Data Mining – Home Assignment 4

Shweta Suran

[174782IAQD]

• Introduction

Analyzing the structure of a social network in collaborative Web-based platforms such as *Q&A* platforms (Reddit & Stack Overflow), *Citizen science* platforms (Zooniverse & MangroveWatch), *Knowledge & Information Exchange* platforms (Climate CoLab & hackAIR); helps in gaining insights into interactions and relationships among users while revealing the patterns of their online behaviour. Regularities, or patterns in relationships between social entities, can be used to characterize the social environment and even predict its further evolution.

For users to engage in a social network of a collaborative Web-based platform, use reputation scores are used. It helps users to find other users that they can trust in some context, as well as provide qualitative information to the researchers. In collaborative Web-based platforms, user reputation scores are generally computed according to *centrality-based reputation (CBR) scores*. In CBR approaches, a "who-trusts-whom" network also known as a trust network is available and the most reputable users occupy the most central position in the trust network, according to some definition of centrality.

Centrality measures are used to find the most reputable users (vertices, or nodes) in a network, which is important for web-based platforms and have a big influence on the dissemination of information over the web. It identifies key users in complex networks, analyze and predict factors such as trust & influencer, identifies potential user for communication activity, influencing users on social media or collaborative platforms.

Why Centrality measures? *Identifying the main influencers in a social network using centrality measures can help in tries to increase the speed of information spreading over the network, which could be used, for example, to decrease it in case of a cyber-attack by a hostile entity using misinformation or fake news.*

In this assignment, we investigate CBR scores based on six popular centrality measures—*Degree Centrality, Closeness Centrality, Betweenness Centrality, PageRank, Hubs-Authorities and Eigenvector Centrality*. Centrality is an important index because it indicates which node takes up critical position in one whole network. CBR are based on the fact that some platforms allow their members to explicitly declare what users they trust. The web of trust relationships can be viewed as a graph (called trust network) in which each user corresponds to a vertex/node and edges encode trust relationships. Furthermore, we also do clustering based on centrality to identify the prominent clusters in the network.

• Program

First, program load the data set in CSV format 'kangaroo_data.csv' and computes the total number of edges 'edgeCnt' and nodes 'nodeCnt' in the given data set. Then, it plots the input graph as shown in Figure 1. 'visIgraph' is used to visualize the one node connectivity with the others as shown in Figure 1.

Furthemore, program compute the clustering for each centrality measure based on Kmeans. To do this program use the inbuilt function for Kmean clustering i.e., 'kmeans'

- **Hyperlink-Induced Topic Search** (**HITS**) / **Hubs & Authorities**

Authorities estimates the node value based on the incoming links. Hubs estimates the node value based on outgoing links. *The idea of a hub is that a good hub points to good authorities and a good authority is pointed to by a good hub*. To compute hub and authority score an inbuilt function is used 'hub_score' & 'authority_score'. Figure 2 shows the clustering based on hubscore and the nodes with the high hubscore.

