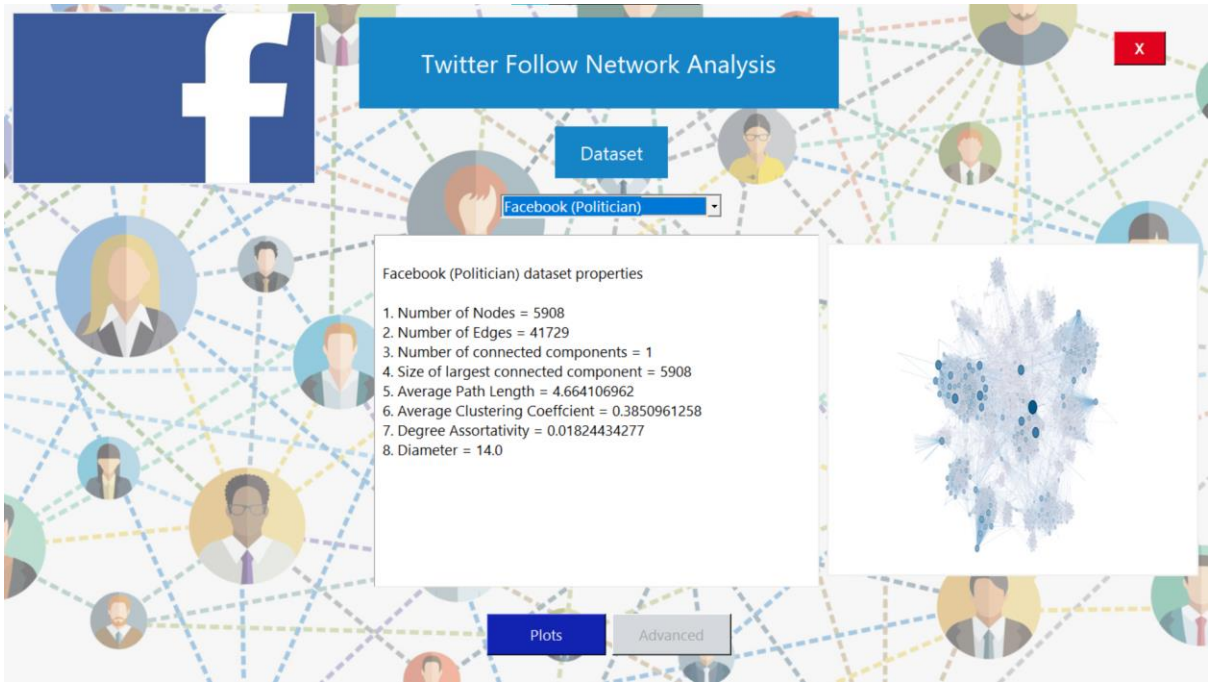
Facebook main screen



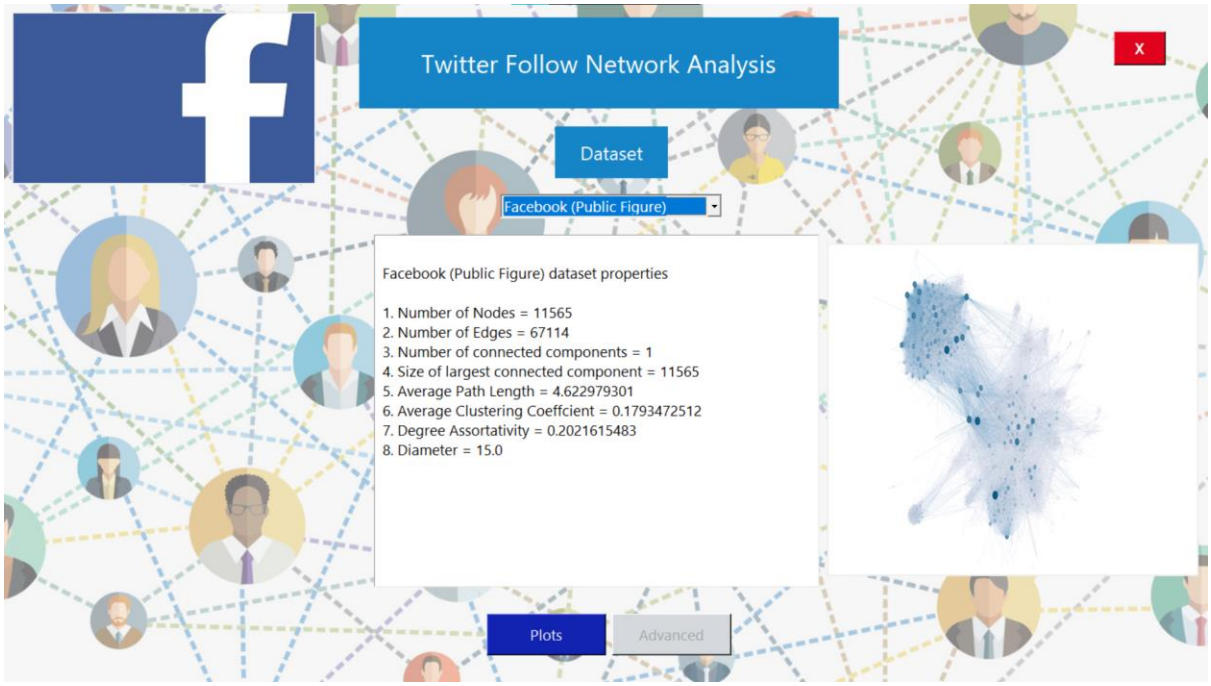
Facebook athletes main screen



