Report On

**“Analysis of Social Network data using GraphX”**

By

Archana Bhatia D17A 10

Ria Dharmani D17A 21

Under the Guidance of

Dr. (Mrs.) Anjali Yeole



Department of Computer Engineering

Vivekanand Education Society’s Institute of Technology 2020-21

**Problem Statement:**

Analysis of social media datasets taken from YouTube and Facebook. Finding the most popular communities on YouTube using the PageRank algorithm and the number of triangles passing through each vertex in the Facebook dataset.

## **Graph Data Processing**

Graph data processing mainly includes graph traversal to find specific nodes in the graph data set that match the specified patterns and then locate the associated nodes and relationships in the data so we can see the patterns of connections between different entities. The data processing pipeline typically includes the following steps:

1. pre-processing of data (which includes loading, transformation, and filtering)
2. graph creation
3. analysis
4. post-processing

## **Spark GraphX**

GraphX is Apache Spark's API for graphs and graph-parallel computation. It extends the Spark RDD by introducing a new Graph abstraction: a [directed multigraph](https://en.wikipedia.org/wiki/Multigraph#Directed_multigraph_.28edges_without_own_identity.29) with properties attached to each vertex and edge.

GraphX library provides graph operators like subgraph, joinVertices, and aggregateMessages to transform the graph data. It provides several ways of building a graph from a collection of vertices and edges in an RDD or on disk. GraphX also includes several graph algorithms and builders to perform graph analytics tasks

**Algorithms Applied**

**1. PageRank**

PageRank algorithm is used to determine the relative importance of an object inside a graph data set. It measures the importance of each node in a graph, assuming an edge from another node to this node represents an endorsement.

Google's search engine is a classic example of PageRank. Google uses PageRank as one of the measures to determine the importance of a web page based on how many other web pages reference it.

Another example is social network websites like Twitter. If a Twitter user is followed by a lot of other users, then that user has a higher influence in the network. This metric can be used for ad selection/placement to the users that follow the first user (100,000 users follow a chef=> probably food lovers)

GraphX provides two implementations of PageRank: Static and Dynamic.

**Static PageRank:** This algorithm runs for a fixed number of iterations to generate PageRank values for a given set of nodes in a graph data set.

**Dynamic PageRank:** On the other hand, Dynamic PageRank algorithm runs until PageRank values converge based on a pre-defined tolerance value.

## 

## **2. Triangle Counting**

Triangle counting is a community detection graph algorithm that is used to determine the number of triangles passing through each vertex in the graph data set. A vertex is part of a triangle when it has two adjacent vertices with an edge between them. The triangle is a three-node subgraph, where every two nodes are connected. This algorithm returns a Graph object and we extract vertices from this triangle counting graph.

Triangle counting is used heavily in social network analysis. It provides a measure of clustering in the graph data which is useful for finding communities and measuring the cohesiveness of local communities in social network websites like LinkedIn or Facebook. Clustering Coefficient, an important metric in a social network, shows how much community around one node is tightly connected.

Other use cases where Triangle Counting algorithm is used are spam detection and link recommendations.

Triangle counting is a message heavy and computationally expensive algorithm compared to other graph algorithms. So, make sure you run the Spark program on a decent computer when you test the Triangle Count algorithm. Note that PageRank is a measure of relevance whereas Triangle Count is a measure of clustering.

**Dataset**

The datasets used for running Spark GraphX algorithms are YouTube and Facebook social networks and circles. These are available on the SNAP(Stanford Network Analysis Project) website, which is hosted by Stanford University.

1. **Youtube:**

In the Youtube social network, users form friendships with each other and users can create groups which other users can join. Such user-defined groups are considered as ground-truth communities. This data is provided by [Alan Mislove et al.](http://socialnetworks.mpi-sws.org/data-imc2007.html)

The YouTube social network data consists of 1134890 nodes and 2987624 edges. There are 8,385 communities formed.

1. **Facebook:**

The Facebook network dataset consists of 'circles' (or 'friends lists') from Facebook. The dataset includes node features (profiles), circles, and ego networks.

It has 4039 nodes and 88,234 edges.