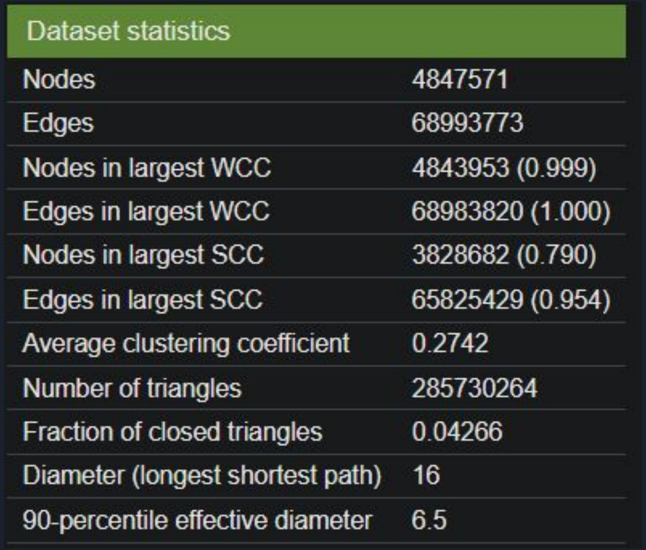
CS6240 Project Team 10

Fall 2020

April Gustafson, Mason Leon, Matthew Sobkowski

Project Overview

- LiveJournal Dataset from Stanford
- Social network graph with over 4 million nodes and 68 million edges
- Spark focused
- Tasks based on graph analytics



Dataset statistics	
Nodes	4847571
Edges	68993773
Nodes in largest WCC	4843953 (0.999)
Edges in largest WCC	68983820 (1.000)
Nodes in largest SCC	3828682 (0.790)
Edges in largest SCC	65825429 (0.954)
Average clustering coefficient	0.2742
Number of triangles	285730264
Fraction of closed triangles	0.04266
Diameter (longest shortest path)	16
90-percentile effective diameter	6.5

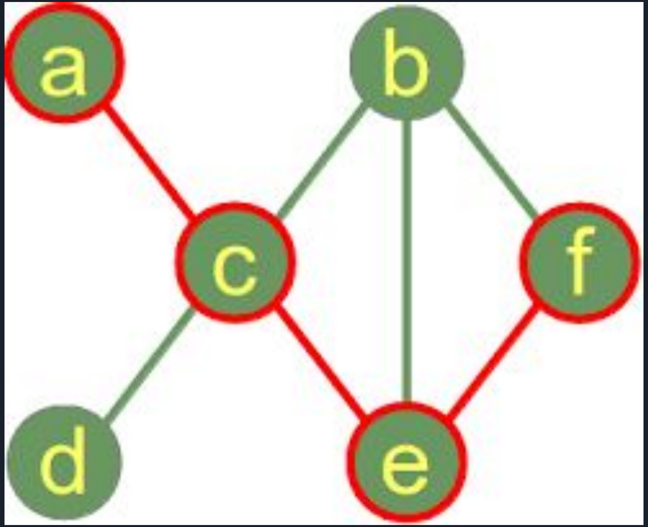
All-Pairs Shortest Path

- Find minimum distance from any node to all other nodes
- APSP represents minimum degree of separation between LiveJournal users
- Goal: translate the optimal algorithm into an optimal parallel program

Version Variation *cluster size and edge filter constant	Runtime
Converge when no distance updates	**02:58:41**
Co-partition joined datasets	**02:39:40**
Eliminate redundant join computation	**02:24:21**
Co-partition joined datasets in every iteration	**02:12:16**

Diameter

- Diameter: the longest shortest-path in a graph
- Also equivalent to the number of iterations required for APSP to converge
- Diameter of LiveJournal network: 16