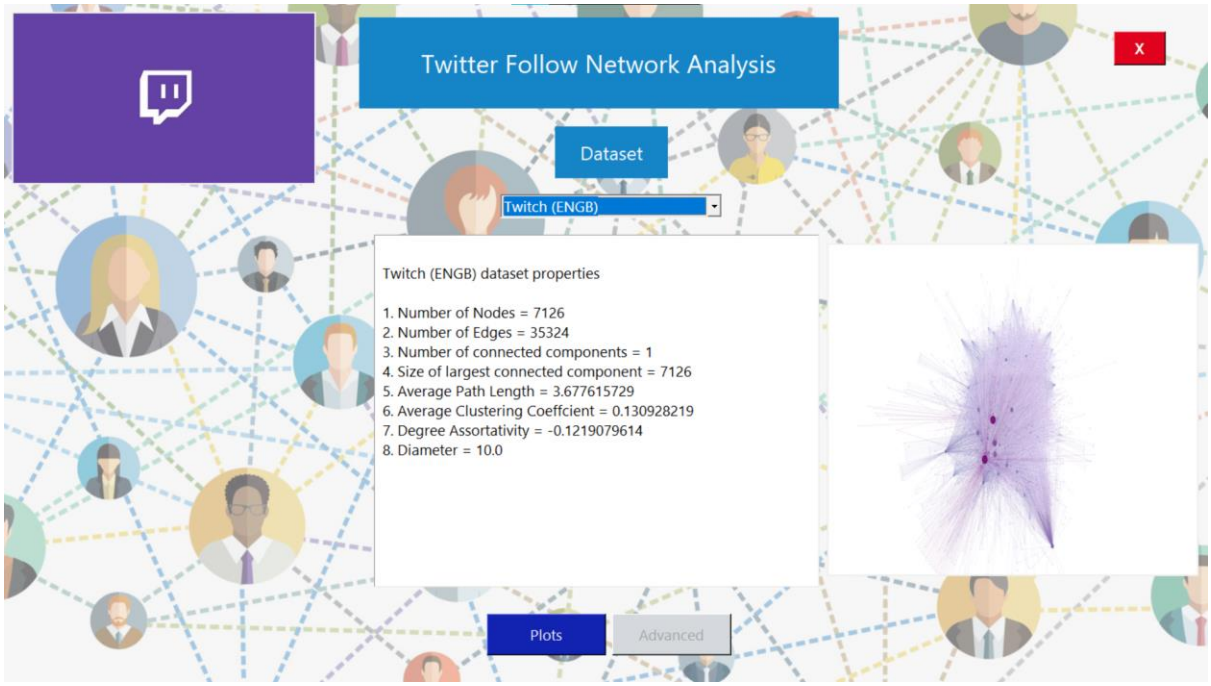
Facebook politician main screen



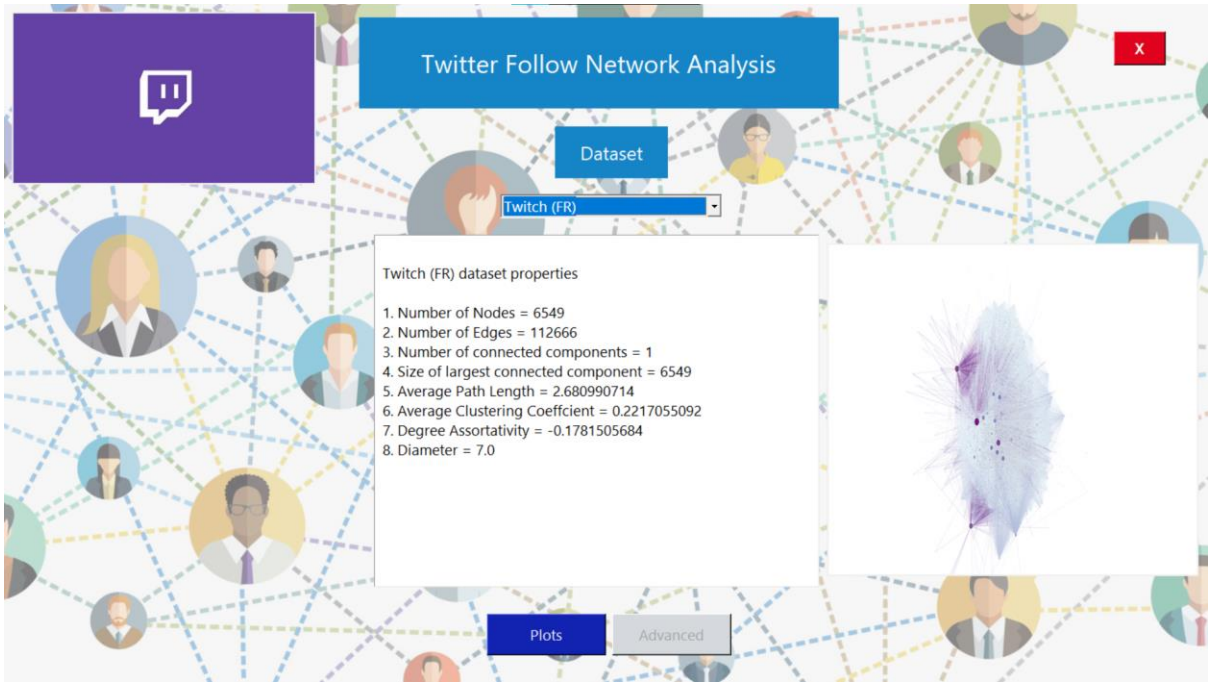
