Facebook Graph Analysis

*Note: Graph Analysis of Facebook Network using GraphX Apache Spark

1st Gurvinder Singh Yadav *20BDS025*
2nd Ankit Kumar *20BDS009*
3rd Ripudaman N Singh *20BDS043*
4th Navneet Sen *20DBS037*
5th Gudibanda Karthik *20BDS023*

***Abstract*—Due to increasing number of users on social media platforms it becomes a necessary endevour for social scientist to analyse these evolving networks. Due to the vast amount of data it becomes almost impossible to use vertical scale up systems to do any useful work, thats where Apache Spark GraphX API comes in. In this report we analyse Facebook network using GraphX and present the findings.**

***Index Terms*—Apache Spark, GraphX, Connected Components, PageRank**

I. Dataset

In the given study Dataset [1] from snap Stanford is utilised. This dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using this Facebook app. The dataset includes node features (profiles), circles, and ego networks.

Feature	Value
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

TABLE I
Dataset Statistics

Given Dataset is in the form an edge list. The methodolgy followed in reading an preprocessing data is as follows:

- Created a local Spark stand-alone cluster.
- Read edgelist using GraphLoader API.
- Created a Directed Graph using Edgelist.

Following Dataset has been extensively studied by the community and acts a good benchmark Dataset hence is choosen for analysis. Dataset is placed in the Dataset folder in the parent directory.

II. Algorithm

The following Algorithms have been used:

A. PageRank [2]

PageRank has been utilised to find important websites which have the most chance of being clicked by any person on internet. Using the same concept here we try to find the most important users in the facebook network who have the most chance of their profile being visited by any other person.

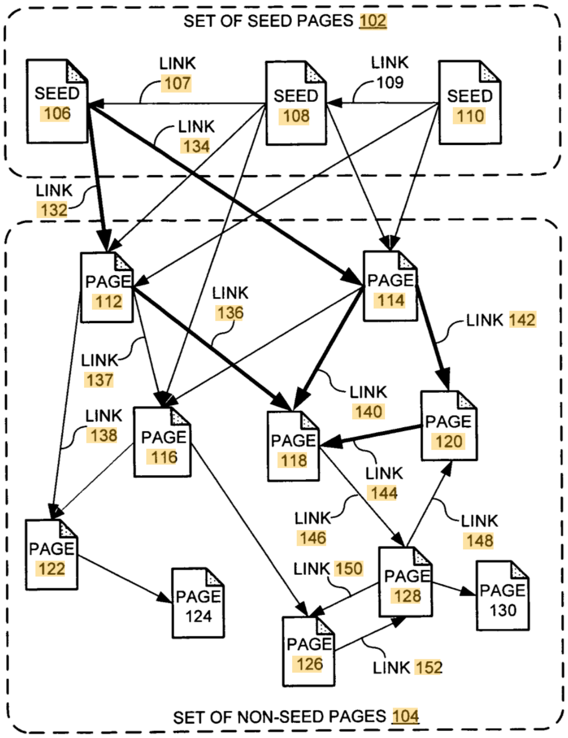


Fig. 1. PageRank

B. Connected Components [3]

We utilised Connected Components algorithm in GraphX to see all the independent subnetworks of users in the network. To see the disjoint isolated users in the network. Here we tried to check if there are users in the network who are isolated are connected amongst a smaller subnetwork.

Identify applicable funding agency here. If none, delete this.

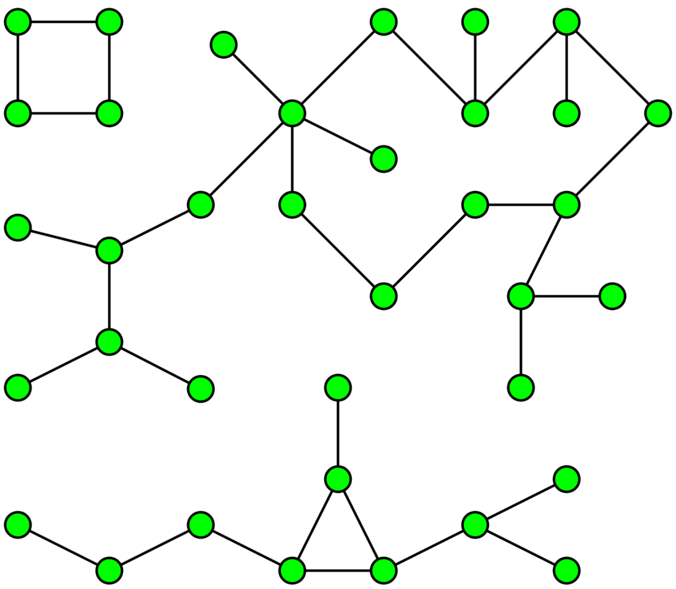


Fig. 2. Connected Components

C. Triangle Count [4]

Triangle count has gained popularity in social network analysis, where it is used to detect communities and measure the cohesiveness of those communities. It can also be used to determine the stability of a graph, and is often used as part of the computation of network indices, such as clustering coefficients.

