

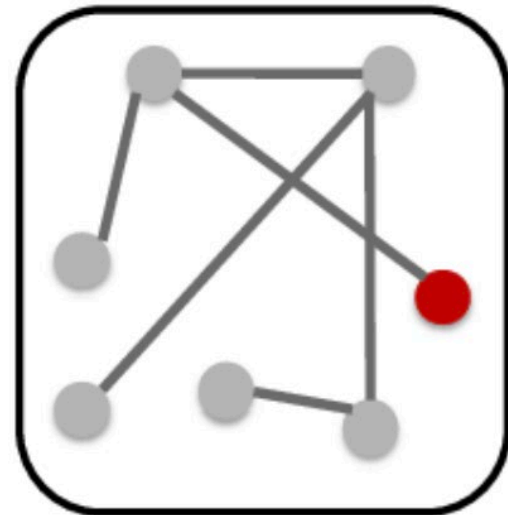
# Link Prediction with GraphX, Spark and MLlib

Brady

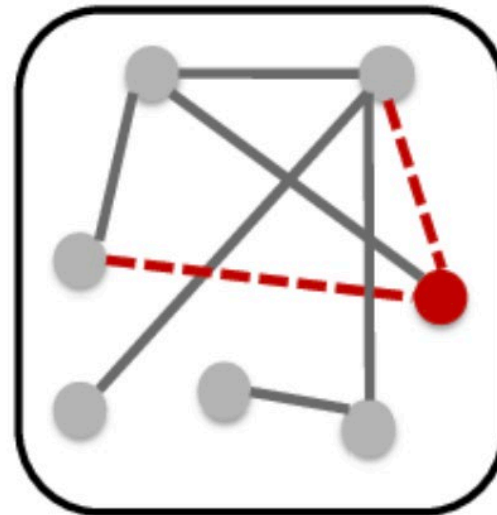


# What is Link Prediction

- Given current state of the graph
- Predict the likelihood of a future association between two nodes
- Application: bioinformatics, e-commerce, security domain
- Difficult Problem: Negative Link  $\gg$  Positive Link (Huge class skew)



