Graph-Based Social Network Analysis: Insights from Facebook's Dataset Using PySpark and GraphFrames

Amber Hasan, Mikiyas Midru

Executive Masters, The University of Texas at Dallas

***Abstract*— This report explores social network analysis techniques applied to a Facebook combined graph dataset using PySpark and GraphFrames. Key graph algorithms, including degree analysis, PageRank, connected components, and triangle counting, were utilized to uncover insights into network structure, node influence, and clustering. The analysis identified the most influential and connected nodes, highlighted significant community structures, and measured local clustering.**

I. Introduction

Social network analysis is crucial for understanding interactions and relationships within a community. In this study, we analyzed a Facebook combined graph dataset using PySpark and GraphFrames to extract meaningful insights about user connectivity, influence, and clustering. This report details the methods, techniques, and results of analyzing nodes and their relationships within the dataset.

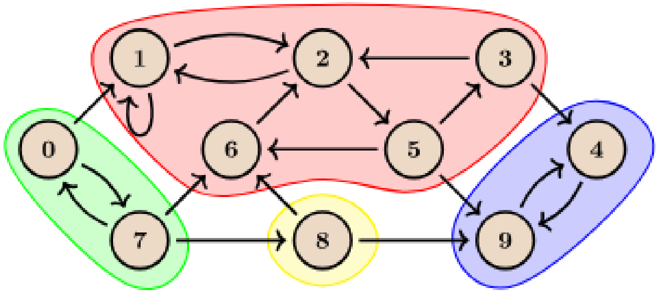


Fig. 1 This is an example of finding connected components. .

II. Background Work

Graph-based representations of data are fundamental in analyzing social networks. Techniques such as PageRank, indegree/outdegree computation, connected components, and triangle counting have been widely used to measure node influence, identify community structures, and evaluate clustering. These methods, developed for applications like web ranking and social media analytics, are employed in this study to analyze Facebook's network graph. GraphFrames, an extension of Apache Spark, provides a framework for scalable graph analysis. It enables the computation of essential graph algorithms efficiently, making it ideal for handling large datasets, such as the Facebook combined graph used here.

Notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3890371488063011/4262159529860767/567627597756756/latest.html

III. Study of the Technique

1. *Outdegree and Indegree Analysis:*

Outdegree identifies nodes with the most outgoing connections, often reflecting influence. Indegree highlights nodes with the most incoming connections, indicating popularity.

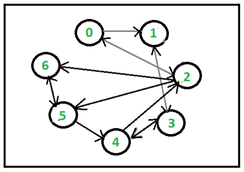


Fig. 2 Example of what outdegrees and indegrees can look like.

2. *PageRank:*

Calculates node influence based on connections, using iterative rank distribution across the graph.

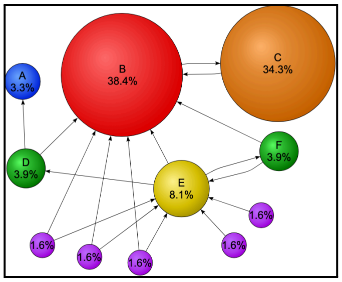


Fig. 3 The "bigger" the node / percentage, the higher it is on page rank.

3. *Connected Components:*

Groups nodes into subgraphs where every node is reachable from any other node in the same group.

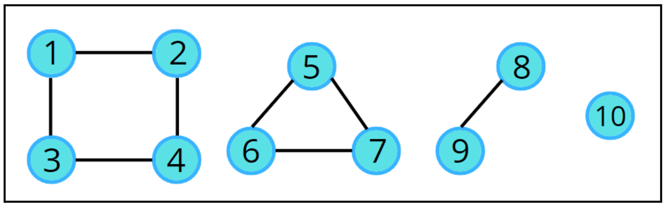


Fig. 4 Connected Components Example.

4. *Triangle Count:*

Counts the number of triangles (closed triplets) involving each node to measure clustering and connectivity.

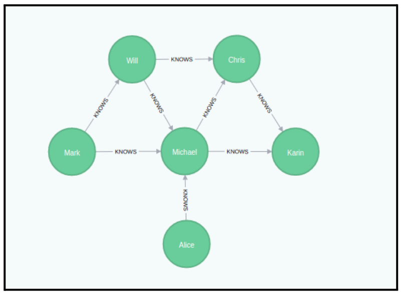
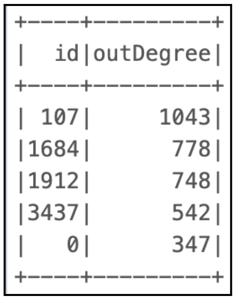


Fig. 5 Triangle Count Example

IV. Results and analysis

1. *Outdegree Analysis:*

Top 5 nodes with the highest outdegree were identified, showcasing the most outwardly connected users. “Who connects the most?” Some people in the network are super active in making connections - like sending friend requests to everyone. One person (Node 107) was friends with over 1,000 others, making them the most outgoing.