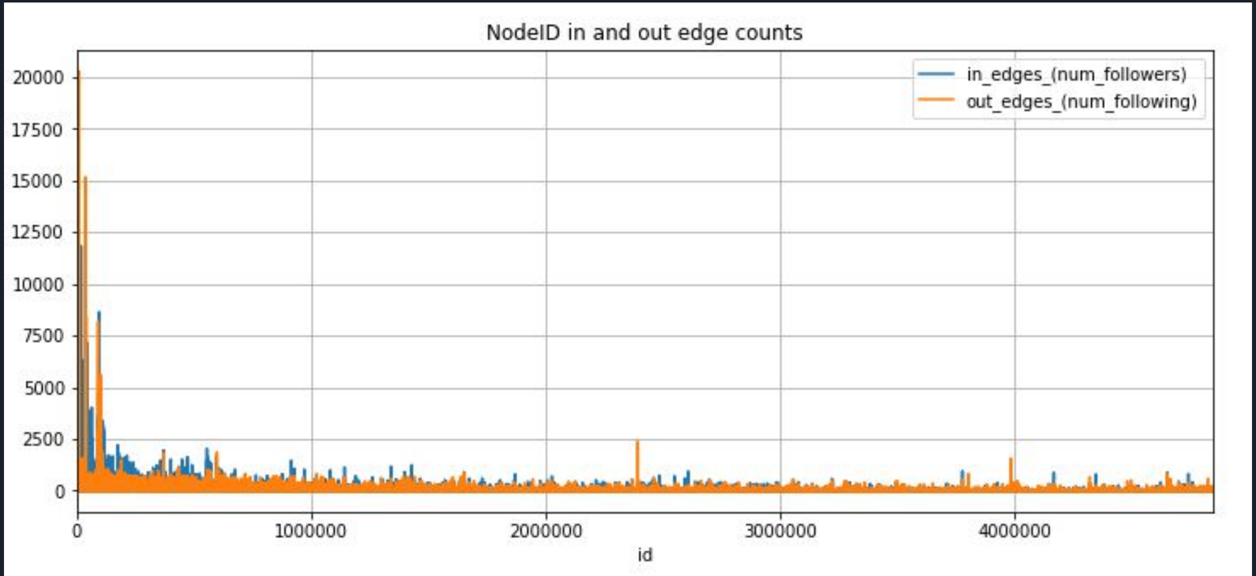
Cycles

- Finding number of distinct cycles with length n
- Filtered out all non-outgoing nodes from edges
- Use same partitioner for both structures

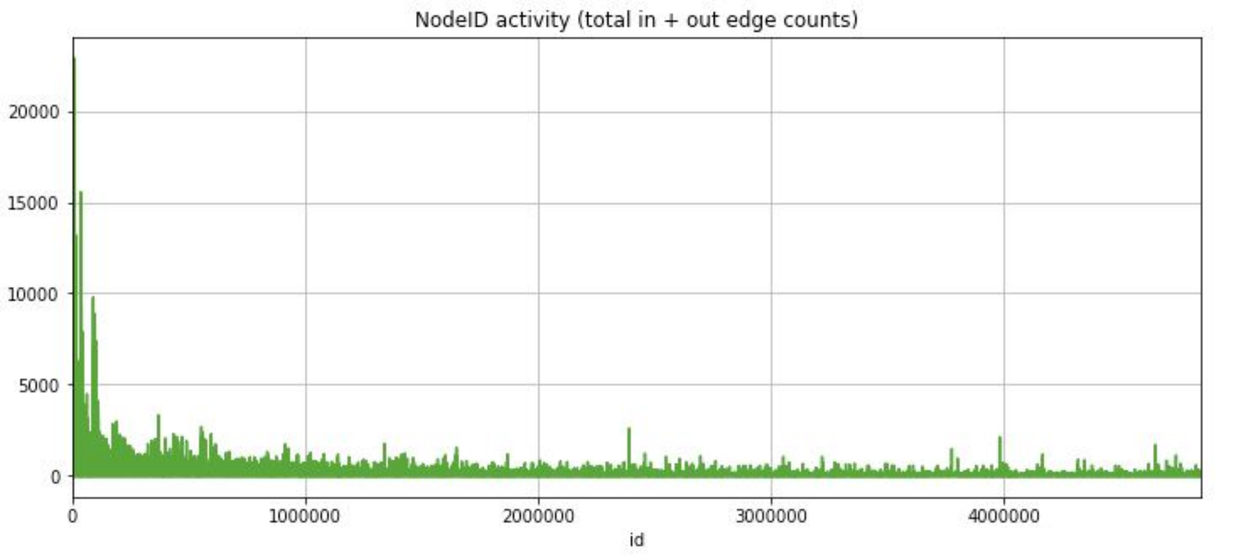
Filter Size	Runtime (min)
10000	1
20000	3
30000	8

Cluster Size with Filter = 30000	Runtime (min)
4 workers | 1 master	12
6 workers | 1 master	8

