SPARK GRAPH ANALYSIS

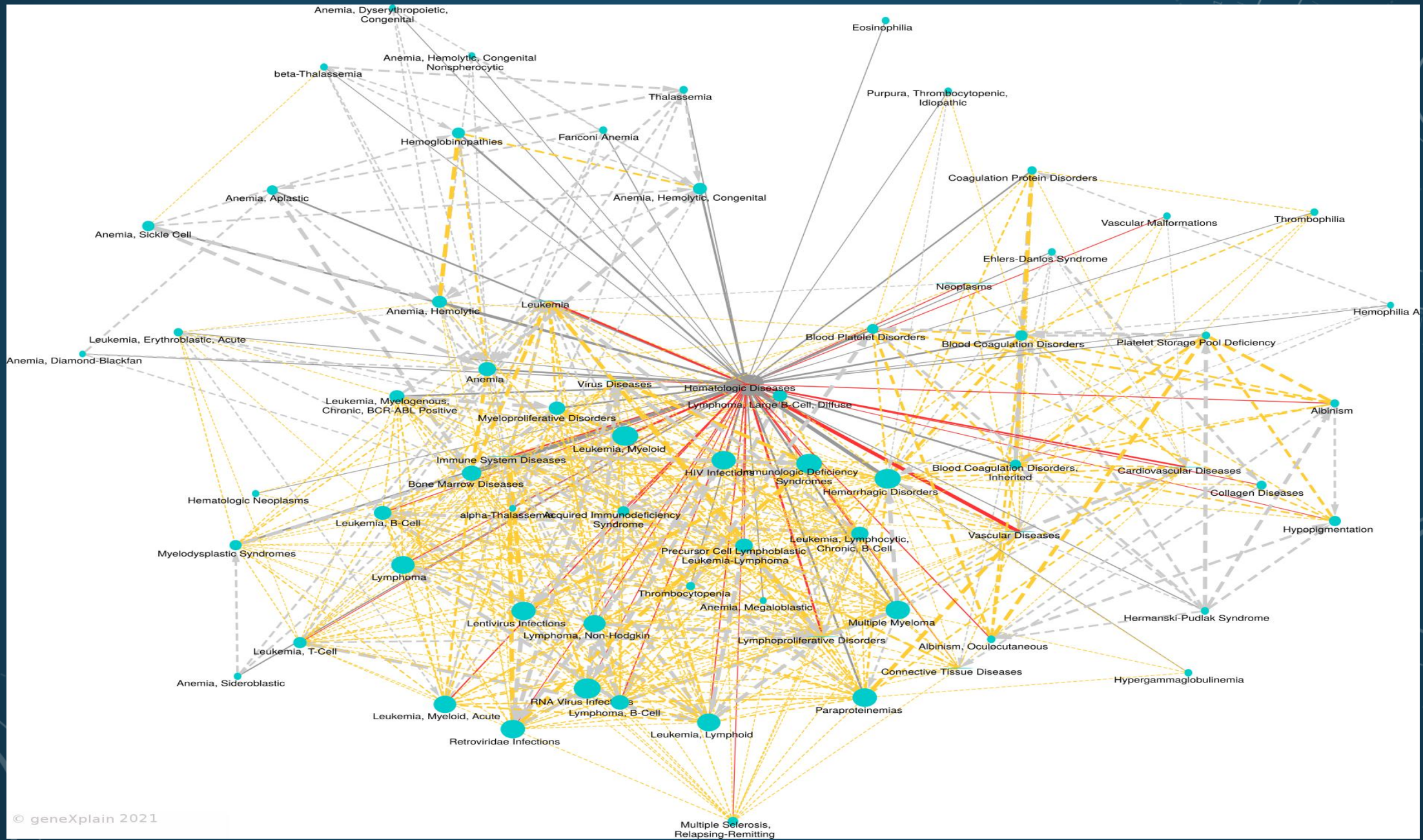
2024

The Hoang

LEARNING OBJECTIVES

- Graph Analysis
- Apache Spark GraphFrames
- Lab practice

GRAPH ANALYSIS – INTRODUCTION



GRAPH ANALYSIS

Definition: *A graph is a non-linear data structure consisting of nodes (also called vertices) and edges that connect them. Think of a graph as a network of interconnected objects, where each object is a node, and the connections between them are edges.*

