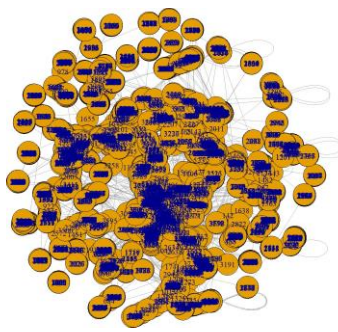
Tutorial 1

1. Original Graph and Induced Subgraph

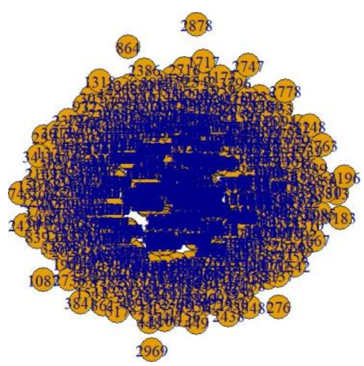
a) Refer to the R programming file Plotting of the original twitter graph



b) Refer to the R programming file

sample_size	253
Vsample	int [1:253] 415 463 179 14 195 426 306 118 299 229 ...

Plot of the induced subgraph:



c)

Graph density

```
> print(density1)
[1] 0.002279749
>
> density2 <- graph.density(networkA_sample)
> print(density2)
[1] 0.002257878
```

Average path length

```
> cat(avg_path_length1, "\n")
6.275911
> cat(avg_path_length2, "\n")
6.810667
```

clustering coefficient

```
> cat("Clustering Coefficient - Original Graph:", clust_coeff1, "\n")
Clustering Coefficient - Original Graph: 0.4433374
> cat("Clustering Coefficient - Induced Subgraph:", clust_coeff2, "\n")
Clustering Coefficient - Induced Subgraph: 0.4636895
```

d)

It is observed the graph density and clustering coefficient roughly remains the same. The similarity in graph density indicates that the induced subgraph preserves the original graph's edge-to-node ratio. This suggests that the processof creating the subgraph has managed to maintain the overall connectivity level of the original graph.

The slight increase in average path length in the induced subgraph suggests thatthe process of inducing the subgraph has led to a network where, on average, nodes are farther apart.

2. measures and correlation

a)

- **Degree:** The number of edges connected to a node is referred to as the degree of a node. Degrees can help in targeting influencers and key opinionleaders for product endorsements.
- **Closeness:** A measure of how close a node is to all other nodes in the network, calculated as the reciprocal of the sum of the shortest path lengthsfrom the node to all other nodes. Closeness of nodes in a transportation network helps identify optimal locations for warehouses or distribution centersto minimize transportation costs
- **Clustering Coefficient:** A measure of the degree to which nodes in a graph tend to cluster together. This insight can improve the accuracy of personalizedrecommendations by suggesting products bought or liked by similar community.

PageRank: An algorithm used to rank nodes in a network based on the number and quality of links to a node, indicating the node's importance orinfluence within the network. Can help us understand the importance of different websites.

- **Eccentricity:** The maximum distance from a node to all other nodes in the network, with distance measured as the shortest path between nodes. It reflects the furthest a node is from any other node in the graph. In telecommunication networks, eccentricity can identify nodes that are farthestfrom others, highlighting areas with potential service delays or lower quality

b) Refer to the R programming filec)

The Pearson correlation of 5 measures:

The correlation between degree and PageRank implies that nodes with a higher number of connections (higher degree) tend to also have higher PageRank values.This indicates that in the network, simply having more connections contributes significantly to a node's perceived importance.

The high negative correlation between closeness and eccentricity indicates that anode on an average is closer to all other nodes. Therefore the information flow within this network is efficient as the distance between nodes are less.

3. Using a random graph for comparison

a) Refer to R programming file

```
> summary(network_random)
IGRAPH 9f5bfe4 U--- 3892 17262 -- Erdos-Renyi (gnm) graph
+ attr: name (g/c), type (g/c), loops (g/l), m (g/n)
```

```
        Wilcoxon rank sum test with continuity correction

data:  clustering_coefficient_networkA and clustering_random
W = 3886, p-value = 0.0006337
alternative hypothesis: true location shift is not equal to 0
```

The random network is assumed to have loops and is undirected

After performing the Mann-Whitney U test , we observe the alternate hypothesis to be true. This indicates there is a difference between the two groups.

Specifically, it suggests that the median of one group is not equal to the median of the other group. This proves that the original graph and the random graph created are structurally different.

c) Refer to R programming file

```
> wilcox.test(degree_networkA, degree_random, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  degree_networkA and degree_random
W = 4752300, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(closeness_networkA, closeness_random, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  closeness_networkA and closeness_random
W = 36581, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> #wilcox.test(clustering_coefficient_networkA, clustering_random, alternative = "two.sided")
> wilcox.test(pagerank_networkA, pagerank_random, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  pagerank_networkA and pagerank_random
W = 6314644, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(eccentricity_networkA, eccentricity_random, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  eccentricity_networkA and eccentricity_random
W = 15147664, p-value < 2.2e-16
```

A similar trend can be seen throughout all 5 node level measures of the two networks. For all the measures, the alternate hypothesis turned out to be true, indicating our random network is structurally different from the original network significantly.

d)Two noticeable things that come in mind while comparing the two networks, are the values of cluster coefficient and average path length. The original network Ahas a higher cluster coefficient and a larger average path length. This says a lot about the local and global structure between two networks. Locally, network hastight clusters of nodes, suggesting strong small-scale interactions (e.g., within communities or groups). However, globally, the connectivity is weaker, as indicated by the longer paths needed to connect different parts of the network. This shows the random network generated does not produce the tight node bondsto similar extent of network A, but the larger network is more interconnected as the average path length is lesser.

This comparison highlights the importance of considering both local and global properties when analyzing or designing networks, as they can offer insights intothe network's functionality, efficiency, and resilience.