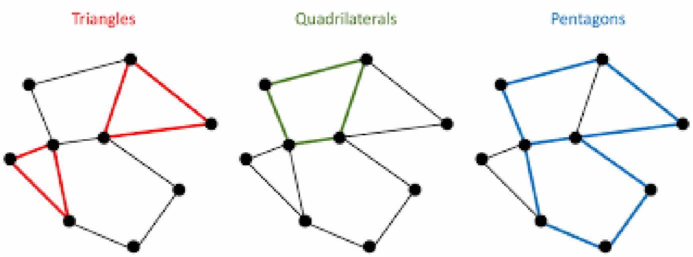
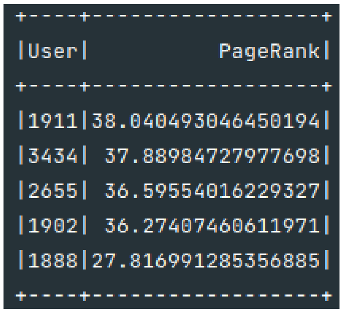


Fig. 3. Triangle count

III. Results

A. PageRank

After running the pagerank algorithm we saved the results in the results directory and made a table of the top 5 page ranks. Seeing fig-4 we can clearly see the top 5 user profiles on facebook who have the highest likelihood of being visited by any new user on facebook.

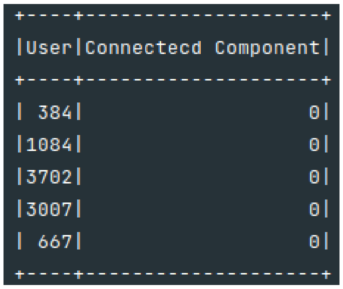


```
+----+------------------+
|User|          PageRank|
+----+------------------+
|1911|38.040493046450194|
|3434| 37.88984727977698|
|2655| 36.59554016229327|
|1902| 36.27407460611971|
|1888|27.816991285356885|
+----+------------------+
```

Fig. 4. PageRank.show(5)

B. Connected Components

Seeing fig-5 it can be clearly seen that the network is highly connected and there is no user in the network that is isolated. Hence there is just one connected component.

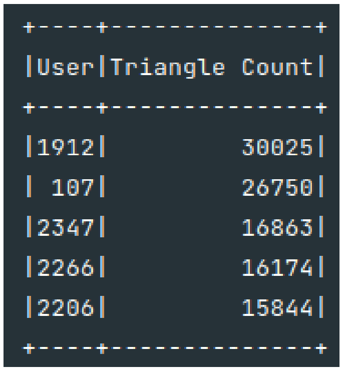


```
+----+--------------------+
|User|Connectecd Component|
+----+--------------------+
| 384|                   0|
|1084|                   0|
|3702|                   0|
|3007|                   0|
| 667|                   0|
+----+--------------------+
```

Fig. 5. ConnectedComponents.show(5)

C. Triangle Count

Seeing fig-6 it is evident that node have a very connected and cohesive neighbourhood. Hence a tightly knit social network.



```
+----+--------------+
|User|Triangle Count|
+----+--------------+
|1912|         30025|
| 107|         26750|
|2347|         16863|
|2266|         16174|
|2206|         15844|
+----+--------------+
```

Fig. 6. TriangleCount.show(5)

References

[1] @articleleskovec2012learning, title=Learning to discover social circles in ego networks, author=Leskovec, Jure and Mcauley, Julian, journal=Advances in neural information processing systems, volume=25, year=2012

[2] https://spark.apache.org/docs/3.0.2/api/scala/org/apache/spark/graphx/lib/PageRank$.html

[3] https://spark.apache.org/docs/latest/api/scala/org/apache/spark/graphx/lib/ConnectedComponents$.html

[4] https://spark.apache.org/docs/3.1.3/api/java/org/apache/spark/graphx/lib/TriangleCount.html