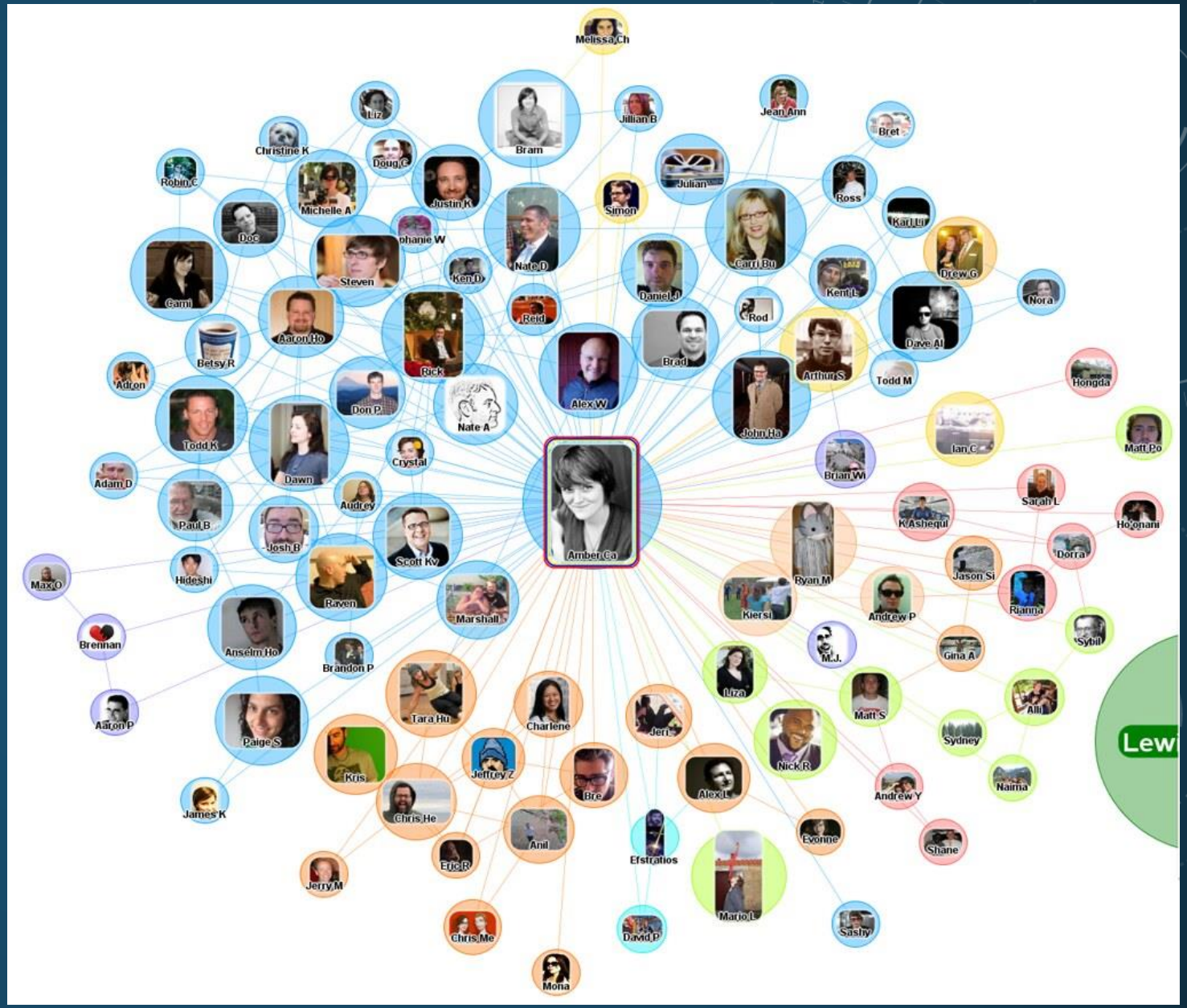
Types of Graphs:

- Undirected Graph: A graph where edges do not have direction. For example, a friendship network where two people are friends with each other.
- Directed Graph: A graph where edges have direction. For example, a social media network where one person follows another.
- Weighted Graph: A graph where edges have weights or values associated with them. For example, a road network where edges represent distances between cities.
- Unweighted Graph: A graph where edges do not have weights or values associated with them. For example, a simple friendship network.