Facebook public figure main screen



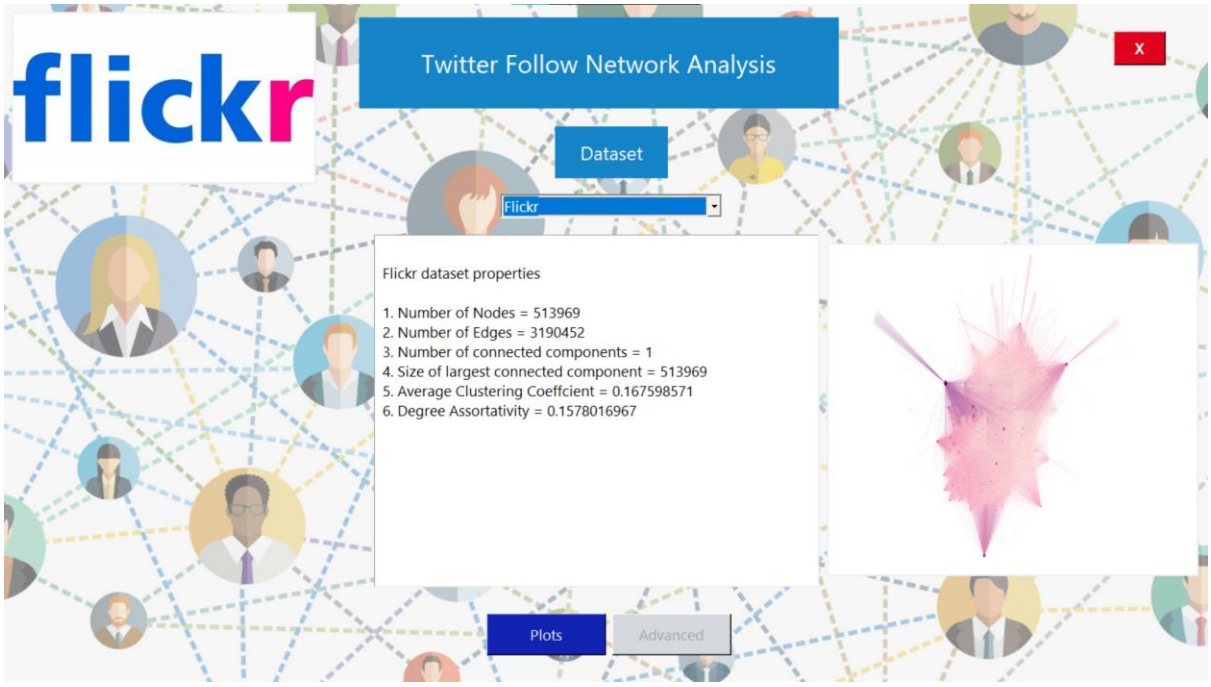
Twitch [ENGB] main screen