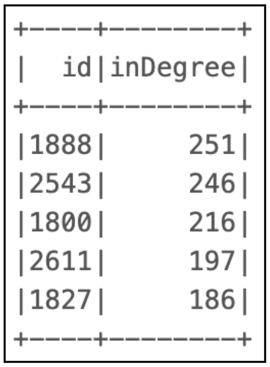


```
+----+---------+
|  id|outDegree|
+----+---------+
| 107|     1043|
|1684|      778|
|1912|      748|
|3437|      542|
|   0|      347|
+----+---------+
```

Fig. 6 Top 5 highest out degrees. For example, node 107 has an out-degree of 1043, indicating it has 1043 outgoing edges. This means node 107 is friends with a lot of people.

2. *Indegree Analysis:*

Nodes with the most incoming connections highlight popular users in the network. “Who is the most popular?” These people have the most followers / sent friend requests their way.

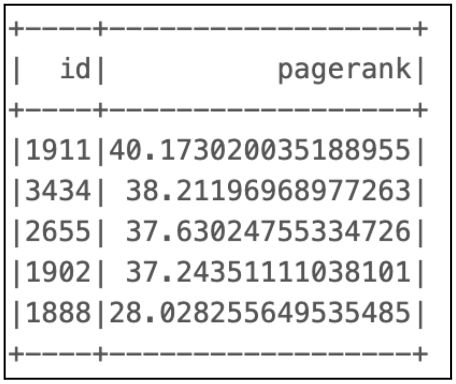


```
+----+--------+
|  id|inDegree|
+----+--------+
|1888|     251|
|2543|     246|
|1800|     216|
|2611|     197|
|1827|     186|
+----+--------+
```

Fig. 7 Shows the top 5 nodes with the highest in-degrees.For example, node 1888 has an in-degree of 251, meaning 251 edges point to it.

3. *PageRank:*

The top 5 influential nodes based on PageRank were identified, indicating key hubs in the network. “Who’s the biggest influencer?”



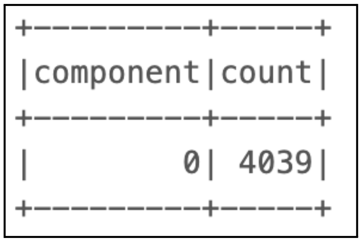
```
+----+------------------+
|  id|          pagerank|
+----+------------------+
|1911|40.173020035188955|
|3434| 38.21196968977263|
|2655| 37.63024755334726|
|1902| 37.24351111038101|
|1888|28.028255649535485|
+----+------------------+
```

.

Fig. 8 These are the top 5 page ranked nodes. We computed PageRank with a damping factor of 0.15 and 10 iterations.

4. *Connected Components:*

The largest connected components were detected, reflecting significant community structures. “Which groups exist?” The biggest group had over 4,000 people all interconnected.



```
+---------+-----+
|component|count|
+---------+-----+
|        0| 4039|
+---------+-----+
```

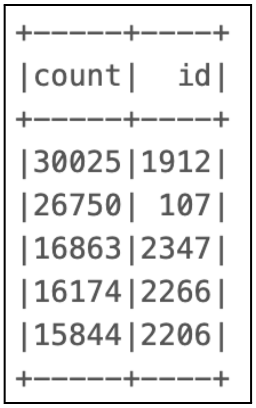
Fig. 9 The connected components dataframe shows one component (0) with 4039 nodes.

.

5. *Triangle Count:*

Nodes involved in the most triangles were identified, suggesting highly clustered regions.

Each technique provided unique insights into the dataset's structure, connectivity, and influential nodes. “Who’s in the cliques?” One person (Node 1912) was part of over 30,000 of these mini cliques, meaning they’re in the middle of tons of overlapping friend groups.



```
+-----+----+
|count|  id|
+-----+----+
|30025|1912|
|26750| 107|
|16863|2347|
|16174|2266|
|15844|2206|
+-----+----+
```

Fig. 10 Shows the top 5 nodes with their respective triangle counts. ID 1912 has 30025 triangles.

V. Conclusions

This study demonstrates the power of graph-based techniques in analyzing large social networks. Key influencers were identified through PageRank and degree metrics, while community structures emerged from connected components analysis. Triangle counting highlighted highly clustered areas, offering insights into local connectivity.

Future work could explore additional metrics or even more data from different areas, instead of just Facebook.

Video Explanation:
https://www.youtube.com/watch?v=1qgiQqwFU2A