Analysis of Facebook and Transportaition Dataset using Graphs Data Structure and Algorithms*

*Note: Conducted an analysis of a Facebook dataset(Stanford Large Network Dataset SNAP)

1st Ayush Singh
IIIT Dharwad(22BDS012)

2nd Avinash Tiwari
IIIT Dharwad(22BDS010)

3rd Nachiket Apte
IIIT Dharwad(22BDS041)

4th Yashraj Kadam
IIIT Dharwad(22BDS0066)

***Abstract*—This research paper compares 4 graph algorithms, Kruskal and Prim's for MST, Dijkstra's and Bellman-Ford for shortest path, on two multigraph datasets. It finds various measures such as averages, minimum/maximum, medians values, while applying the PageRank algorithm for page ranking based on mutual likes. Community detection is performed using the Louvain algorithm, and the role of closeness centrality is in formulating comprehensive results.**

I. Introduction

This report analyzes six graph algorithms: Kruskal's, Prim's for MST, Bellman-ford, and Dijkstra's for shortest path and PageRank[1] and Louvain[2] algorithm. It evaluates their efficiency and effectiveness in handling two different graph datasets with multiple edges. The study also examines graph operations like medians, minimum values, maximum values, and averages, providing statistical insights into graph topologies. The goal is to advance understanding and practical use of graph algorithms, facilitating decision-making in fields like network optimization, resource allocation, and route planning.

II. Description of datasets

A. Verified Facebook Page Networks[3]

These datasets represent blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them.

TABLE I
Classification of Datasets

Table Sr. No.	Verified Facebook Page Networks		
	Categories	***Nodes***	**Edges**
1)	Government[a]	7,057	89,455
2)	Public Figures [a]	11,565	67,114
3)	Politicians[a]	5,908	41,729

[a]3 different datasets

B. Multi-Layer Transport Network Dataset[4]

This dataset represents 3 multigraphs i.e. roadways, railways, airways connecting different cities. Here cities are represented by nodes, the edges represent the routes between them and edge weight represents time taken to travel between corresponding nodes(cities).

Prof. Animesh Chaturvedi, Researcher, Teacher, and Educationist

III. Analysis of Dataset[5]

Statistical measures such as mean, median, maximum, and minimum of the MST weight and Shortest path and practical applications in social network graphs.

A. Multi-Layer Transport Network Dataset

Minimum Spanning Tree

i) Average of MSTs: It gives us average cost/distance/time to establish an efficient system within network(nodes/cities).

ii) Maximum of MSTs: It shows the most costly or distance/time-consuming route needed to connect any two cities in the dataset.

iii)Minimum of MSTs: It displays the cheapest and best connection needed to connect any two cities in the dataset.

iv)Median of MSTs: It provides a balanced perspective on the transportation costs/distance/time in the network.

Performance Analysis(Shortest Path/Time)

i) Average Travel Time: It provides measures of typical time required to travel within the network(cities).

ii) Maximum Travel Path/Time:The maximum travel time represents the worst-case scenario where the longest time is required to travel within the network. It also suggests the presence of distant or less accessible city pairs within the transportation network.

iii) Minimum Travel Path/Time): It represents the existence of highly efficient routes or direct connections between specific city pairs. It also indicates the presence of well-established and direct transportation links and minimum possible time to travel between major cities.

iv) Median Travel Path/Time: It is the measure of central tendency. A lower median shortest path/time suggests that a significant portion of city pairs have relatively short distances between them, indicating good overall connectivity.

B. Verified Facebook Page Network Dataset

Minimum Spanning Tree

i) Average of MSTs: Typical distance or effort to connect pages within the category.

ii) Maximum of MSTs: Longest distance between any two pages, indicating potential barriers or influential connections.

iii) Minimum of MSTs: Shortest distance between any two pages, highlighting closely connected subgroups or clusters within the category.

iv) Median of MSTs: A higher median MST weight indicates stronger relationships or respect between page pairs, while a smaller weight indicates fluctuations in network connections or mutual liking.