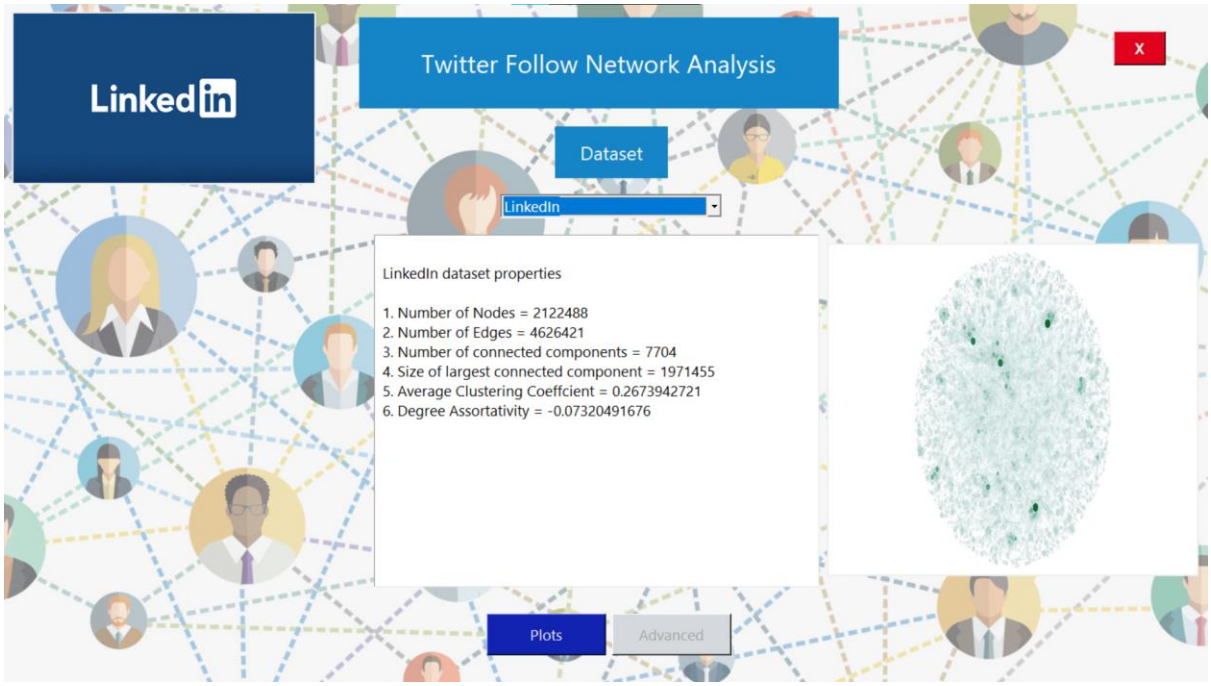
Twitch [FR] main screen



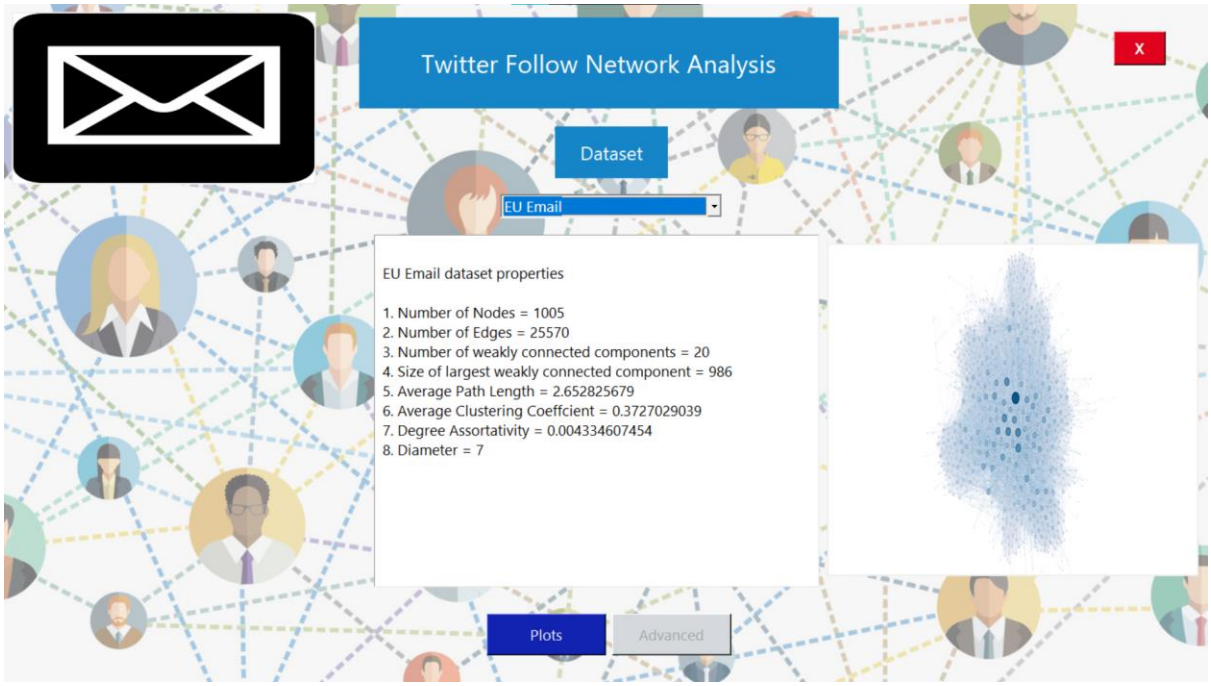
Flickr main screen



