Assignment - 01

“Web and Social Computing (IT752)”

Submitted by
Tarushi Jat (202IT029)

Submitted to
Dr. Sowmya Kamath

MASTER OF TECHNOLOGY
in
INFORMATION TECHNOLOGY

Department Of Information Technology

National Institute Of Technology, Surathkal

Task Description

Consider real-world information networks of different scale (at least 3 different standard datasets with nodes and edges in the order of thousands, tens of thousands and millions). Design an information network analysis strategy for measuring the various network properties listed below.

1. Average degree
2. Degree distribution
3. Sparseness
4. Diameter
5. Geodesic path length
6. Strongly connected components (SCC)
7. SCC properties
8. 1-connectedness to k-connectedness; what is k for your graph?
9. Clustering coefficient & average clustering coefficient

Dataset Used for the Task

Dataset 1 : Social Networks Dataset : soc-sign-bitcoin-otc

Nodes are the people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputation to prevent transactions with fraudulent and risky users. Members of Bitcoin OTC rate other members in a scale of -10 (total distrust) to +10 (total trust).
Therefore, there will be an **edge** between any 2 nodes if one user gives rating to the other user.

Number of Nodes = 5881
Number of Edges = 21492

Dataset 2 : Collaboration Networks Dataset: ca-GrQc

This dataset covers scientific collaborations between authors papers submitted to the General Relativity and Quantum Cosmology category. In this collaboration

Network, **nodes** are the authors in the network and If an author *i* co-authored a paper with author *j*, the graph contains an undirected **edge** from *i* to *j*.

Number of Nodes = 5242
Number of Edges = 14496

Dataset 3 : Communication Network : email-Eu-core

This dataset contains the information of incoming and outgoing email between members of a European research institution. There is an **edge** (u, v) in the network if person u sent person v at least one email.

Number of Nodes = 1005
Number of Edges = 16706

Information Network Analysis

1. Average Degree

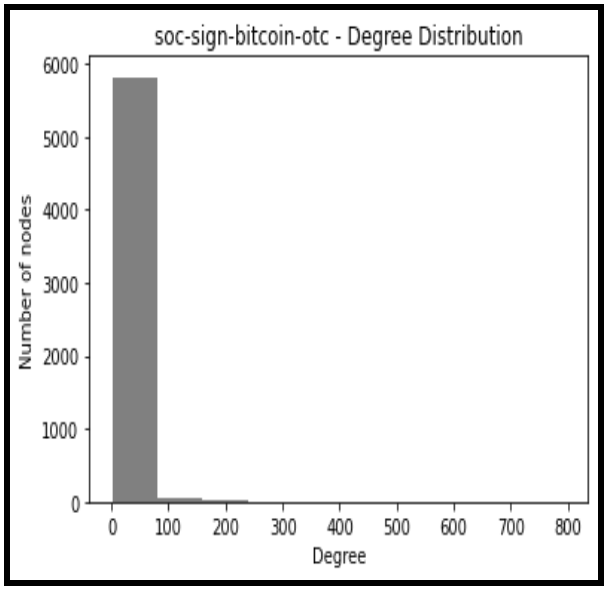
Average Degree is the average number of edges per node in a graph. It can be analyzed that communication network dataset - email-Eu-core (Dataset 3) has the highest average degree with a value of **33.2458**. On the other hand, collaboration network dataset - ca-GrQc (Dataset 2) has the lowest average degree of **5.5307**. And the social network dataset of bitcoin-otc has an average degree of **7.3090**. It can be concluded that most interaction happens in email communication network where edge represents the email sent from one person to the other.