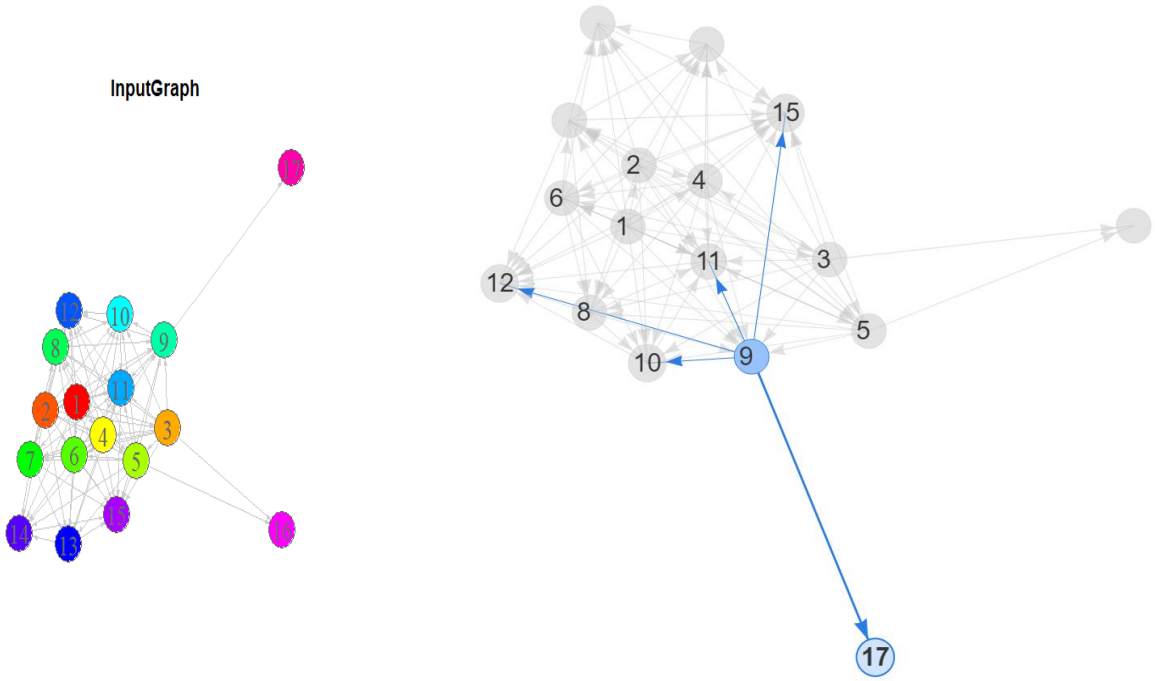

Figure 1: Input Graph (Left Side) and Node Connectivity of Node:17 (Right Side)

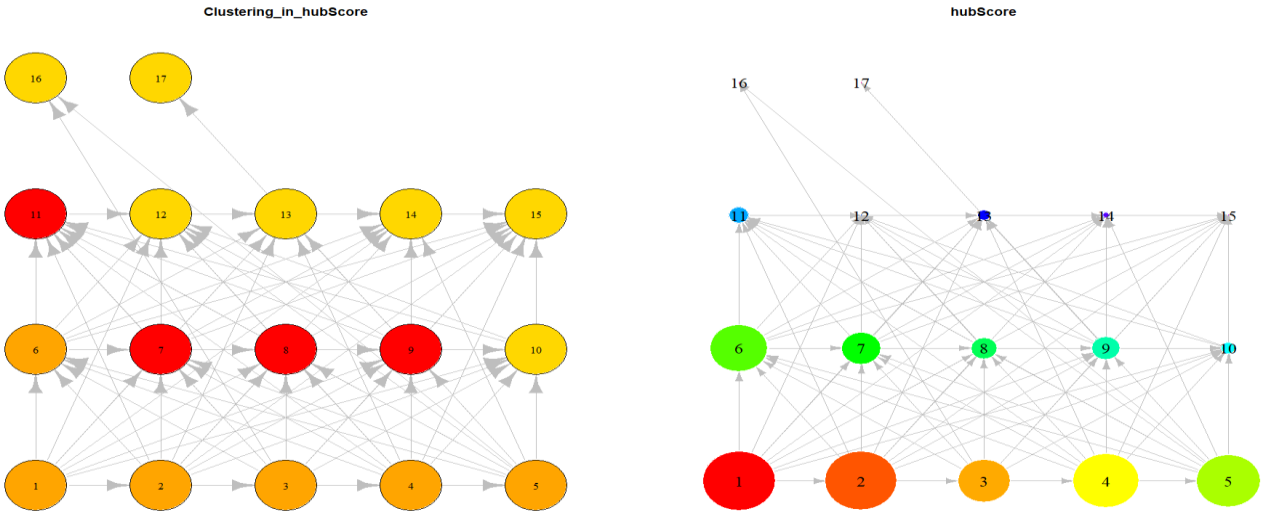

Figure 2: Results of Hub Score

An entity with higher authoritive score indicates: if an entity has a high number of relationships pointing to it has high authoritive value and acts as definitive source of information as shown in Figure 4.