GRAPH ANALYSIS

Undirected Graph Example

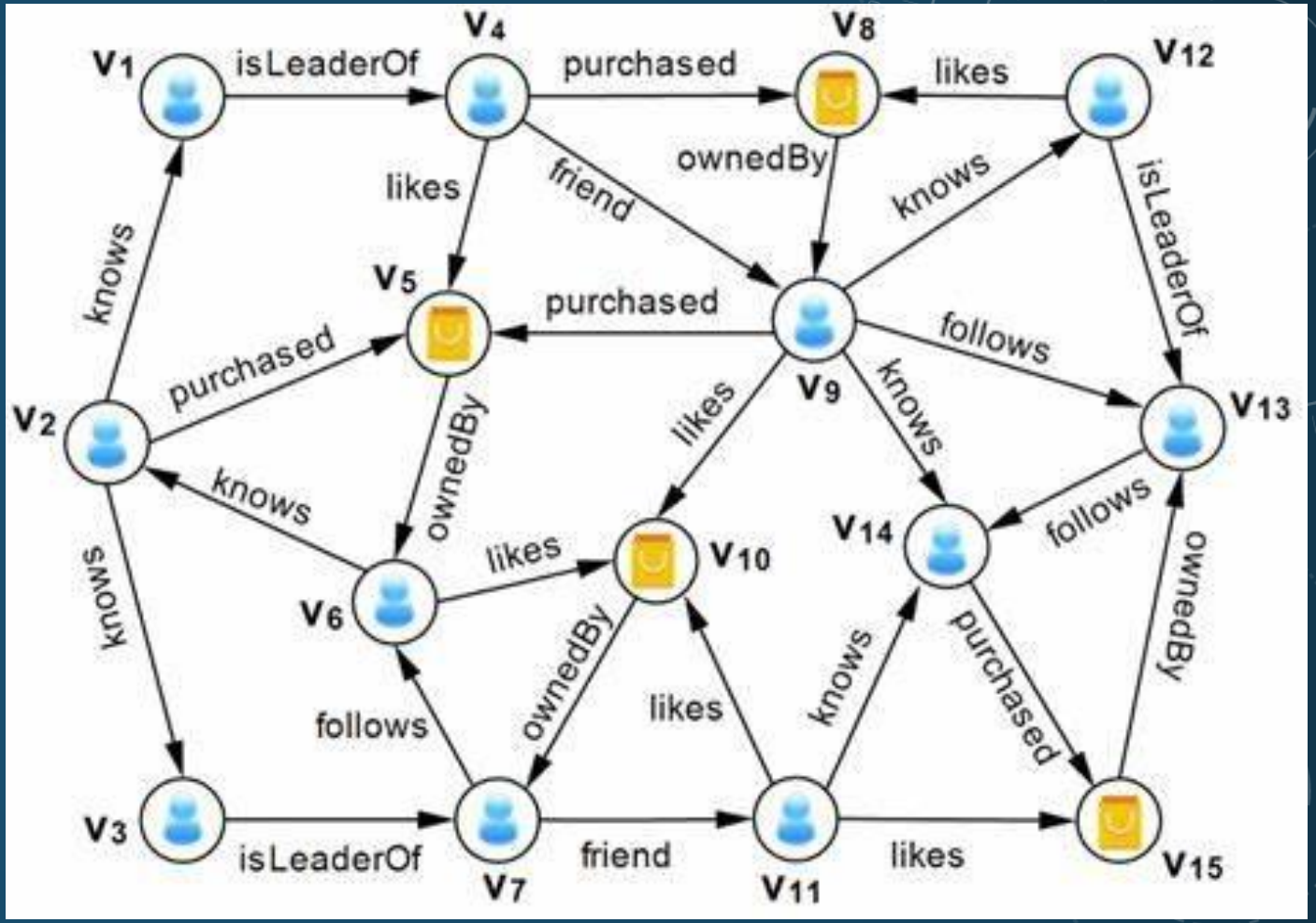
- Friendship Network



GRAPH ANALYSIS

Directed Graph Example

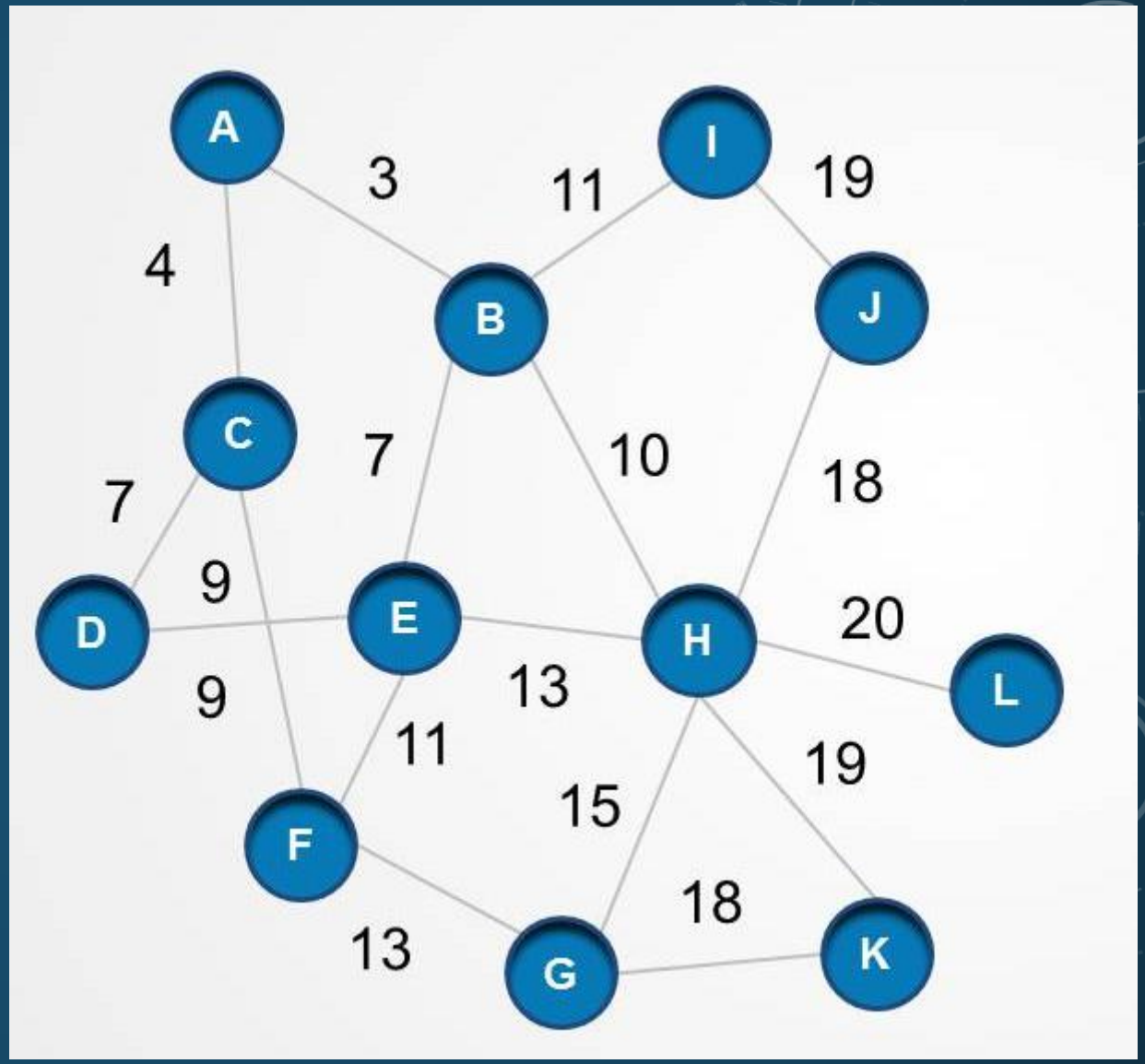
- Social Shopping Network



GRAPH ANALYSIS

Directed Graph Example

- Road Network



GRAPH ANALYSIS

Graph Analytics Concepts

- Node Degree: The number of edges connected to a node.
- Edge Weight: The value or weight associated with an edge.
- Shortest Path: The minimum number of edges required to travel between two nodes.
- Clustering Coefficient: A measure of how closely connected a node is to its neighbours.
- Centrality Measures: Measures of a node's importance or influence in the graph, such as PageRank or Betweenness Centrality.

GRAPH ANALYSIS

Graph Analytics Techniques

- Graph Search: Finding a specific node or path in a graph.
- Graph traversal to explore all nodes connected to a certain node.
- Finding shortest paths between nodes.
- Detecting communities or clusters of nodes that are densely connected.
- Identifying important nodes using centrality measures (like PageRank).
- Graph Clustering: Grouping similar nodes together based on their connections.
- Graph Embeddings: Representing nodes as vectors in a high-dimensional space to capture their relationships.

GRAPH ANALYSIS

Real-World Applications

- Social Network Analysis: Analyzing relationships between people in social media networks.
- Recommendation Systems: Recommending products or services based on user behaviour and preferences.
- Traffic Optimization: Optimizing traffic flow in transportation networks.
- Network Security: Identifying vulnerabilities in computer networks.
- Biological Network Analysis: Analyzing relationships between genes, proteins, and other biological molecules.

GRAPH ANALYSIS

Graph Analysis using Apache Spark

- Apache Spark GraphX, **GraphFrames**: A graph processing engine built on top of Apache Spark.

Feature	GraphX	GraphFrames
Underlying Structure	RDD-based	DataFrame-based
API Level	Lower-level	Higher-level, more user-friendly
Performance	Optimized for graph-parallel computations	Leverages DataFrame optimizations
Complex Queries	More manual; requires RDD manipulations	Supports SQL-like queries
Ease of Use	More complex for beginners	Easier for users familiar with DataFrames
Graph Algorithms	Basic algorithms provided	Rich set of algorithms with easier access

GRAPH ANALYSIS

Lab practice