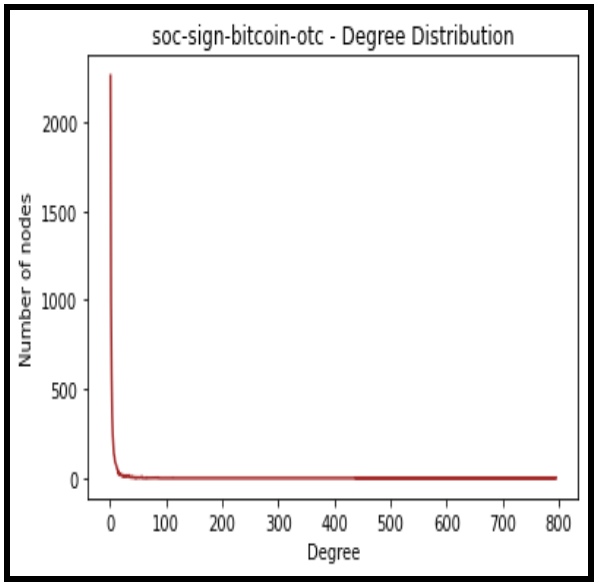
Average Degree		
Dataset 1	soc-sign-bitcoin-otc	7.3090
Dataset 2	ca-GrQc	5.5307
Dataset 3	email-Eu-core	33.2458

2. Degree Distribution

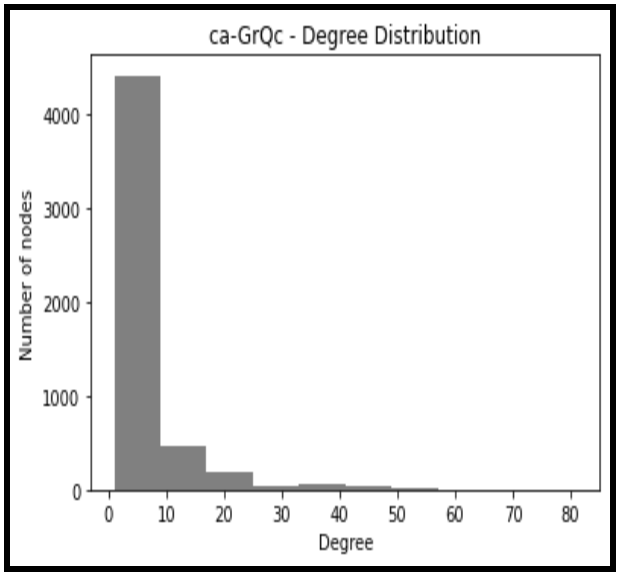
The degree of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network. Below are the plots for degree distribution for each of the 3 dataset.

