COS424 HW3: Leveraging Network Analysis and Clustering to Find Meaningful Patterns within Complaints Against NYPD officers

Co Author: Lucas Irwin **Co Author: Darin Avila**

Abstract

Police brutality and power abuse are prescient topics in modern American political life, as evidenced by the recent trial of Derek Chauvin over the brutal murder of George Floyd. Given the real harms which police brutality poses to people (especially people of color) on a daily basis, the task of predicting future abuse and identifying ”bad apples” amongst active officers is paramount to the safety of minority communities. In this project, we use data taken from the NYPD Misconduct Complaint Database to carry out two key unsupervised learning tasks: network analysis and clustering. To conduct network analysis, we construct a graph of police officers as nodes and shared complaints as edges. We run three centrality models (degree centrality, eigenvector centrality, Pagerank) on the graph and draw graphs using subsets of the data involving officers with the highest number of race-based and physical abuse allegations against them. For clustering, we use various clustering models (K-means, Birch) to separate the officers into groups based on their allegations and investigate similarities within those groups. In our results, we present visualizations of our graphs and analyze their properties, and also provide tables outlining the difference in conditional distributions of allegation clusters based on an officer’s rank and command.

1 Introduction:

Police brutality is one of the most serious issues deserving attention in American political life in the 21st century. In an attempt to address this problem, we use data from the New York City Civilian Complaint Review Board (CCRB) comprised of complaints against individual officers to carry out network analysis and clustering in order to identify meaningful patterns between NYPD officers involved in complaints in the New York City area. For network analysis, we were interested in identifying the officers with the highest centrality in a graph of officers. We therefore created a graph of officers as nodes and shared complaint ids as edges, and ran three centrality measure models (degree centrality, eigenvector centrality and Pagerank) on the resulting graph. This allowed us to draw the graphs of the 100 officers with highest scores across different centrality measures to visualize how strongly connected these graphs were. Additionally, we selected the officers with the highest number of allegations described as ”black”, ”hispanic” and ”punch/kick” and created sub-graphs of these officers to investigate how connected these were.

We considered desirable results of centrality measures to be highly connected components which would allow us to identify a large proportion of the complaints. This would allow for the shrinking of the data set to focus only on those officers who were involved in the majority of complaints. We expected eigenvector centrality to be most effective at this task since it considers both the centrality of the node as well as the centrality of the node’s neighbors, meaning subgraphs of nodes with the highest eigenvector centrality would likely be highly connected.

For clustering, we used K-means and BIRCH clustering to cluster based on the type of abuse and the allegations themselves. Desirable behavior for clustering would be an algorithm which placed officers of similar categories (such as rank or command) in the same cluster to investigate whether these attributes factor into an officer's list of allegations. We also use K-means clustering with principal component analysis (PCA) to reduce the allegation features into a much smaller dataset which still contains the vast majority of the variance necessary to cluster. After clustering using these techniques, we were able to investigate conditional distributions of these clusters among officers based on their rank and command and discovered that these distributions were indeed affected by these properties.

2 Related Work:

Network analysis is a broad field which spans many areas of discipline. We were interested in focusing on centrality measures. The most commonly referenced were the Pagerank algorithm, Degree centrality and Eigenvector centrality[1][2]. We were inspired to use Eigenvector centrality in particular because it considered the influence of both the node and the node's neighbors in a network. When it came to clustering, K-means clustering was by far the most popular model[3] among related projects, so we decided to leverage the algorithm to cluster our own data. PCA was referenced as a useful tool[4] for improving the performance of K-means clustering, and we took note that the elbow method was also a useful evaluation model[5] for determining the optimal number of clusters in a data set.

2.1 Data Processing:

We utilize two data sets taken from the New York City Civilian Complaint Review Board (CCRB). The first is an NYPD Misconduct Complaint Database consisting of complaints made by the public against police officers in New York City boroughs, spanning two different periods: (i) prior to 1994 when the CCRB operated within the NYPD and (ii) since 1994 when it has operated as an independent agency. It includes 323,911 unique complaints against 81,550 active or former NYPD officers. The second is a CapStat.NYC Police database consisting of 12,450 police officers along with attributes including their rank, NYC district, salary etc.

For our network analysis tasks, we used the complaint data set to construct a graph of police officers as nodes and shared complaint IDs as edges. To prep the data, we dropped columns consisting of ”N/A”, and created a data set consisting of a column for the complaint Id and a column for the combined first name and last name of the officer. We then use the Python Networkx library to construct a graph by adding edges between all officers with shared complaints. We remove all edges which are self-loops to create a simple graph, and were left with a graph of 35760 nodes (representing 35760 officers) and 71425 edges.

