19OH01 - SOCIAL AND ECONOMIC NETWORK ANALYSIS

Group Project
"Instagram Network Analysis"

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING

April 2021

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

Arun Prasad P - 18z207
Jagathees S - 18z219
Navenkumar D - 18z234
Tadepalli Siva Koushik - 18z258
Lokesh R - 19z434

CONTENTS

INSTAGRAM NETWORK ANALYSIS

Problem Statement

Considering an instagram account, find the list of users who follow the account holder and being followed back. Establish links between these nodes based on their following. Visualize this network in the form of a graph. Analyze factors such as average degree, average path length, network diameter, modularity, centrality measures, number of connected components etc. Find communities between the users. Identify the hubs among the users to broadcast information or advertise elements.

Dataset Description

The dataset has been obtained by scrapping the instagram profile through phantom buster. Phantom buster (following collector or follower collector) when provided with the profile link it scraps all the followers/following of that particular profile. The resulting dataset can be downloaded in the form of .csv format.

Data Gathering

The list of followers and following of the profile aruntrendzzz were downloaded in the form of csv through phantom buster. Let the term best-friend of a particular profile be described as the users who follow the account holder and being followed back. The profile aruntrendzzz is being followed by 248 profiles and follows 418 profiles. Among the 418 following profiles **72 profiles** were public and followed aruntrendzzz profile back. Private profiles were removed as they don't provide any more details such as followers or following which is required for analysis. These 72 profiles are considered to be the best friends of the profile aruntrendzzz.

For each of the best friends' profiles the following profiles are extracted through phantom buster. The following of all the best friends' profiles are merged into a single .csv file (followingMerged.csv). All the columns except profileURL and user are dropped. The tuple (user, profileURL) indicates that the user follows profileURL. There are a total of **53595 such tuples (edges) between 42281 users (nodes)**.

Another data set (bestFriendNetwork.csv) is made by establishing edges among the 72 nodes (best friends). This is done by performing an inner join on the best friends' dataset and followingMerged dataset on the user column. This results in a dataset with 459 rows (edges) with 67 nodes. Adding isolated nodes results in **564 rows and 72 nodes**. The datasets can be viewed through the link.

Tools Used

Networkx

Networkx is a python package for creating and manipulating graphs and networks.

Igraph

Igraph is a python library and has high performance towards graph data structure and algorithms.

Pandas

Pandas is a Python library used for analyzing, cleaning, exploring, and manipulating datasets.

Gephi

Gephi is open-source software for visualization of graphs and networks.

Jupyter Notebook

Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

Challenges Faced

The dataset required was not readily available. So, a suitable dataset had to be created that matches the requirements. Collecting the details of the accounts which were private was not possible, which would have given better results if used in the network. While collecting the data few of the users were either making their accounts private or changing their usernames, which resulted in few errors while collecting data related to those accounts.

Scrapping of instagram profiles using selenium web driver takes too much time. When error occurred the code stops executing and requires restart, which led to multiple logins. Instagram finds this activity to be suspicious and prevents logging in into that profile for several hours.

Initially edges between users were established based on their followers which lead to the omission of accounts that have higher indegree (i.e. being followed by many) and has zero out degree (i.e. does not follow anyone back among the profiles collected).

Girvan Newman community detection requires the calculation of edge betweenness for all the edges which is highly time consuming for large graphs. So establishing edges only between those 72 best friends profile a new graph was created.

Contribution of Team Members

Roll No	Name	Contribution
18z207	Arun Prasad P	bestfriendsgraph.csv, centralities, PageRank
18z219	Jagathees S	bestfriends.csv, community detection
18z234	Navenkumar D	suggestions users to follow, Gephi
18z258	Siva Koushik T	followingmerged.csv, hubs and authorities
19z434	Lokesh R	SCC, Giant component, avg. path length