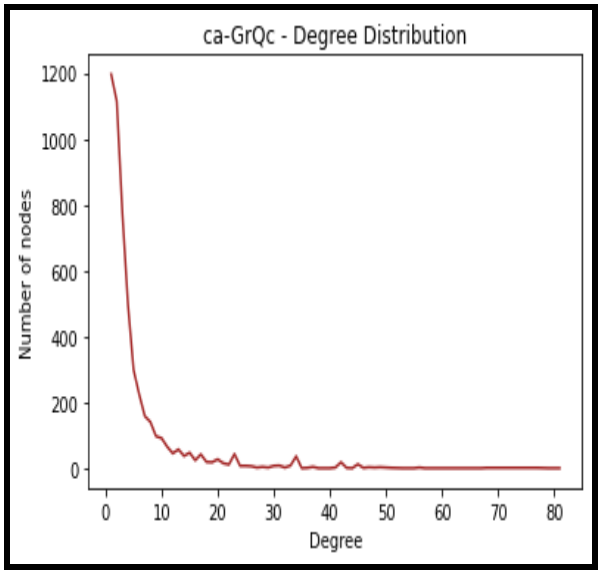
Dataset 1 : soc-sign-bitcoin-otc



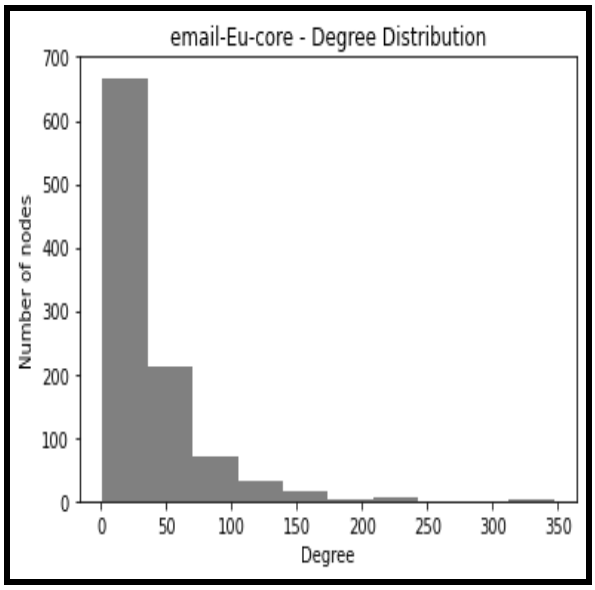


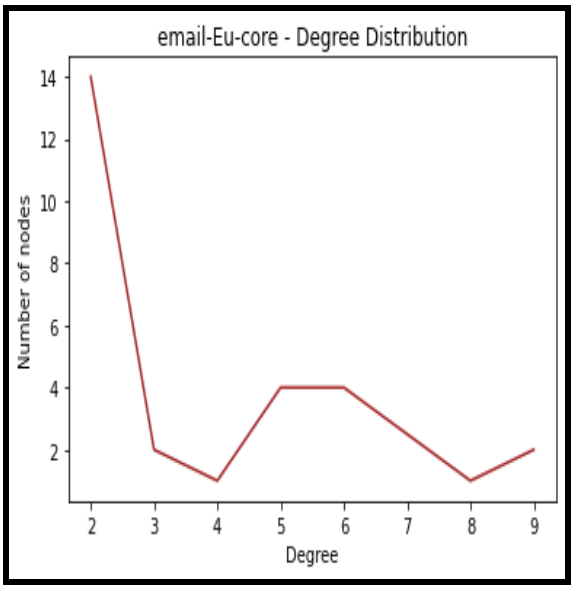
Dataset 2 : ca-GrQc





Dataset 3 : email-Eu-core





3. Sparseness

Sparseness means the network graph has much fewer links than the possible maximum number of links within the network. A dense network contains the maximum number of edges. By finding the edge density of a network graph we can know whether the graph is sparse or dense. The value of edge density varies between 0 and 1. If the value of edge density is more toward 0, then it is a sparse network and if the value of edge density is more towards 1, then it is a dense network.

In the experiment with 3 network datasets, social network dataset “soc-sign-bitcoin-otc” [Dataset -1] have sparseness value of **0.001243020588612932**, collaboration network dataset “ca-GrQc” [Dataset -1] have sparseness value of **0.001055278280507905** and communication network dataset “email-Eu-core” [Dataset - 3] have the sparseness value of **0.03311331787278746**.

Sparseness		
Dataset 1	soc-sign-bitcoin-otc	0.0012430205886129 32
Dataset 2	ca-GrQc	0.0010552782805079 05
Dataset 3	email-Eu-core	0.0331133178727874 6

4. Diameter

Once the shortest path length from every node to all other nodes in a network graph is calculated, the diameter is the longest of all the calculated path lengths. It is the shortest distance between the two most distant nodes in the network.

Since all of the 3 datasets that I have used in this assignment contains multiple components, thus I have calculated the diameter for the largest component present in the network. For soc-sign-bitcoin-otc dataset, diameter for the largest component is **9.** For ca-GrQc dataset, diameter for the largest component is **17** and for email-Eu-core dataset, diameter for the largest component is **7**.

Diameter [For the largest component present in the network]		
Dataset 1	soc-sign-bitcoin-otc	9
Dataset 2	ca-GrQc	17
Dataset 3	email-Eu-core	7

5. Geodesic Path Length

A geodesic path length between any two nodes in a network graph is a path with the minimum number of edges. There exist various edges between different nodes and geodesic path length can be calculated between any two given pair of nodes.