In order to cluster the police officers based on the nature of the allegations lodged against them, we first had to process the data such that each officer had all of the complaints of a certain type consolidated into one variable which contained the quantity of these complaints against this officer. This was accomplished by first creating dummy variables for the FADO Type and Allegation column. Then, because many officers had multiple complaints, each officer was taken as their own data frame (based on their Unique ID), and had these dummy variable rows summed up. This created a data set with over 100 features describing the type and quantity of abuse perpetrated by each officer. We could then use these features to cluster the officer into groups based on their offences.

2.2 Methods:

Link Analysis:

1. *Degree Centrality:* measures the number of edges/the degree of a node in a network. The greater the degree, the more central the node is.

2. *Eigenvector Centrality:* A measure of the node's influence in a network. Eigenvector centrality measures the importance of a node while also giving consideration to the (eigenvector) centrality of its neighbors. For instance, an officer who is involved in allegations with officers who have many

allegations against them themselves will have a higher score than an officer who shares allegations with officers who have few allegations against them.

3. *Pagerank Algorithm:* PageRank works by counting the number and quality of edges to a ”page” (or node) to determine a rough estimate of how influential the node is. The underlying assumption is that more influential nodes are likely to receive more links from other nodes.

We used the Networkx library in Python to calculate the three different types of centrality for officers in the graph. For each type of centrality, we sorted the officers by centrality score and drew a graph of the 100 officers with highest scores. We also evaluated the average degrees of subgraphs of nodes with high degree centrality to see whether the average degrees of these sub-graphs was greater than the average degree of the entire graph.

Clustering:

1. *K-means:* Clustering algorithm which calculates centroids for each cluster, and assigns each data point into the closest cluster using the metric of Euclidean distance in the feature space, and then recalculates the position of the centroids.

2. *K-means with PCA:* This method uses the K-means algorithm described above on a data set with a dimensionality which has been reduced via principal component analysis, a method which utilizes a singular value decomposition of the data matrix to produce a matrix with a smaller dimension whose set of fewer components can explain much of the variance found in the original matrix. This allows us to work with a much smaller data set without losing much information from the original data set.

3. *BIRCH:* BIRCH clustering is a four-step clustering process which creates a tree of sub-clusters and their summary statistics, and then uses an existing clustering algorithm to cluster the leaves of the tree to the desired number of clusters.

3 Spotlight Model:

K-means clustering is an unsupervised machine learning model which takes a set of data points containing $x_i \dots x_n$ vectors of p features, and group them into k clusters.

$$\text{Partition } D = \{x_1, \dots, x_n\} \text{ into } K \text{ clusters}$$

The goal is to create a centroid for each cluster, or the arithmetic mean of all of that' cluster's coordinates in feature space, and to group the coordinates such that the sum of the distance between the points and the cluster centroids is minimized. The metric to choose this distance may depend on the data, but oftentimes a p-norm is used, and most often p is set to be 2, giving us the Euclidean distance, defined as follows:

$$d(x_i, \eta_k) = \sqrt{\sum_{j=1}^{p} (x_{i,j} - \eta_{k,j})^2} \tag{1}$$

The algorithm begins with an initialization of k centroids, which can be accomplished by a variety of means. One can choose points in D, sample the feature space with a uniform distribution, or use some other function of the data. It may take multiple trials to determine which of these initilization methods yields the best performance.

After initiliziation, the algorithm will assign each point to the cluster it is closest to under the chosen metric. It will then recalculate each of the k means as the average of the location of the points in that cluster. It will repeat these two steps until the next clustering is the same as the previous one, indicating that the centroids will not change, and thus the clusters will remain the same as well. We can prove the convergence of this algorithm by simply noting that the total error must decrease with each iteration, and so reaching a point where the clusters do not change indicates that our error must be at a minimum.

3.1 Assumptions:

1. *Variance:* Assumes the variance of the distribution of the variables is spherical and that all variables have equal variance.
2. *Prior probability:* The prior probability for all k clusters are the same, i.e. each cluster has roughly equal number of observations[1]

3.2 Spotlight Evaluation:

Elbow Method: We created a curve which plots the inertia, or the sum of the distances of each data point from the center of its respective cluster, against the number of clusters when using K-means clustering. It is clear that as we use more clusters, inertia will decrease, but the goal of the elbow method is finding the point of concavity in this curve, as this is the point where adding more clusters is ”not worth” the corresponding decrease in inertia. In our case, this point happens at around 4 clusters, so we decided to cluster all of our data into 4 groups.