Output: hubScore

1	2	3	4	5	6	7	8	9
1.0000000	0.9851625	0.7265822	0.9154720	0.8855735	0.7946946	0.5416624	0.3695520	0.3919813
10	11	12	13	14	15	16	17	
0.1952697	0.2752224	0.0000000	0.1795161	0.1008463	0.0000000	0.0000000	0.0000000	

> authSore

1	2	3	4	5	6	7	8
0.00000000	0.14693025	0.29168043	0.39843733	0.53294787	0.66306542	0.77983009	0.85941669
9	10	11	12	13	14	15	16
0.83412847	0.89172238	1.00000000	0.93368160	0.79309829	0.77903612	0.99864282	0.23687445
17							

0.05759391

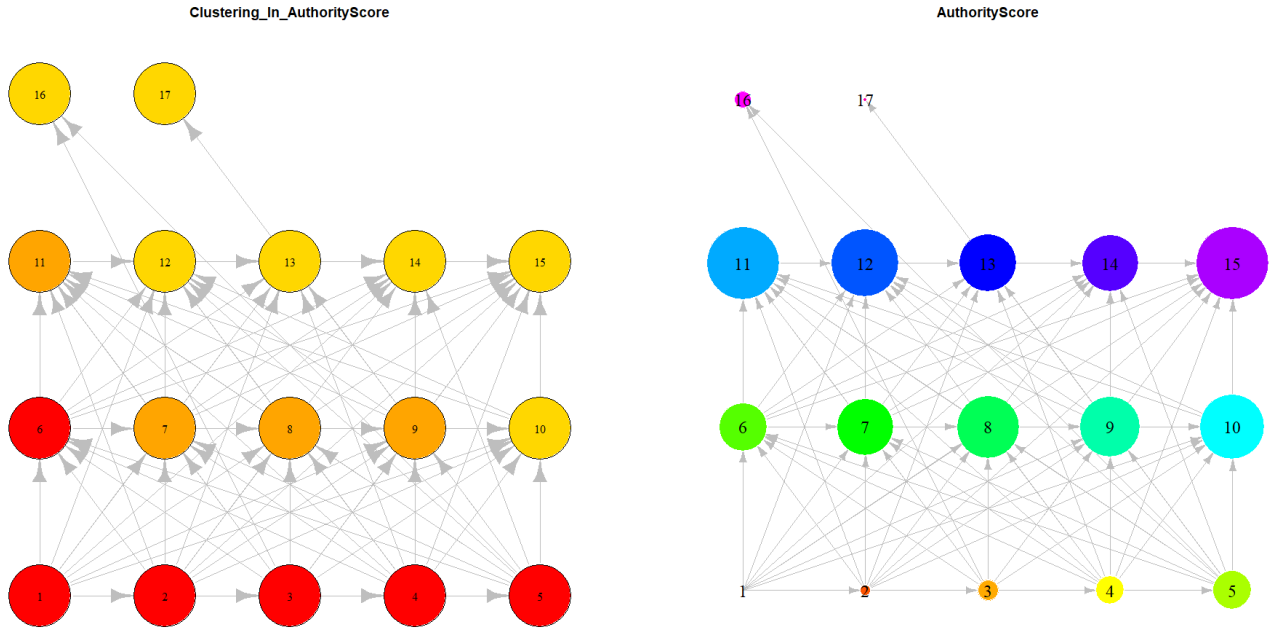

Figure 3: Results of Authority Score

- **Page Rank**

Page rank assigns a sore o find importance to each node. Important nodes are those with many in-links form important pages (nodes with many incoming links are influential).

An entity with higher page rank score indiactes*:* node with high score considered as influential. To compute the PageRank an inbuilt function is used 'page_rank' and result is shown in Figure 4.

Output:

1	2	3	4	5	6	7	8
0.02697552	0.02861331	0.03048418	0.03307534	0.03563116	0.03838448	0.04200968	0.04796105
9	10	11	12	13	14	15	16
0.05220140	0.06107564	0.09298416	0.11673852	0.07171541	0.07584894	0.17813145	0.03231999
17							
0.03584975							

- **Eigenvector Centrality (EC)**

EC is a method of computing the approximate importance of each node in a network. The intuition behind the eigenvector centrality is that a node is thought to be more important if it is directly connected to more important nodes. This metric identifies the most influential node, which is connected to other important nodes.