Thus to calculate geodesic path length for the 3 datasets that are used in this assignment, I have randomly picked 2 nodes in each dataset to find geodesic path length between the two selected nodes. In **soc-sign-bitcoin-otc** dataset, the source ID = 6 and destination ID = 371 and geodesic path length between the two nodes is **2** and average geodesic path length of the entire graph is **2.3433333333333333**. In **ca-GrQc** dataset, the source ID = 0 and destination ID = 796 and geodesic path length between the two nodes is **3** and average geodesic path length of the entire graph is **2.4444444444444446**. In **email-Eu-core** dataset, the source ID = 3466 and destination ID = 17038 and geodesic path length between the two nodes is **1** and average geodesic path length of the entire graph is **2.47008547008547**. This analysis if formulated in below table:-

Geodesic Path Length					
S.No.	**Dataset name**	**Src ID**	**Dstn ID**	**Geodesic Path Length**	**Average Geodesic Path Length**
Dataset 1	soc-sign-bitcoin-otc	6	371	2	2.3433333333333333
Dataset 2	ca-GrQc	0	796	3	2.4444444444444446
Dataset 3	email-Eu-core	3466	17038	1	2.47008547008547

6. Strongly connected components (SCC)

A graph is said to be strongly connected if there is a path in each direction between each pair of vertices of the graph. A pair of vertices *u* and *v* are said to be strongly connected to each other if there is a path in each direction between them. The details of the number of strongly connected component with respect to each of the 3 datasets that I have used is formulated in the table below:

Strongly Connected Components		
Dataset 1	soc-sign-bitcoin-otc	1144
Dataset 2	ca-GrQc	355
Dataset 3	email-Eu-core	203

7. SCC properties

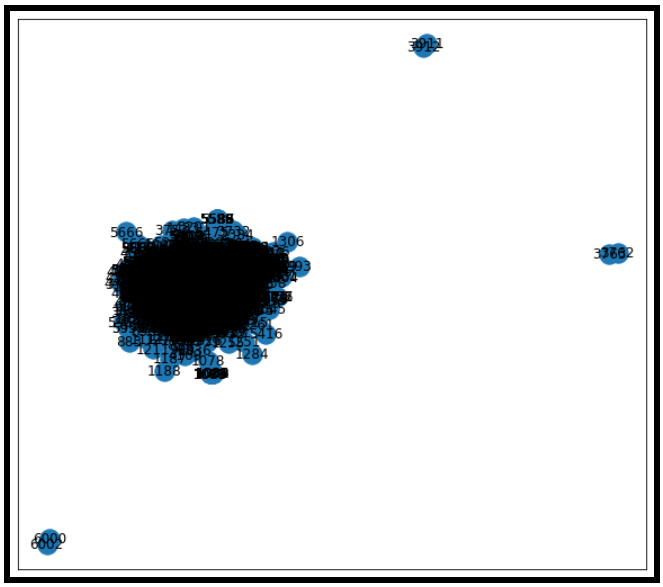
A directed graph is strongly connected if there is a path between all pairs of vertices. A strongly connected component of a directed graph is a maximal strongly connected subgraph. Dataset -1 “soc-sign-bitcoin-otc” has a maximum number of strongly connected components as compared to the other 2 datasets. Number of connected components in the dataset - 1 is 4 with a total of 5881 nodes. Number of connected components in the dataset - 2 is 355 with a total of 5242 nodes. Number of connected components in the dataset - 3 is 20 with a total of 1005 nodes.

8. 1-connectedness to k-connectedness; what is k for your graph?

