Checkpoint-4

Getting Started

We've created a csv of the data we'll be using for this checkpoint, which can be found at **src/supervisor_officer_coaccusals.csv.** If you'd like to recreate this CSV yourself, you may do so by running our SQL code at **src/checkpoint4.sql**.

Our databricks notebook is the html file **src/checkpoint4.html**. You should be able to upload this html file to databricks (as we saw in the demo in class) to see our workbook. Or, you can view a static version of the workbook with results by right clicking on the html file and opening it in your favorite web browser.

If you would like to re-run our code, just upload the csv to databricks before running anything.

src/pagerank.csv contains the pagerank results.

Our visualization was too much for databricks, but runs in the browser. You can run our visualization by opening the html files **src/entireNetwork.html** and **src/zoomedClusters.html**. Right click the file and select "open" to open it in your favorite web browser. It is a lot of data, so please give it a few minutes to load. We also created a large window to view the data so please scroll around to explore (you'll find most of the data by scrolling down and to the right).

These are visualizations of officers and supervisors that have been co-listed on allegations. The node color indicates unit, the size indicates if they are a supervisor (larger nodes are supervisors), and hovering over the node will display the officer's complaint percentile, unit, and ID.

entireNetwork.html shows all the officers in a condensed view, while **zoomedClusters.html** is a zoomed in view that allows you to see individual clusters more easily.

Our questions

We can identify clusters of officers who frequently collude (i.e. who are co-listed on allegations), focusing on the connections of supervisors to any other officers. This analysis will focus on the supervisors direct impact over those they manage over. We will also look at complaint percentiles to better understand the behavior of the officers in these clusters.

Using PageRank, we can identify the connections officers have among them, emphasizing the connections supervisors have with other officers. We can use the weights that PageRanks assigns to the connections to determine how much influence a supervisor has on other officers