An entity with higher eigenvector centrality indiactes: informs about how close an entity is to other highly close entities within a network. To compute the Eigenvector Centrality an inbuilt function is used 'eigen_centrality' and result is shown in Figure 5.

Ouput:

1	2	3	4	5	6	7	8
0.98824648	0.98824648	0.84270596	0.98824648	1.00000000	0.98824648	0.86628306	0.82750130
9	10	11	12	13	14	15	16

0.82829693 0.76102329 0.94164173 0.76283258 0.67957314 0.60731216 0.80780483 0.15316160
17
0.06884619

- **Closeness Centrality (CC)**

CC helps in finding the nodes that are closest to the other nodes in a network. This centrality measures the speed at which a piece of information can reach other entities within the network from a given entity.

An entity with higher closeness centrality indicates*:* has quick access to other nodes and nodes with a high closeness value have a shorter distance to all the other nodes & are therefore considered to be efficient broadcasters of information. To compute the closeness centrality an inbuilt function is used 'closeness' and result is shown in Figure 6.

Output:

1 2 3 4 5 6 7 8
0.05555556 0.05555556 0.05000000 0.05555556 0.05882353 0.05555556 0.04761905 0.04761905
9 10 11 12 13 14 15 16
0.05000000 0.04545455 0.05263158 0.04545455 0.04166667 0.04000000 0.04761905 0.03225806
17
0.02857143

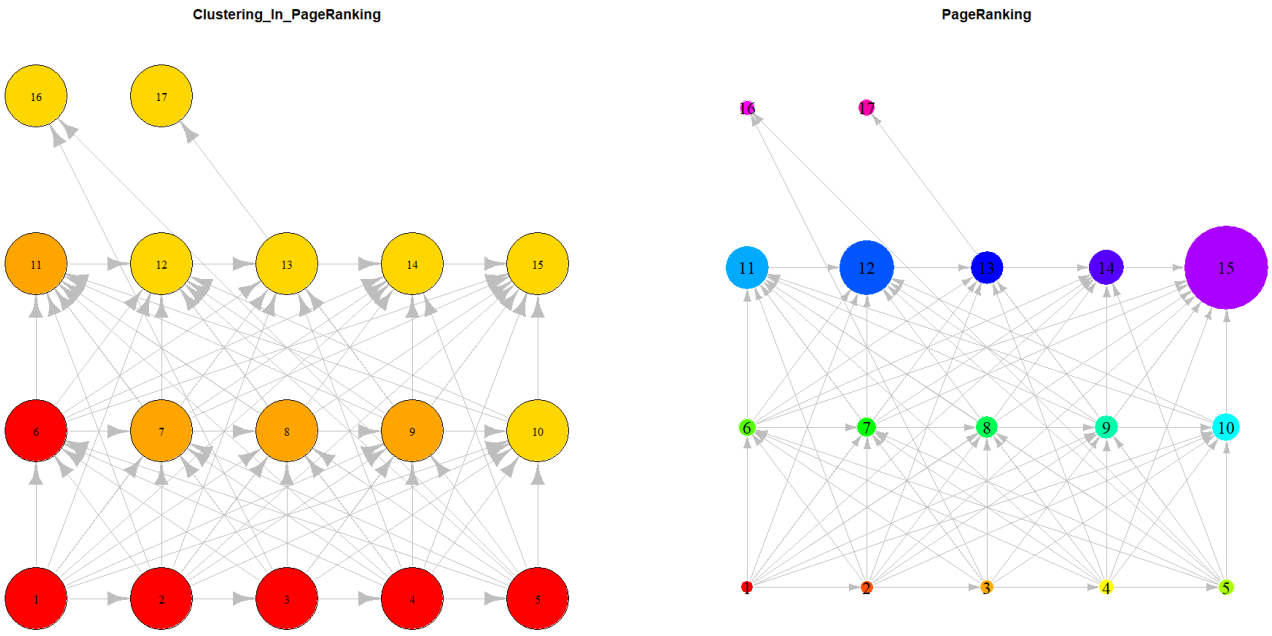

Figure 4: Page Ranking Results

- **Betweenness Centrality (BC)**

BC is to measure one node undertaking 'mediation' role in a network. If one node locates in the only way which other nodes have to go through, such as communication, connection, transportation or transaction, then this node should be important and very likely have a high betweenness centrality.