Annexure I: Code

https://github.com/navenduraisamy/INSTAGRAM-NETWORK-ANALYSIS

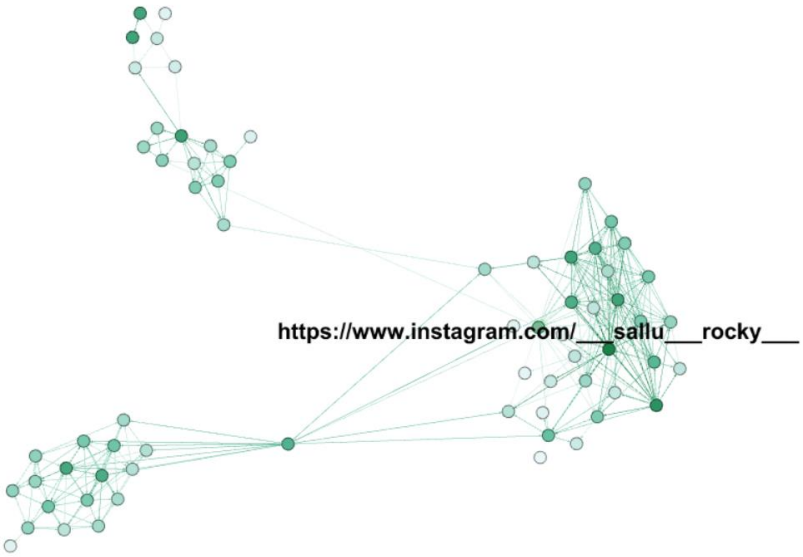
Annexure II: Snapshots of the Output


```
['https://www.instagram.com/_.b.k.g._', 0.3775208461572099]
['https://www.instagram.com/___sallu___rocky___', 0.37041778281288784]
['https://www.instagram.com/lordboros2000', 0.2594988344988345]
['https://www.instagram.com/ilayaraja033', 0.1839487477686778]
['https://www.instagram.com/sri_ram_aravind_sai', 0.18362470862470862]
```

The user https://www.instagram.com/_.b.k.g._ has the highest betweenness centrality score of all nodes because it serves as the bridge node between the school community and college community.

```
Clustering with 67 elements and 4 clusters
[0] https://www.instagram.com/prasannavijay92,
    https://www.instagram.com/prp_raj_praveen_12321,
    https://www.instagram.com/vijai_krissh,
    https://www.instagram.com/__nanthu__,
    https://www.instagram.com/lewin_kingsly,
    https://www.instagram.com/terific_eswar,
    https://www.instagram.com/srv2018passedouts,
    https://www.instagram.com/b.nandagopal,
    https://www.instagram.com/_well_wisher.__,
    https://www.instagram.com/__a.l.f.r.e.d,
    https://www.instagram.com/parker_prasath__,
    https://www.instagram.com/kavii.07, https://www.instagram.com/_.b.k.g._,
    https://www.instagram.com/_nithish_the_legend,
    https://www.instagram.com/manikandan_thamaraiselvan,
    https://www.instagram.com/sudar7san_vpm,
    https://www.instagram.com/ghost_gopal_2.o,
    https://www.instagram.com/naveen_domnic
[1] https://www.instagram.com/ilayaraja033,
    https://www.instagram.com/_the_eminent_artist_,
    https://www.instagram.com/____preetham____,
    https://www.instagram.com/sleeper_cellz_of_psg,
    https://www.instagram.com/rj_mokshith,
    https://www.instagram.com/manojh_sivakumar,
    https://www.instagram.com/dhayanandh_at,
    https://www.instagram.com/book_ishq,
    https://www.instagram.com/mrwirelessbrain,
    https://www.instagram.com/sri_nagul_baskar,
    https://www.instagram.com/a_shibu_musical_official,
    https://www.instagram.com/ttrnni,
    https://www.instagram.com/__.v_.i_.s_.h_.n_.u__,
    https://www.instagram.com/events_at_coimbatore,
    https://www.instagram.com/ajay133734,
    https://www.instagram.com/adhavanalexander,
    https://www.instagram.com/the_hacker_2.o,
    https://www.instagram.com/_rishikeshwaran_,
    https://www.instagram.com/nitin_sundararaj,
    https://www.instagram.com/___sallu___rocky___,
    https://www.instagram.com/akil_yadhav,
    https://www.instagram.com/rohith_ck_,
    https://www.instagram.com/sanjay_sandy3,
    https://www.instagram.com/projects_360_,
    https://www.instagram.com/__the_lonely_king__,
    https://www.instagram.com/ajey729, https://www.instagram.com/srinath_nsk_,
    https://www.instagram.com/an_engineer_sketch,
    https://www.instagram.com/_pirate_10,
    https://www.instagram.com/sak_ajmal298,
    https://www.instagram.com/ashwin7818,
    https://www.instagram.com/techy_trolls
[2] https://www.instagram.com/sri_ram_aravind_sai,
    https://www.instagram.com/ak__photography0007,
    https://www.instagram.com/c_a_p_t_u_r_e__2081,
    https://www.instagram.com/lordboros2000,
    https://www.instagram.com/being_h_uman,
    https://www.instagram.com/camera_2081,
    https://www.instagram.com/ananth_wolf,
    https://www.instagram.com/_bad_revan_,
    https://www.instagram.com/aakash1072000,
    https://www.instagram.com/h_e_a_v_e_n_b_o_y_0001,
    https://www.instagram.com/newbharathschooltiruvarur
[3] https://www.instagram.com/arunbarathdhon,
    https://www.instagram.com/arunbarath53,
    https://www.instagram.com/_dhavaseelan,
    https://www.instagram.com/arun.barath.54,
    https://www.instagram.com/_dhavadeep_, https://www.instagram.com/asridhar8
```

