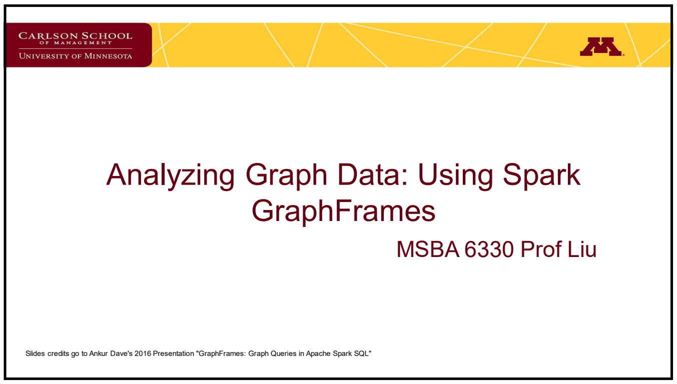
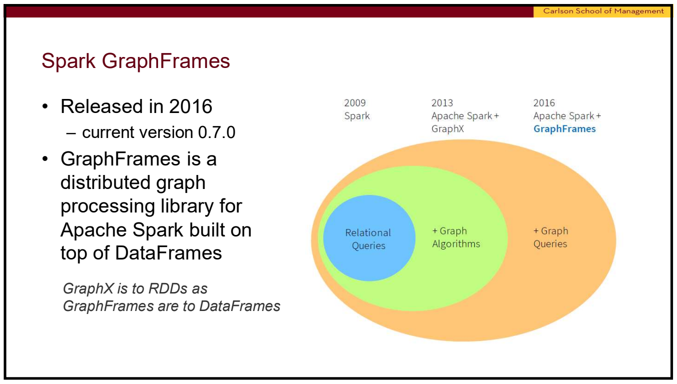


Carlson School of Management
University of Minnesota

# Analyzing Graph Data: Using Spark GraphFrames

MSBA 6330 Prof Liu

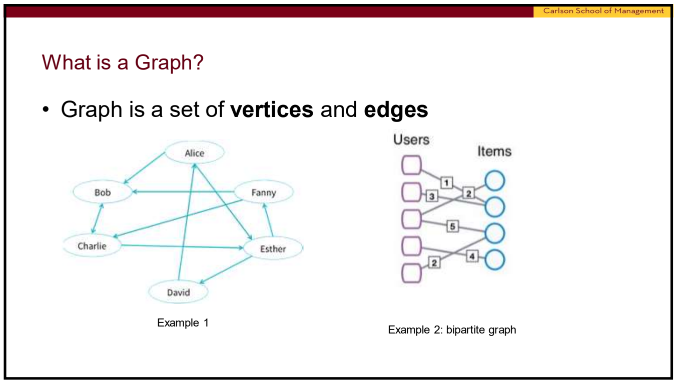
Slides credits go to Ankur Dave's 2016 Presentation "GraphFrames: Graph Queries in Apache Spark SQL"



## Spark GraphFrames

- Released in 2016
  - current version 0.7.0
- GraphFrames is a distributed graph processing library for Apache Spark built on top of DataFrames