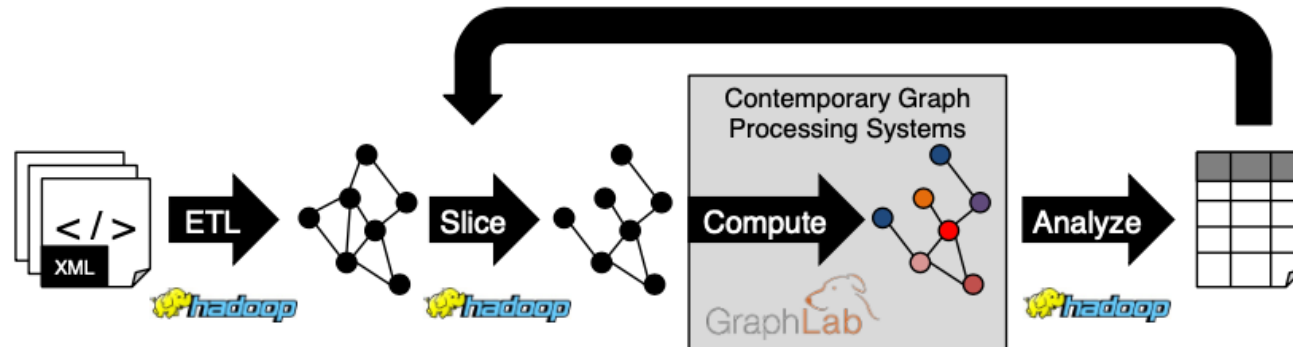
Time t



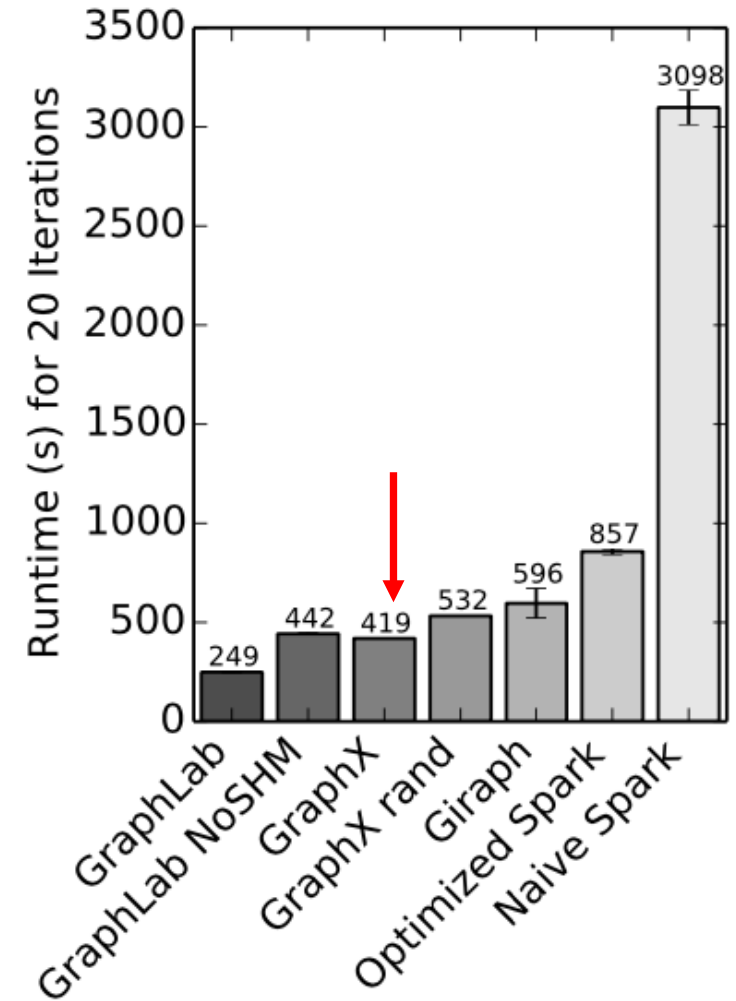
Time t+a

# Why GraphX

- View the same data as both graphs and collections



- Support from Spark
  - lineage-based fault tolerance
  - Benefit from Spark ecosystem
- Performance Comparable to other Frameworks
  - Giraph, GraphLab



(b) PageRank Twitter

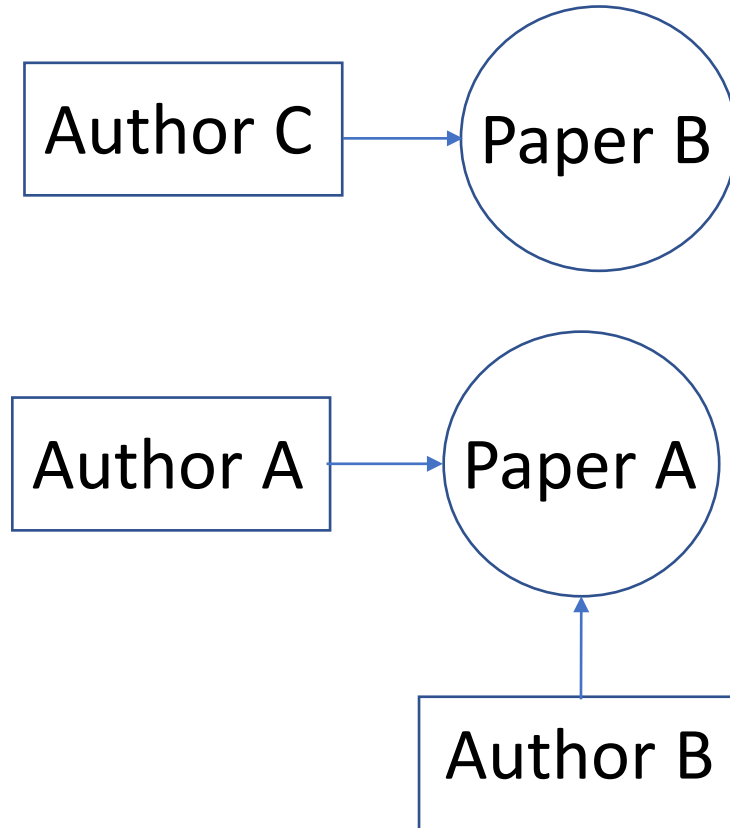


# Project

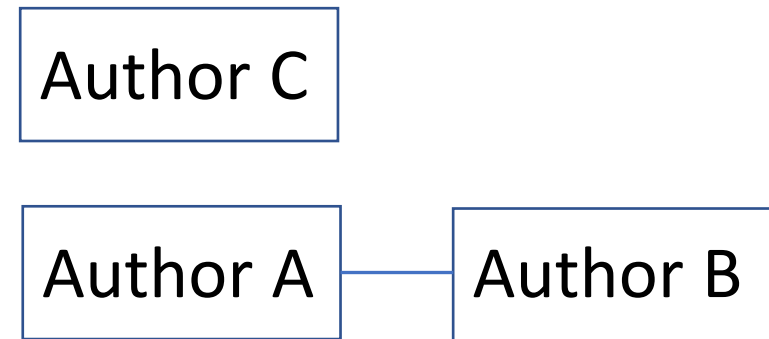
- Goal: Predict future co-authorships using DBLP citation dataset
- Tool: GraphX, Spark, MLlib
- Process
  - Pre-processing
    - Load Data
    - Build Graph
  - Actual Prediction
    - Unsupervised Learning
    - Supervised Learning
  - Evaluation
    - Unsupervised vs Supervised Learning

# Pre-processing

DBLP



Co-Author Graph






# Actual Prediction

- Unsupervised Learning (Similarity Metrics)
  - Common Neighbors (CN)
  - Jaccard's coefficient (JC)
  - Adamic/Adar (Adar)
  - preferential attachment (PA)
- Supervised Learning (Decision Tree - MLlib)
  - Feature Vector

<b>Node A</b>	<b>Node B</b>	<b>CN</b>	<b>JC</b>	<b>Adar</b>	<b>PA</b>	<b>Label (0/1)</b>
---------------	---------------	-----------	-----------	-------------	-----------	------------------------



# Work Plan

- Literature Review (2 weeks)
-  • Pre-processing (6 Dec – 12 Dec)
- Implement Similarity Metrics Algorithms (6 Dec – 12 Dec)
- Implement Supervised Learning (13 Dec – 19 Dec)
- Evaluation (20 Dec – 26 Dec)
- Project Report (27 Dec – 2 Jan)