CS 6350-002
Assignment-3
Part-2 Report

Analyzing Social Networks using GraphX

In this project, we used Spark GraphX to analyze social network data. The 'ego-Twitter' dataset was chosen from the SNAP repository: https://snap.stanford.edu/data/index.html#socnets

Details about the dataset:

The edges in the graph are directed. There are 81,306 nodes and 1,768,149 edges in the dataset which represent social circles from Twitter.

Analysis details

The following analysis was done on the twitter data:

a. To find the top 5 nodes with the highest outdegree and find the count of the number of outgoing edges in each

Analysis: Node with the highest out-degree is 3359851. This suggests that the node is following as many other Twitter identities as possible.

b. To find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each

Analysis: Node with the highest out-degree is 3359851. This suggests that the node is following as many other Twitter identities as possible.

c. To calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values.\

Analysis: The node with the highest pagerank is 115485051. This suggests that this profile is quite significant and that its tweets have a significant impact.

d. To run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

Analysis: We can state that the whole graph is connected because there is just one component. The graph contains no independent components. This suggests that every user on Twitter has a connection to every other user through one or more nodes.

e. To run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, randomly select the top 5 vertices.

Analysis: With a maximum triangle count of 774424, node 40981798 suggests that a large number of users share this user in their mutually connected networks with three nodes.