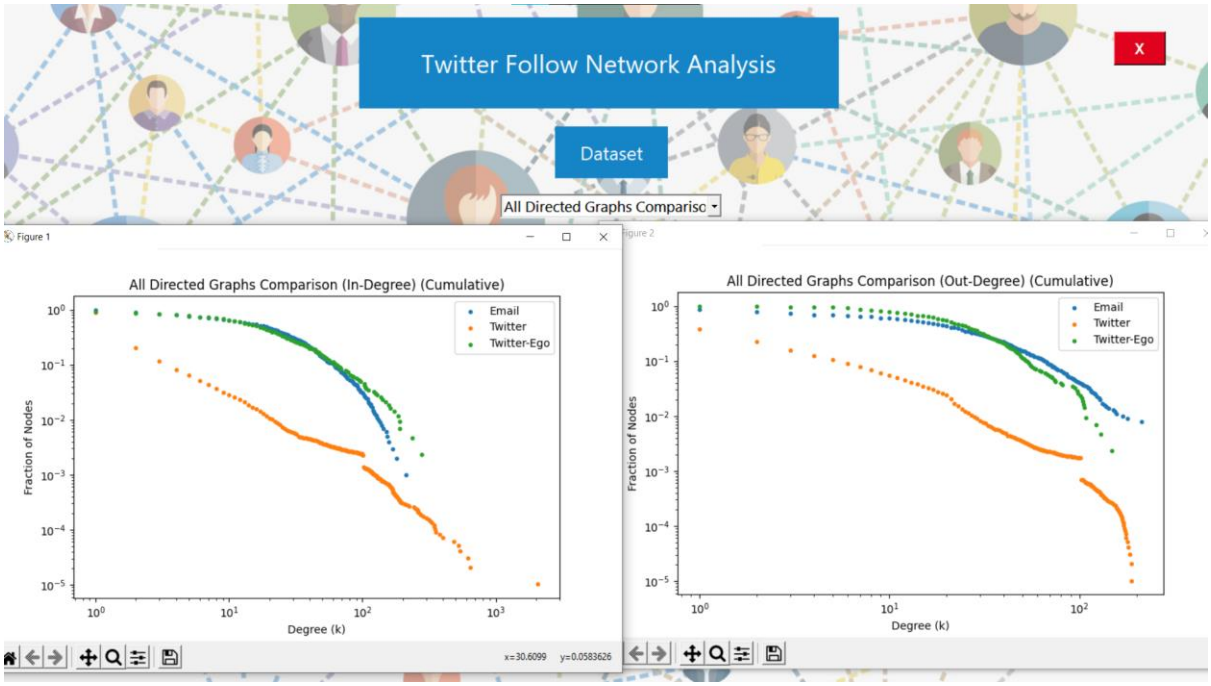
Linkedin main screen



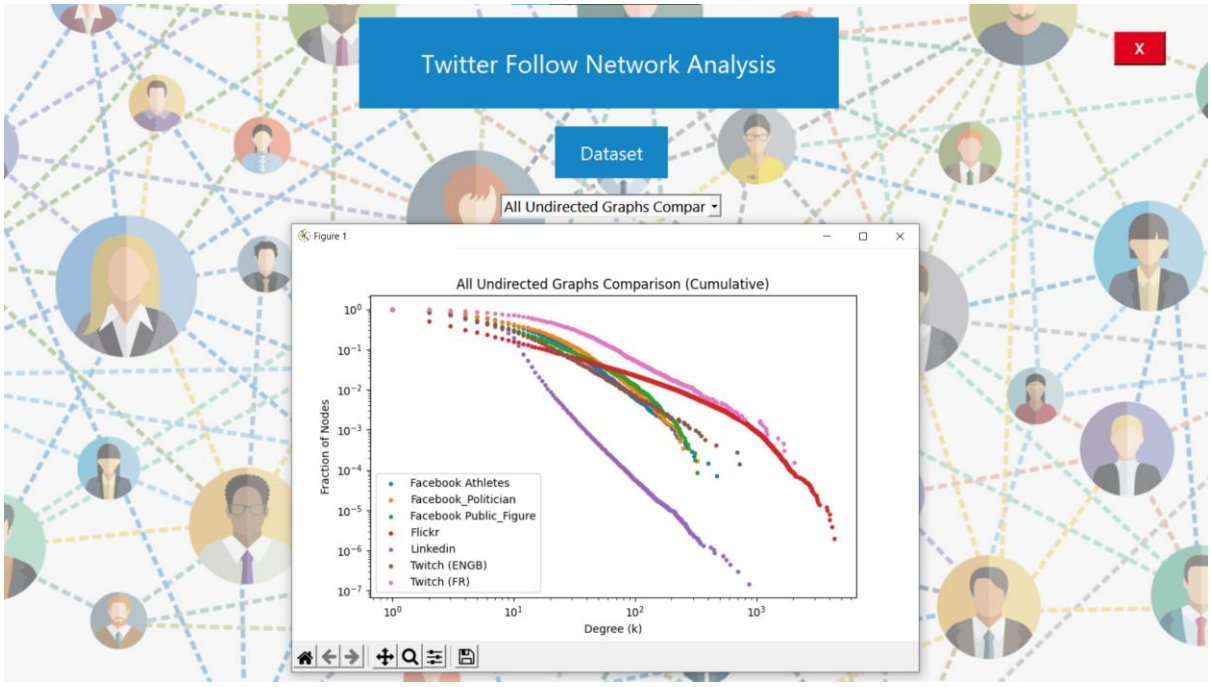
EU Email main screen



