Web and Social Computing (IT752)
Lab Assignment 1

Submitted by: Sampat Kr Ghosh
Roll number: 192IT020

Dataset 1: Twitter
Dataset 2: Facebook
Dataset 3: Gnutella peer-to-peer network, August 4 2002

The above 3 datasets were used and I concluded various factors like:

1. Degree Distribution
2. Diameter
3. Geodesic path length
4. Clustering Coefficient
5. Strongly Connected Components
6. Sparseness
7. k-connectedness

1. Degree Distribution

Finding degree distribution for all the 3 datasets are plotted as shown:

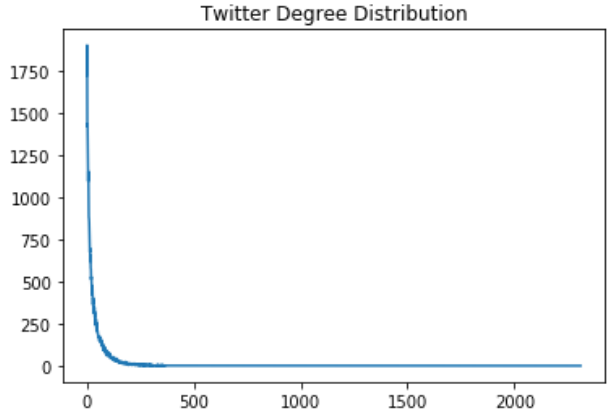

Fig 1: Degree distribution of dataset 1

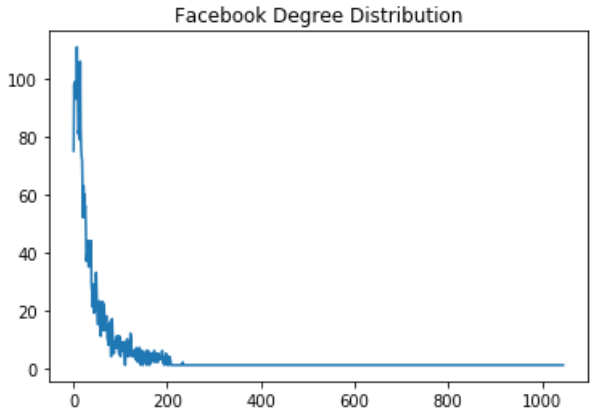

Fig2: Degree distribution of dataset 2

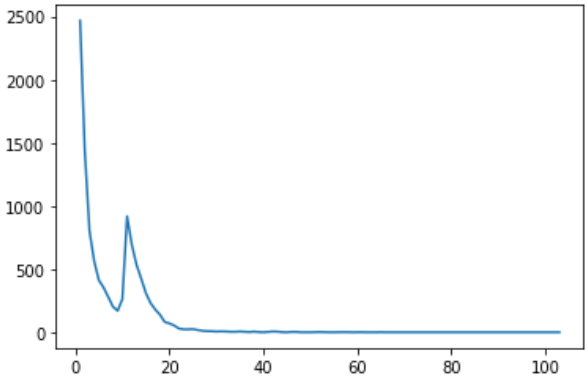

Fig3: Degree distribution of dataset 3

From the above degree distributions, we can easily say that the dataset 3 has very few nodes with very high degree and a large amount of nodes have very small degree.

2. Diameter

The maximum eccentricity from all the vertices is considered as the *diameter* of the *Graph* G or it can be defined as the maximum shortest path between any pair of nodes. For dataset 1, I got an infinite diameter, for dataset 2, I got diameter as 8 and for dataset 3 I got 10. As the path length between some nodes of dataset 1 may be infinite, that is there is no path from some vertex to another.

3. Geodesic path length

It is defined as the number of edges along the shortest path connecting a given pair of nodes. As there are various edges between different nodes, geodesic path length can be calculated for different vertices. I have randomly taken 2 nodes to find the geodesic path length. For dataset 1, source node id is 19933035 and destination node id is 113420831 and geodesic path length between them is 3 and the average geodesic path length for the whole graph could not be calculated since it was disconnected. For dataset 2, source node id is 0 and destination node id is 4038 and geodesic path length between them is 5 and the average geodesic path length for the whole graph is 3.6925068496963913. For dataset 3, source node id is 0 and destination node id is 721 and geodesic path length between them is 5 and the average geodesic path length for the whole graph is 2.693598381757999.

4. Clustering Coefficient

It is a measure of how connected node i's neighbors are connected to each other. Clustering coefficient is found for various node. I have taken few outputs for various nodes.

<u>For dataset 1:</u>
(214328887: 0.36926611561517947)
(34428380: 0.03861899964317948)
(17116707: 0.37572393822393824)
(28465635: 0.09665891308944809)
<u>For dataset 2:</u>
(0: 0.04196165314587463)
(1: 0.41911764705882354)
(2: 0.8888888888888888)
(3: 0.6323529411764706)
<u>For dataset 3:</u>
(0: 0.011029411764705883)
(1: 0.005494505494505495)
(2: 0.013888888888888888)
(3: 0.004166666666666667)
Average clustering coefficient for dataset 1 is 0.3850872080950751, for dataset 2 is 0.6055467186200876 and for dataset 3 is 0.0031087663857330312.

5. Strongly connected components

A graph is said to be strongly connected or disconnected if every vertex is reachable from every other vertex. For dataset 1, number of strongly connected components are 6503, for dataset 2 is 1 and for dataset 3 is 6560.

6. Sparseness

A graph with only a few edges is a sparse graph. By finding the edge density of the graph we can how much the graph is dense. Edge density lies between 0 and 1, going more towards 0 is sparse and going more towards 1 is a dense graph. Edge density of dataset 1 graph is 0.0005210070796435007, for dataset 2 is 0.010819963503439287 and dataset 3 is 0.0003381398671756435 Here, dataset 3 is the sparsest graph, dataset 1 is the sparser one and dataset 2 is the least sparse graph.

7. Finding K for K-Connectedness graph

A K-component is a maximal subgraph of a graph G that has, at least, node connectivity K, we need to remove at least K nodes to break it into more components. For dataset 1 value of K is 24 and for dataset 2 and dataset 3 is 2. As dataset is having only 1 component, the diameter was 10 and for other dataset 1 and 2, the graph is disconnected due to which the value of diameter is showing infinite.