*GraphX is to RDDs as GraphFrames are to DataFrames*

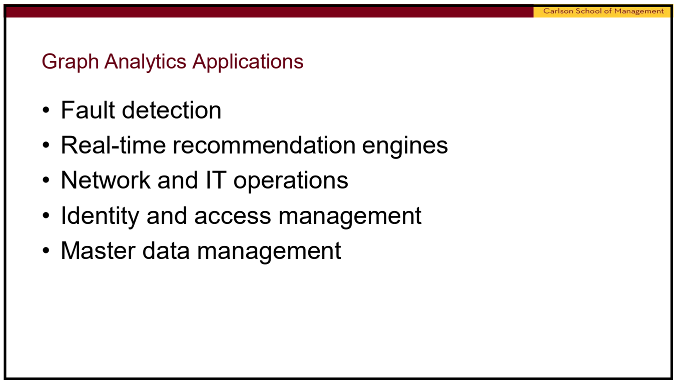


## What is a Graph?

- Graph is a set of **vertices** and **edges**

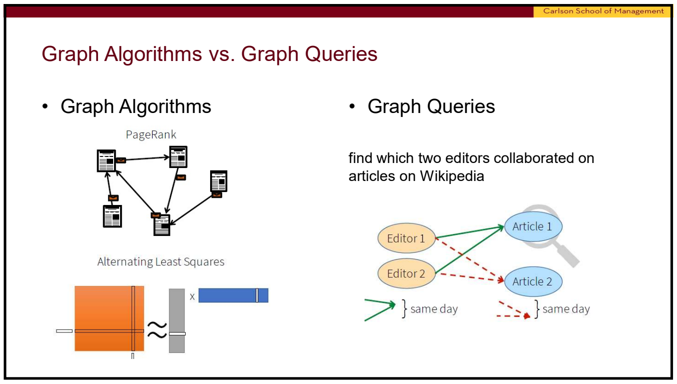
Example 1

Example 2: bipartite graph



## Graph Analytics Applications

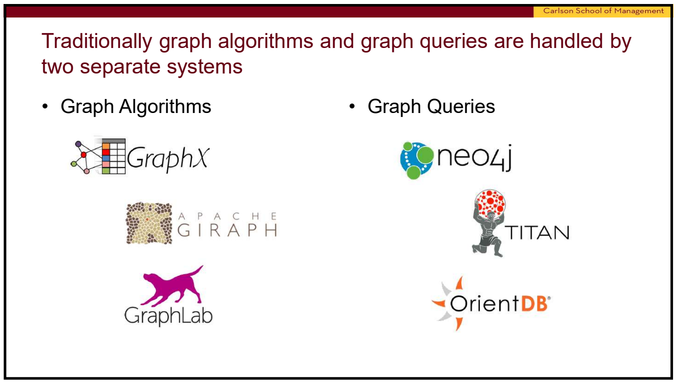
- Fault detection
- Real-time recommendation engines
- Network and IT operations
- Identity and access management
- Master data management



## Graph Algorithms vs. Graph Queries

- Graph Algorithms
- Graph Queries

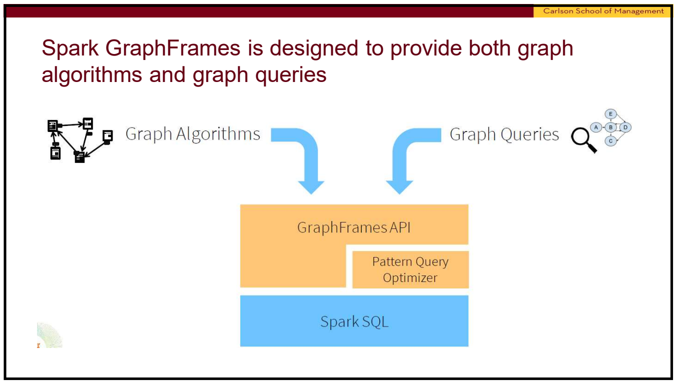
find which two editors collaborated on articles on Wikipedia



## Traditionally graph algorithms and graph queries are handled by two separate systems

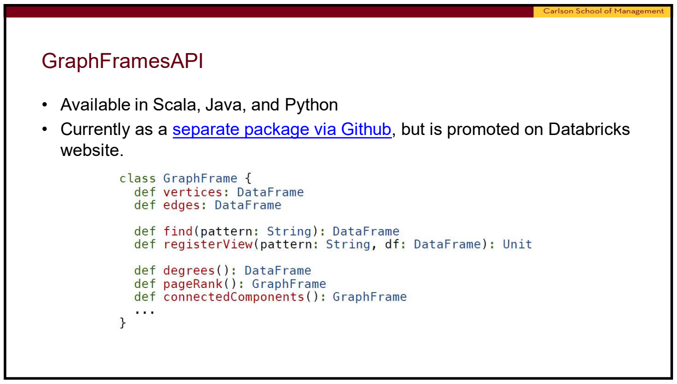
- Graph Algorithms
- Graph Queries

GraphX
APACHE GIRAPH
GraphLab
neo4j
TITAN
OrientDB



Carlson School of Management

## Spark GraphFrames is designed to provide both graph algorithms and graph queries



Carlson School of Management

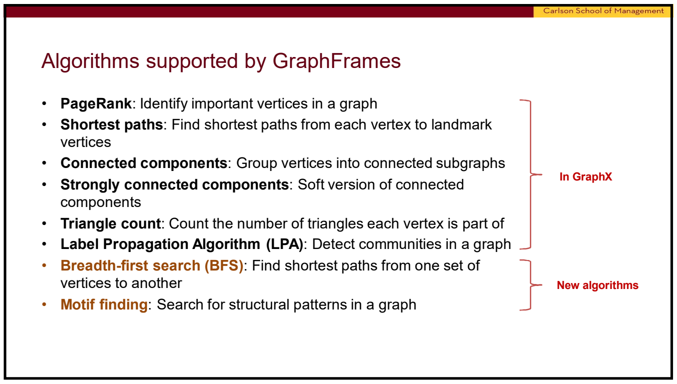
## GraphFramesAPI

- Available in Scala, Java, and Python
- Currently as a separate package via Github, but is promoted on Databricks website.

```
class GraphFrame {
  def vertices: DataFrame
  def edges: DataFrame

  def find(pattern: String): DataFrame
  def registerView(pattern: String, df: DataFrame): Unit

  def degrees(): DataFrame
  def pageRank(): GraphFrame
  def connectedComponents(): GraphFrame
  ...
}
```



Carlson School of Management

## Algorithms supported by GraphFrames

- **PageRank**: Identify important vertices in a graph
- **Shortest paths**: Find shortest paths from each vertex to landmark vertices
- **Connected components**: Group vertices into connected subgraphs
- **Strongly connected components**: Soft version of connected components
- **Triangle count**: Count the number of triangles each vertex is part of
- **Label Propagation Algorithm (LPA)**: Detect communities in a graph

In GraphX

- **Breadth-first search (BFS)**: Find shortest paths from one set of vertices to another
- **Motif finding**: Search for structural patterns in a graph

New algorithms