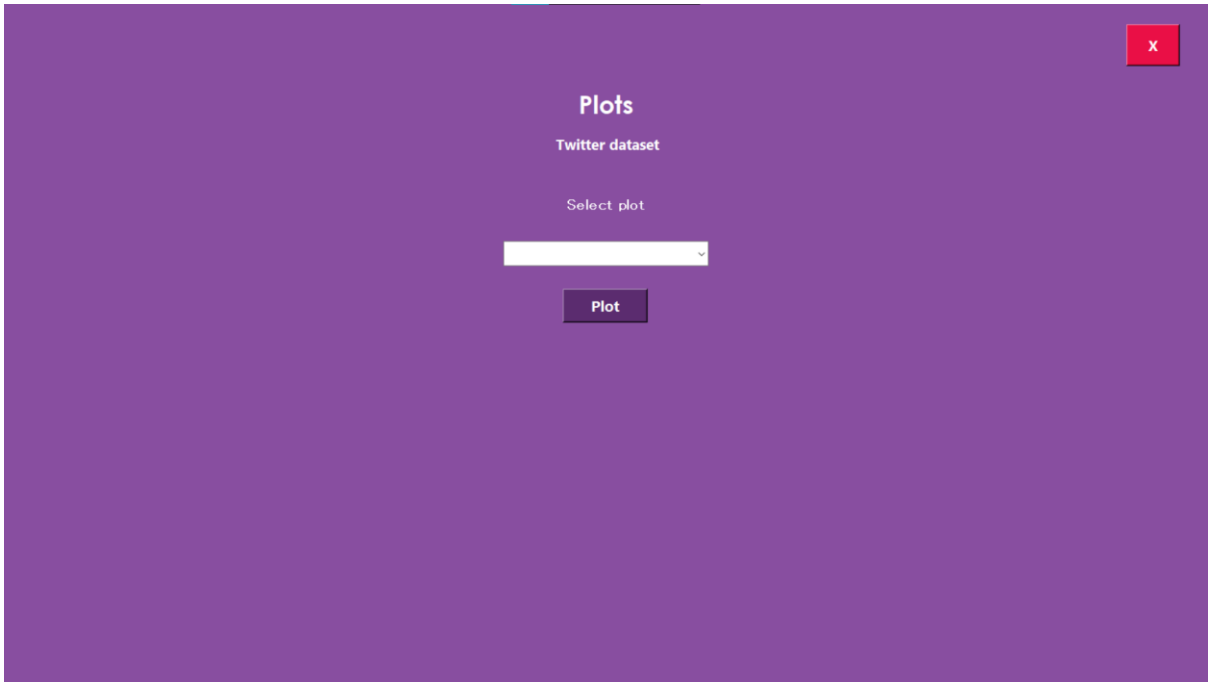
All directed graphs cumulative degree distribution comparision

