Assignment 3 Part 2

Analyzing Social Networks using GraphX/GraphFrame

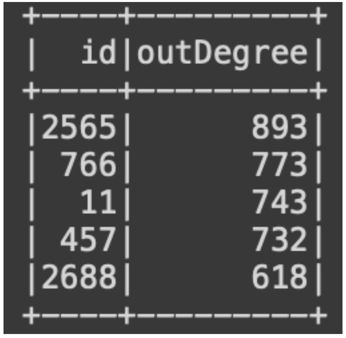
Problem Statement

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]()

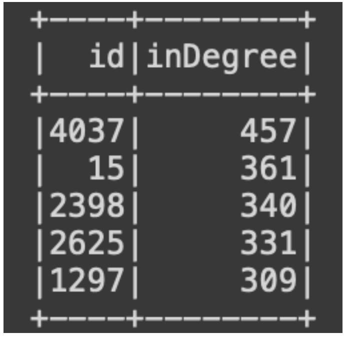
Output of Queries

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



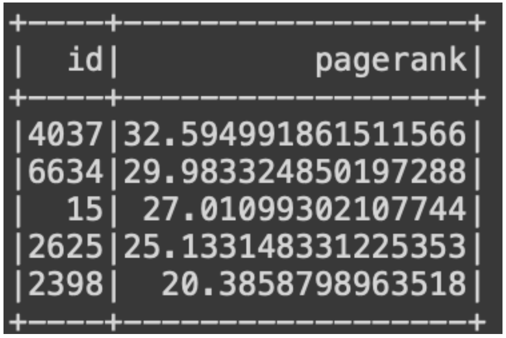
```
+----+---------+
|  id|outDegree|
+----+---------+
|2565|      893|
| 766|      773|
|  11|      743|
| 457|      732|
|2688|      618|
+----+---------+
```

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



```
+----+--------+
|  id|inDegree|
+----+--------+
|4037|     457|
|  15|     361|
|2398|     340|
|2625|     331|
|1297|     309|
+----+--------+
```

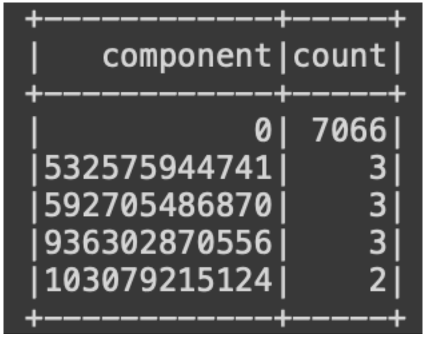
3. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters.



```
+----+------------------+
|  id|          pagerank|
+----+------------------+
|4037|32.594991861511566|
|6634|29.983324850197288|
|  15| 27.01099302107744|
|2625|25.133148331225353|
|2398|  20.3858798963518|
+----+------------------+
```

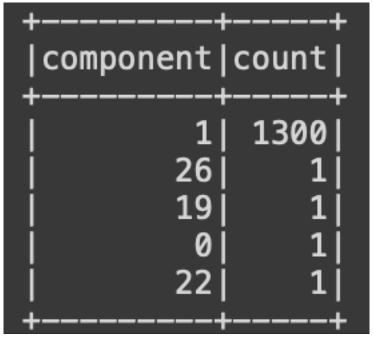
4. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

Connected Components:



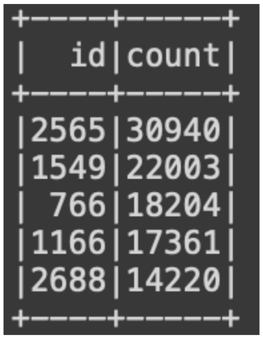
```
+------------+-----+
|   component|count|
+------------+-----+
|           0| 7066|
|532575944741|    3|
|592705486870|    3|
|936302870556|    3|
|103079215124|    2|
+------------+-----+
```

Strongly Connected Components:



```
+---------+-----+
|component|count|
+---------+-----+
|        1| 1300|
|       26|    1|
|       19|    1|
|        0|    1|
|       22|    1|
+---------+-----+
```

5. 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.



```
+----+-----+
|  id|count|
+----+-----+
|2565|30940|
|1549|22003|
| 766|18204|
|1166|17361|
|2688|14220|
+----+-----+
```

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.