An entity with higher betweenness centrality indicates: holds a powerful position in the network and represents a single point of failure. To compute the betweenness centrality an inbuilt function is used 'betweenness' and result is shown in Figure 7.

Output:

```
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15
0.000  0.000  1.000  0.575  2.575  0.575  0.575  3.125  8.625  0.125  7.825  0.000  4.000  0.000  0.000
    16     17
0.0    0.000
```

- **Degree Centrality (DC)**

DC helps in finding the nodes with the highest number of links to other nodes within a network. Hence, it measures the popularity and/or influence factor of the entity in the network. It is often used in identifying the entities that are central with respect to spreading news and influencing other entities in the network. ***An entity with higher degree centrality indicates***: is consider influential/active player and strategically important node for communication. To compute the degree Centrality an inbuilt function is used 'centr_degree' and result is shown below.

Output:

Degreecent : 14 14 12 14 15 14 12 11 12 10 13 10 9 8 11 2 1

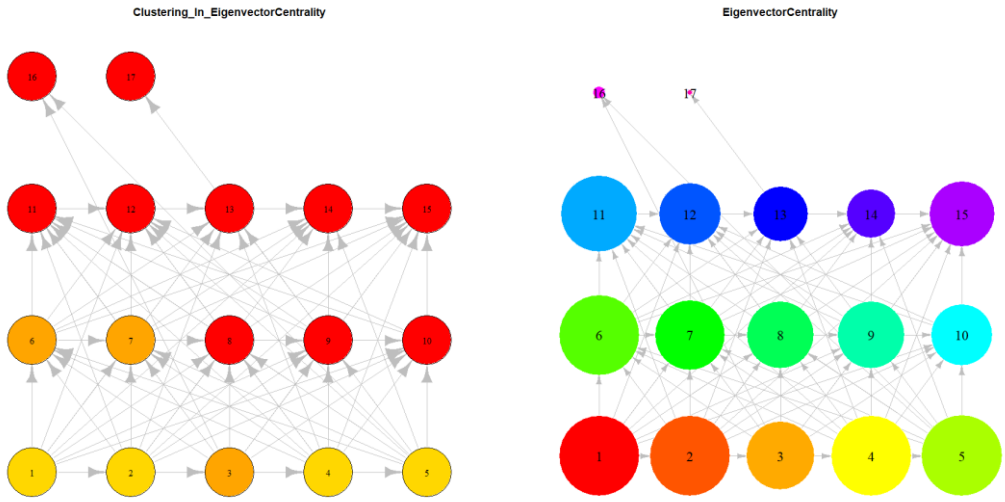