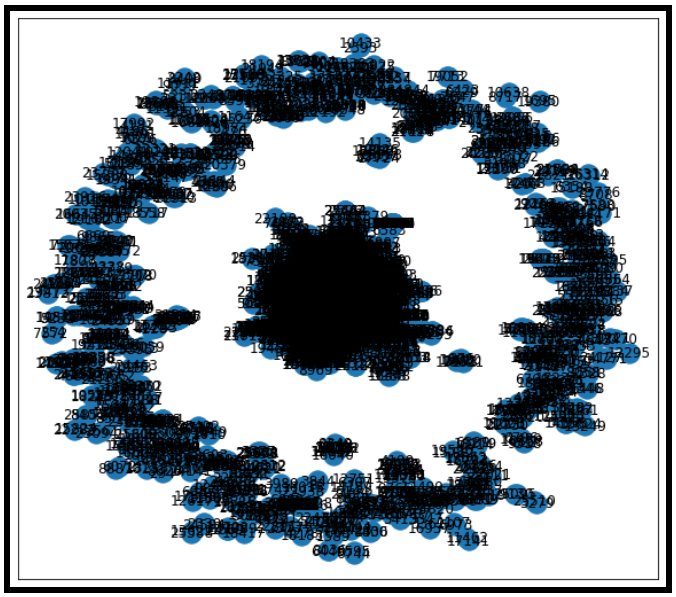
A graph G on more than two vertices is said to be K-connected if there does not exist a vertex cut of size k-1 whose removal disconnects the graph. The complete graph with n vertices has connectivity $n - 1$. All the three network graph datasets that I have used in this assignment are not connected graphs. The Graphs of all the 3 datasets are already **disconnected**. “soc-sign-bitcoin-otc” dataset has **4** connected components, “ca-GrQc” dataset has **355** connected components and “email-Eu-core” dataset has **20** connected components.

Visualization of connected components

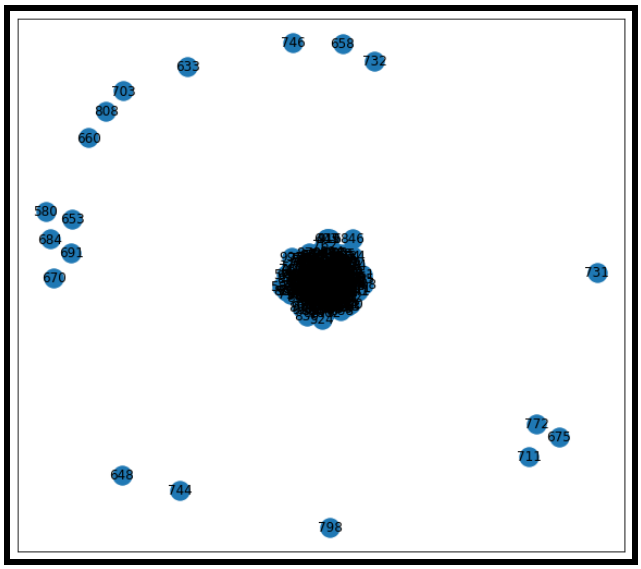
Dataset 1: Social Networks : soc-sign-bitcoin-otc



Dataset 2 : Collaboration Networks : ca-GrQc



Dataset 3 : Communication Networks : email-Eu-core



9. Clustering coefficient & average clustering coefficient

The clustering coefficient of a node is the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of such links. The clustering coefficient for the entire network is the average of the clustering coefficients of all the nodes. A high clustering coefficient for a network is another indication of a small world network. Clustering coefficient of a node is calculated as:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

For the 3 datasets that I have used in this assignment, I have calculated clustering coefficient for each node in all 3 datasets and calculated average clustering coefficient for each of the 3 network datasets. Below are the details formulated in table, also output of clustering coefficient of few nodes are shown in the table.

Clustering coefficient & average clustering coefficient			
S.No.	Dataset Name	Avg Clustering Coefficient	Clustering coefficient of few nodes
Dataset 1	soc-sign-bitcoin-otc	0.1775044940528951	('936', 0) ('937', 0.05263157894736842) ('885', 0.06666666666666667) ('938', 0) ('923', 0.06060606060606061)
Dataset 2	ca-GrQc	0.529635811052136	('2654', 0.2503556187766714) ('4748', 0.8245614035087719) ('5672', 1.0) ('10549',0.6666666666666666) ('12928', 1.0)
Dataset 3	email-Eu-core	0.3993549664221539	('0', 0.2764227642276423) ('1', 0.2653061224489796) ('2', 0.2978027115474521) ('3', 0.38491048593350385) ('4', 0.31869137497140243)