The theme of the project is to analyse the trends before and after the cpdp went public. For example: to find changes in general behaviour of the officer, settlement amounts etc. and to predict for upcoming years. The questions are focussed towards finding out the things that have been tried to be hidden in the official record. For example, finding the relationship between the arrests and the allegation, because there are chances that if no arrest took place then the action taken by the officer might be illegal or looking for cases where things common between investigating officer and allegated officer impacted the case. So, for every question mentioned below, findings will be based on before and after cpdb release.

* Can officers who have been co-accused for misconduct be clustered into groups?

Are there any common attributes for the most important nodes from the graph? (Using page rank to find the most import nodes)

**Graph structure:**

Each officer is a node. Nodes have an edge between them if they are co-accused in any allegations.

* Can the officers who are charged for the same allegation category be clustered into groups? Are there any common attributes for the most important nodes from the graph? (Using page rank to find the most import nodes)

**Graph structure:**

Each officer. would be a node. A link will be formed between the officers if they have been charged for the same allegation category.

**Execution instructions:**

* Just run the .html files provided in the src folder by opening them in your browser.
* If you wish to execute the code yourself, then import the Ipython notebooks provided into Databricks.
  + GraphFrame must be installed under the libraries section of the Cluster on which you plan to run the notebook with. (Type Maven)
  + If GraphX Maven distribution doesn’t work, then the .jar can be downloaded from here: (<http://spark-packages.org/package/graphframes/graphframes>) and installed by uploading the .jar file in libraries section of the Cluster.
  + NetworkX must be installed under PyPi distribution.
  + Additional tips: If CREATE TABLE queries aren’t working then run the following query:
    - dbutils.fs.rm("dbfs:/user/hive/warehouse/", true) Replace the path by the path shown in the error. **Only run if CREATE TABLE query isn’t working.**

**Additional instructions:**

If the graph forming cell isn’t working, then re-run the previous cells. This happens as we are choosing random samples and some random samples may contain null values.