Page Ranking(PageRank Algorithm)

i) Assessing Popularity and Engagement: It measures/ranks the popularity and engagement of government pages, politicians, and public figures by considering factors like followers, likes, comments, shares, indicating public interest and support.

ii) Personalized Recommendations: It helps users make personalized recommendations based on interests and preferences, enhancing engagement and experience by considering government pages, politicians, and public figures' rankings.

Category	**Node**
Governments	5417
Politicians	4902
Public Figures	3353

TABLE II
MOST POPULAR FACEBOOK PAGES

Community Detection(Louvain Algorithm)

i)Finding Groups of Pages: It identifies closely connected pages within a network, allowing identification of pages in various "clubs" or teams within the community of governments, politicians, and public figures.

ii) Following Information and Influence: It helps us see how information are passed around within and between groups. We can find important pages that connect different communities and make information flow between them.

Category	**Size**
Governments	6330
Politicians	4782
Public Figures	8506

TABLE III
LARGEST COMMUNITY SIZE

Closeness Centrality

It tells how close a node(user) is close to other nodes(users), which means whom a common man trusts more- Governments, Politicians or Public figures.

IV. CONCLUSION

The evaluation of graph metrics and algorithms reveals crucial information about multigraph datasets' effectiveness, connectivity, and properties. These findings apply to resource allocation, route planning, network design, and decision-making in areas with common graph structures. Understanding the linkages and connectivity within these networks enables wise decisions to increase efficiency and improve connectivity among entities.

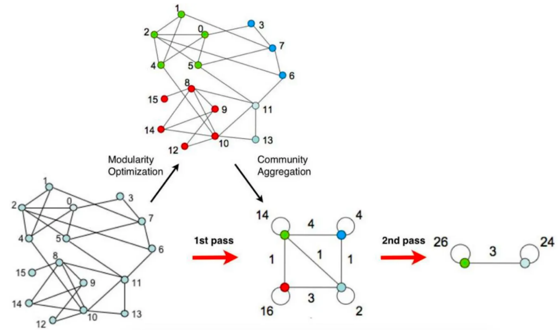

Fig. 1. Sequence of steps followed by Louvain algorithm

$$Closeness\ Centrality\ Score(u) = \frac{number\ of\ nodes - 1}{\sum\ (distance\ from\ u\ to\ all\ other\ nodes)}$$

Fig. 2. Normalized Centrality Score

TABLE IV
MULTI-LAYER TRANSPORTATION NETWORK

Property Name	Algorithms			
	Prim's	Kruskal's	Dijkstra	Bellmanford
Average	10.67	10.67	0.25427428	0.25427428
Minimum	10	10	0.22544909	0.22544909
Median	10	10	0.2686869	0.2686869
Maximum	12	12	0.2686869	0.2686869

TABLE V
FACEBOOK NETWORK(MST)

Property Name	Algorithm	
	Prim's algo.	Kruskal's algo.
Average	8175.67	8175.67
Minimum	7056	7056
Median	5907	5907
Maximum	11564	11564

TABLE VI
VERIFIED FACEBOOK PAGE NETWORK

Property Name	Algorithm	
	Dijkstra's	Bellmanford
Average gov.	0.26960224	0.26960224
Average pol.	0.21876992	0.21876992
Average pub.	0.22130418	0.22130418
least closeness centrality	0.21876992(pol.)	0.21876992(pol.)
median closeness centrality	0.22130418 (pub.)	0.22130418 (pub.)
maximum closeness centrality	0.26960224(gov.)	0.26960224(gov.)

REFERENCES

[1] https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af
[2] https://towardsdatascience.com/louvain-algorithm-93fde589f58c
[3] https://snap.stanford.edu/data/gemsec-Facebook.html https://arxiv.org/pdf/1802.03997.pdf
[4] https://drive.google.com/drive/folders/1rlx4OAD-a-gsGQZNVHZN6xKmHTnYwaix?usp=sharing
[5] https://algs4.cs.princeton.edu/40graphs/

https://github.com/AyushSingh916/Graph_Data_Structure_Project.git