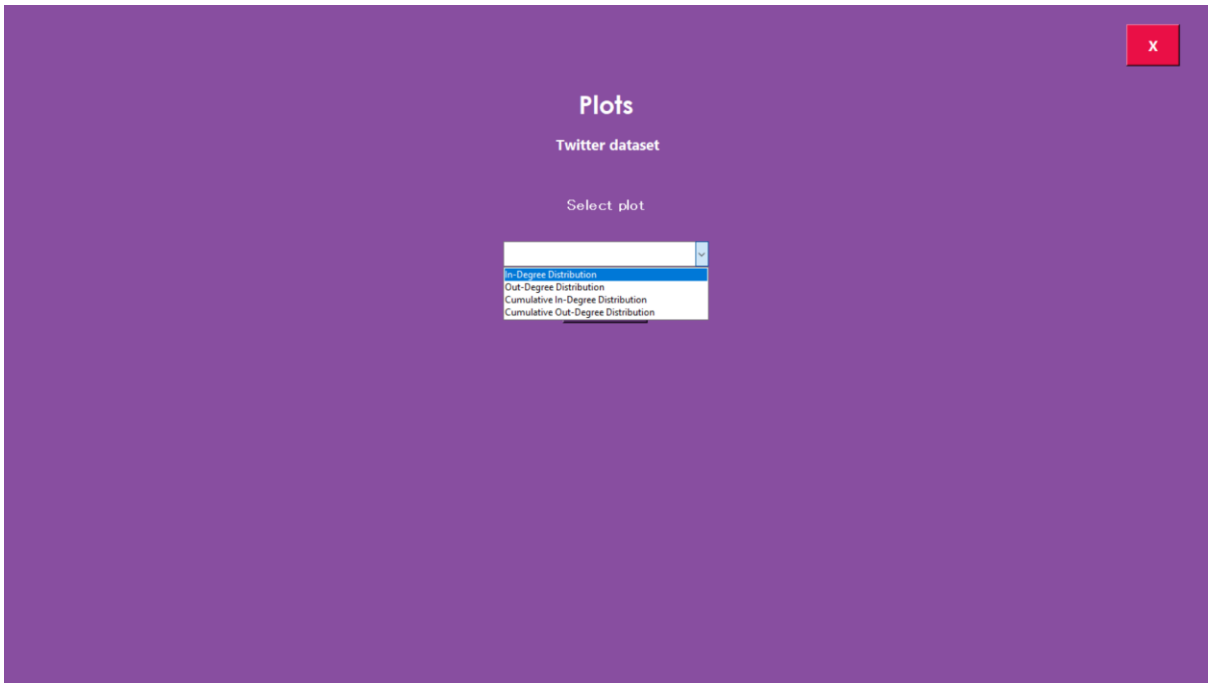


All undirected graphs cumulative degree distribution comparision



Plots screen





Advanced screen