The people falling under cluster 0 studied in school with the profile holder. The users in cluster 1 are his college mates whereas cluster 2 and 3 are the users in his living area. The modularity score for this partition is **0.531172.**




```
PANGERANK:
['https://www.instagram.com/___sallu___rocky___', 0.03917349293278592]
['https://www.instagram.com/adhavanalexander', 0.034427201725232046]
['https://www.instagram.com/rj_mokshith', 0.030811868393061666]
['https://www.instagram.com/lordboros2000', 0.027975026561298296]
['https://www.instagram.com/sri_nagul_baskar', 0.027565728948938398]
```

The PageRank algorithm measures the importance of each node within the graph, based on the number incoming relationships and the importance of the corresponding source nodes.

```
('https://www.instagram.com/mahi7781', 23)
('https://www.instagram.com/rashmika_mandanna', 19)
('https://www.instagram.com/kp_1820', 18)
('https://www.instagram.com/mokkamemes', 18)
('https://www.instagram.com/akash_nambi_', 18)
('https://www.instagram.com/virat.kohli', 17)
('https://www.instagram.com/vijaytelevision', 16)
('https://www.instagram.com/meghaakash', 16)
('https://www.instagram.com/videomemes.vm', 16)
('https://www.instagram.com/chennaiipl', 16)
```

These nodes could be suggested to follow for the account holder aruntrendzzz. These nodes have been followed by many accounts that are being followed by aruntrendzzz.

References

https://github.com/mharkus/instagram-network-analysis/blob/master/Instagram%20Network%20Analysis.pdf

https://medium.com/@maximpiessen/how-i-visualised-my-instagram-network-and-what-i-learned-from-it-d7cc125ef297

https://www.youtube.com/watch?v=a11igVVDT4Y

https://www.alphr.com/how-does-instagram-know-friends/#:~:text=Phone%20Contacts%20%E2%80%93%20Instagram%20will%20also,make%20friend%20suggestions%20for%20you.&text=Mutual%20Friends%20%E2%80%93%20Instagram%20often%20suggests,your%20list%20of%20suggested%20friends.

https://www.python-course.eu/networkx.php

https://igraph.org/python/doc/tutorial/tutorial.html

https://www.datacamp.com/community/tutorials/social-network-analysis-python

https://gephi.org/tutorials/gephi-tutorial-quick_start.pdf?utm_source=dlvr.it&utm_medium=twitter

https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python

Plagiarism Scan Report