CS 6350: Big Data Management and Analytics

Assignment 3

Names of students in your group:

1. Kushal Choudhary (kxc240000)

Number of free late days used: 0

Part 2: Analyze Networks with GraphFrame

Code: https://github.com/Kushal3121/2. Analyze Networks with GraphFrame

1. Data Source

- The dataset used for this analysis is the Epinions "Who-trusts-whom" social network, publicly available from the Stanford SNAP repository.
- It represents a directed trust relationship among users, where each edge (src, dst) means user src trusts user dst.
- The graph consists of approximately 75,879 nodes and 508,837 directed edges, making it well-suited for social network analysis using GraphFrames.

2. Processing and Visualization

The dataset was loaded into Apache Spark and analyzed using the GraphFrames API.
The following algorithms and queries were executed:

- Outdegree and Indegree Analysis: to identify the most active and most trusted users.
- PageRank: to measure overall influence or importance of users.
- Connected Components: to detect groups of users connected directly or indirectly.
- Triangle Count: to find tightly connected communities (mutual trust circles).

The computation was performed on a local Spark environment (Spark 3.5.0, GraphFrames 0.8.3) using Python and PySpark.

3. Results and Interpretation

Metric	Key Findings
Outdegree	Node 645 had the highest outdegree (1,801), meaning it trusted the most users.
Indegree	Node 18 had the highest indegree (3,035), indicating it was the most trusted user in the network.
PageRank	Node 18 again ranked highest, confirming it as a central and influential node.

Connected Components	Almost all users (75,877 out of 75,879) were part of a single large component, showing a highly connected trust network.
Triangle Count	Nodes 645, 18, 27, 634, and 44 had the most triangles, forming small clusters of mutual trust.

These results make logical sense in a trust network: a few nodes emerge as highly influential "hubs," while most users form part of a large, interconnected community. The overlap between high PageRank and indegree nodes validates the consistency of centrality measures.

4. Conclusion

This analysis confirms that:

- A small fraction of users dominate trust relationships.
- The network exhibits **scale-free** behavior, a few nodes have very high connectivity.
- The existence of triangles indicates strong local clustering (trusted friend groups).
- Nearly all nodes belong to a single connected component, showing a cohesive community structure.

Overall, the results are both statistically and conceptually consistent with expectations for online social trust networks. GraphFrames proved efficient for analyzing large-scale graph data, enabling computation of influence and structure metrics with minimal code.