Figure 5: Eigen Vector Centrality Results

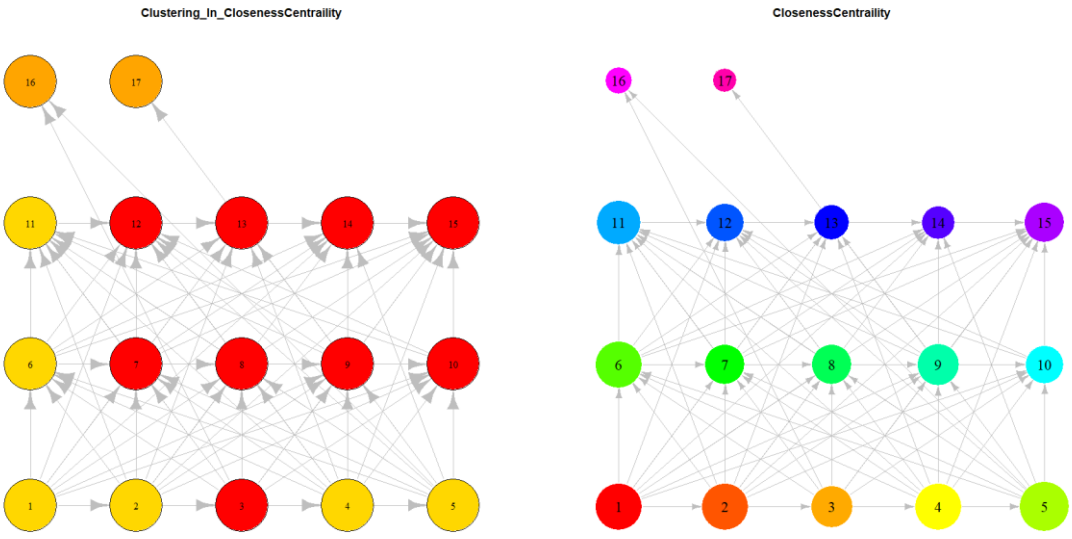

Figure 6: Closeness Centrality Results

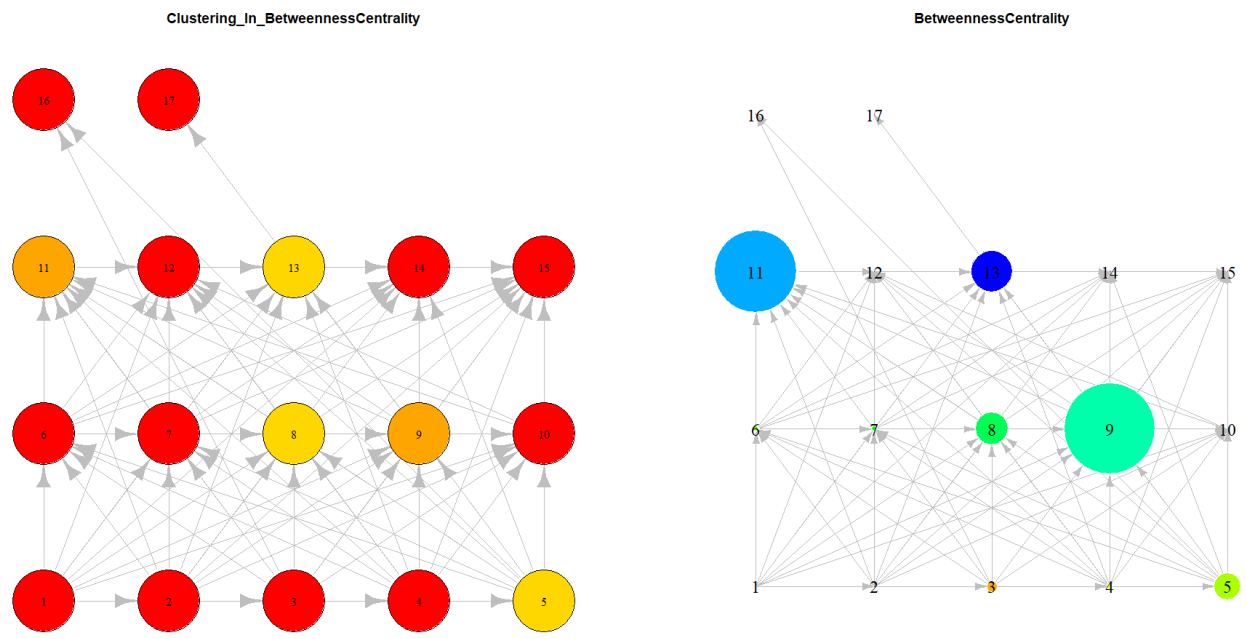

Figure 7: Betweenness Centrality Results

- **References**

- Meo, P. D., Musial-Gabrys, K., Rosaci, D., Sarne, G. M., & Aroyo, L. (2017). **Using centrality measures to predict helpfulness-based reputation in trust networks.** *ACM Transactions on Internet Technology (TOIT)*, *17*(1), 8.in Trust Networks

- Kumar Behera, R., Kumar Rath, S., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2019). **Distributed Centrality Analysis of Social Network Data Using MapReduce**. *Algorithms*, *12*(8), 161.

- Zhang, J., & Luo, Y. (2017, March**). Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network**. In *2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*. Atlantis Press.

- Social-network-analysis
- https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

- Network analysis and mnaipulaton
- http://www.sthda.com/english/articles/33-social-network-analysis/136-network-analysis-and-manipulation-using-r/

- https://kateto.net/wp-content/uploads/2016/01/NetSciX_2016_Workshop.pdf