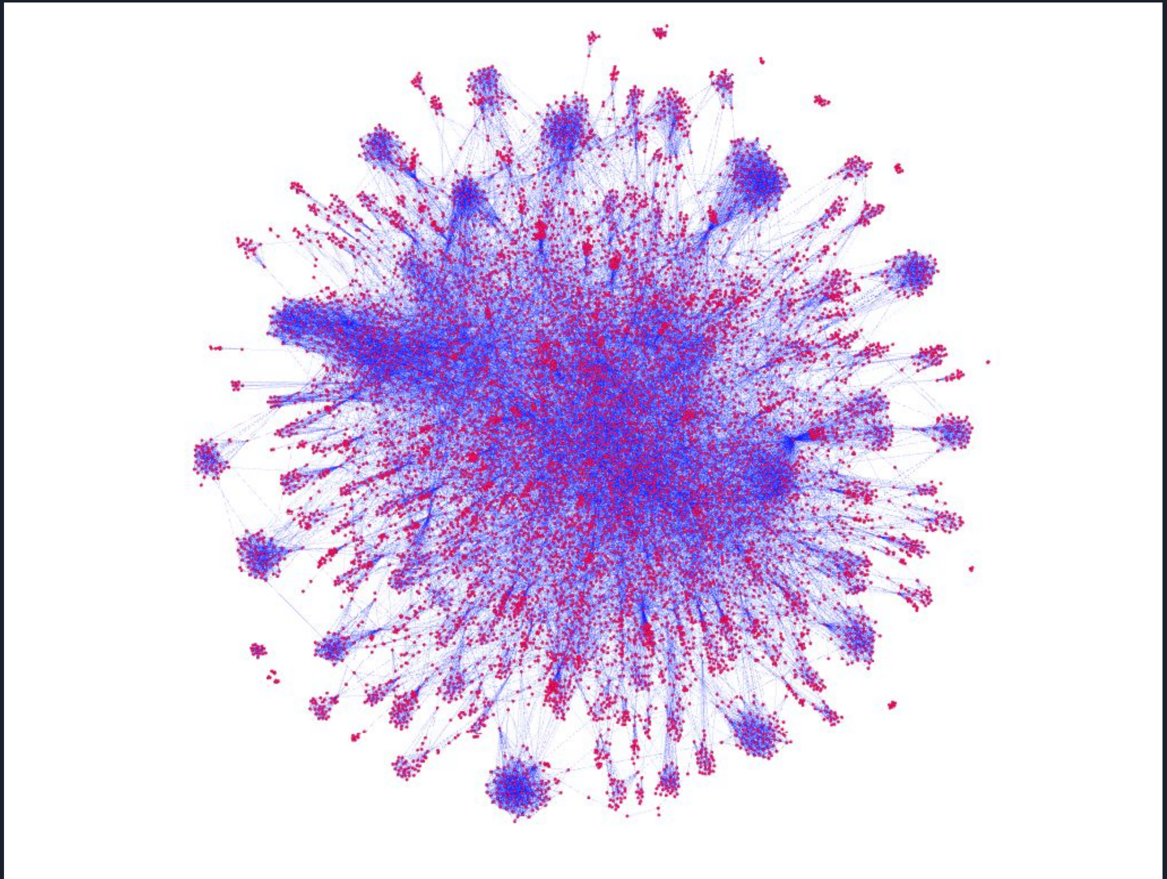
Investigation: Aggregated Node Activity



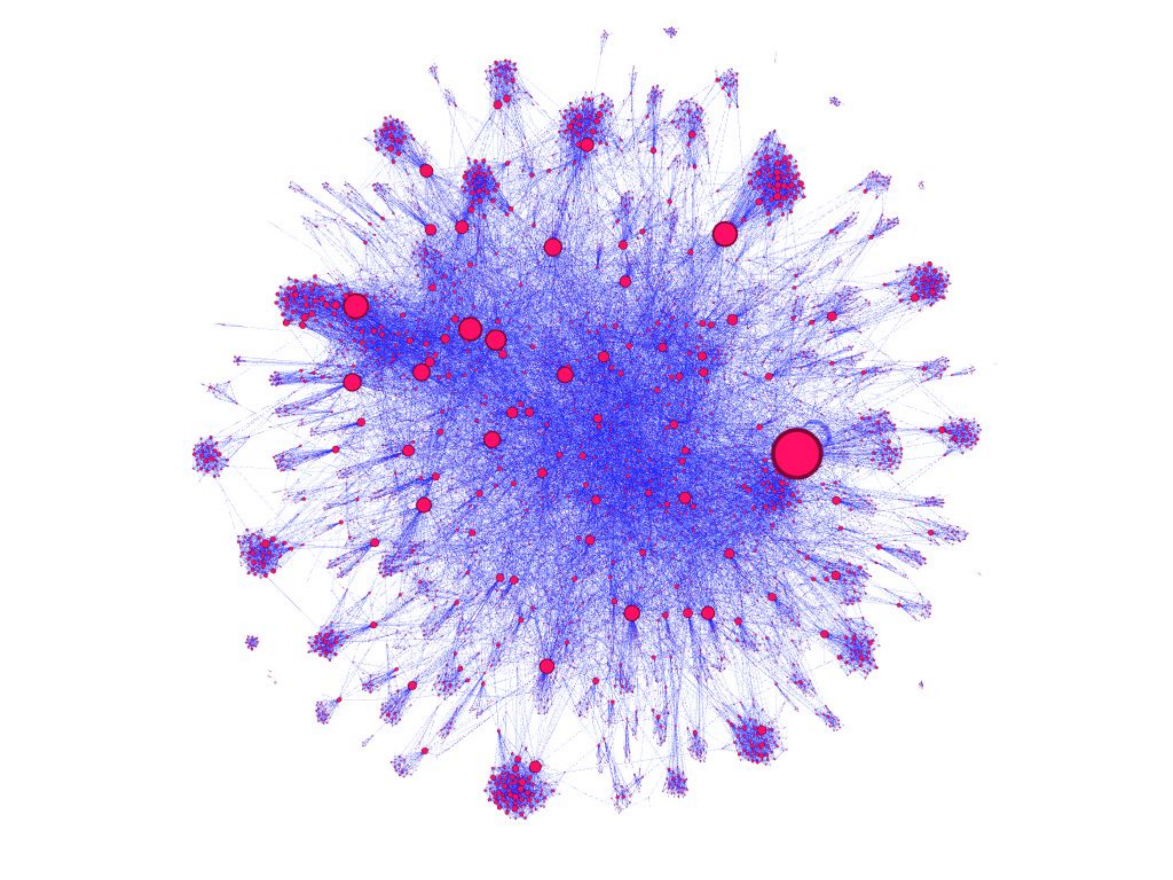
Investigation: Aggregated Node Activity



