Assignment 3 Part 2

Analyzing Social Networks using GraphX/GraphFrame

**Problem Statement**

**In this part, you will use Spark GraphX/GraphFrame to analyze social network data. You are free to choose any one of the social network datasets available from the SNAP repository.**

**You will use this dataset to construct a GraphX/GraphFrame graph and run some queries and algorithms on the graph.**

**Solution:** [Colab Notebook](https://colab.research.google.com/drive/1yP-7jeaYKonvGEgznyWzRck4rDjr39nS?usp=sharing)

**Output of Queries**

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

A black background with white text

Description automatically generated

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

A black and white image of numbers

Description automatically generated

1. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters.

A black and white screen with numbers and letters

Description automatically generated

1. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

*Connected Components: Strongly Connected Components:*

A black and white screen with numbers and a white text

Description automatically generated with medium confidence A black and white screen with white text

Description automatically generated

1. 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, you can randomly select the top 5 vertices.

A black background with white text

Description automatically generated

**Summary:**

1. **The indegree signifies the number of votes received.**
2. **The outdegree signifies the number of votes given by the person.**
3. **A higher PageRank indicates a higher level of importance. This is based on the idea that ids that are linked to by many other votes are likely to be more important.**
4. **The connected components signify the voting groups, ie. people generally reach vote within the same set of ids.**
5. **Triangle count suggests that 2 ids have cast votes for the same id.**