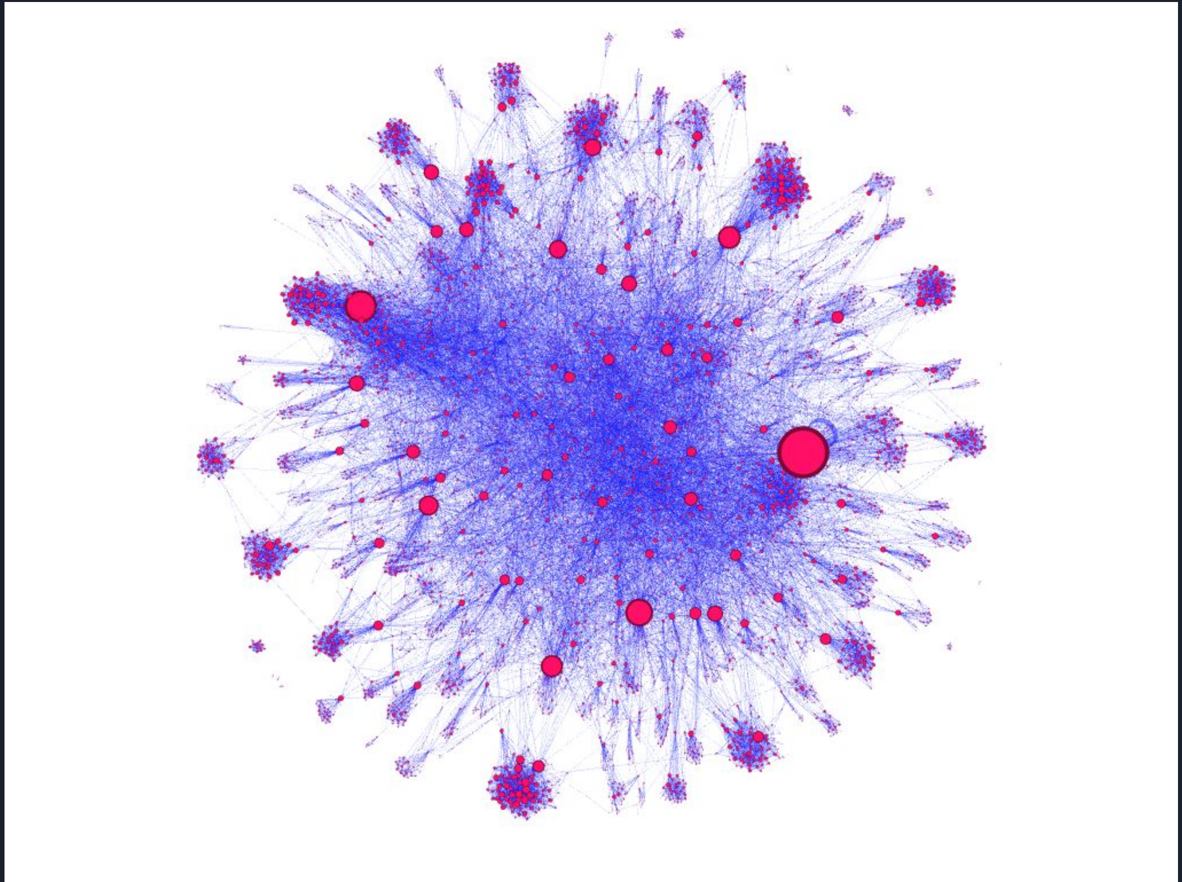
Gephi Visualization



Scale Node Diameter by In Degree



Scale Node Diameter by Out Degree



Total Degree