4 Results:

4.1 Link Analysis:

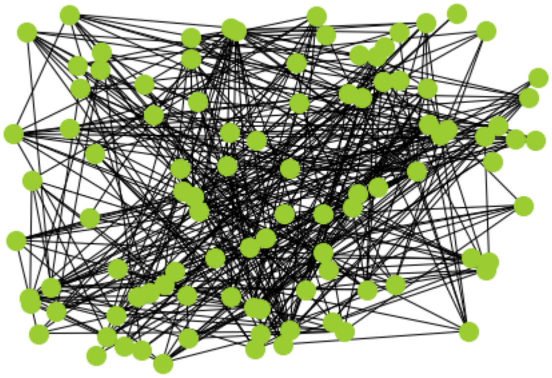


Figure 1: Eigenvector Centrality: Nodes = 100 Edges = 364 Average Degree = 7.28

For our three measures of centrality, we drew the graphs for the top 100 officers sorted by highest centrality. Figure 1 displays the results of eigenvector centrality. As expected, the average degree was 7.28 which was higher than the average degree for the entire graph. This makes intuitive sense since eigenvector centrality considers the nodes which themselves are connected to nodes with high centrality scores. This suggests that eigenvector centrality is a useful means of identifying officers who together account for a large number of the overall complaints. Once we scaled the graph up to 10% of all nodes (3576 nodes), 10% of the nodes accounted for 15.4% of the shared complaints, reaffirming the effectiveness of eigenvector centrality at identifying so-called ”bad apples”.

Figure 2(a) and Figure 2(b) display the results of degree centrality and Pagerank respectively. The average degrees for these graphs were lower than the average degree for the entire graph, suggesting that degree centrality and Pagerank might be less effective means of identifying ”bad apples” compared to eigenvector centrality. However once we scaled our graphs up to include 10% of all nodes, degree centrality accounted for 13.8% of complaints and Pagerank for 10.7% of complaints. This suggested that the two measures could also be useful at identifying the officers with most shared complaints, albeit to a lesser extent than eigenvector centrality.

We also sorted the officers based on the type of complaints against the officers. We did this for ”black”, ”hispanic” and ”punch/kick” complaints. We selected the top 1000 nodes and drew the graphs for officers with the highest number of such complaints. The data on the edges and average

[1] http://varianceexplained.org/r/kmeans-free-lunch/

(a) Nodes = 100 Edges = 50 Average Degree = 1.00 (b) Nodes = 100. Edges = 38 Average Degree = 0.76

Figure 2: Degree Centrality (a) and Pagerank (b)

Type of Complaint	Edges	Average Degree
“punch/kick”	526	1.052
“black”	442	0.884
“Hispanic”	440	0.880
“black” & “punch/kick”	536	1.072
“black” & “Hispanic”	440	0.880

Figure 3: Graphs for different race-based complaints, Nodes = 1000

degrees of each of these graphs is presented in Figure 3. We found it interesting that the average degree of the first 1000 officers with ”black” and ”punch/kick” complaints (500 nodes for each) was higher than both the average degrees for the first 100 officers with ”black” and ”punch/kick” complaints alone. This suggests that officers who are involved in racially motivated abuse are more likely to be involved in the same complaints as officers with physical abuse allegations, suggesting that these two types of abuse may go hand-in-hand.

4.2 Clustering:

After clustering officers into four clusters based on allegations as discussed above using K-means and BIRCH clustering, we decided to focus our analysis on the K-means clustering results (both of these cluster distributions are presented in figure 6), and we split the officers into groups based on their rank as well as their command. Within each one of these groups, the distribution of these clusters was examined and compared to the distribution of the clusters for the entire set of officers. If we assume that an officer’s rank and command are statistically independent from the type of allegations he or she is accused of (or in this case, the group based on allegations he or she is clustered into), then each conditional distribution of clusters should be almost exactly the same as the overall distribution. However, this is not at all what we observe. The distribution changes significantly with rank and command, and figure 5 (see Appendix) shows, for each rank and command, the sum of the absolute values of the differences between the proportions of clusters. This metric shows how strongly the given ranks and commands influence the type of allegations an officers faces, or in other words, how an officer’s position and direct co-workers affect his or her behavior.

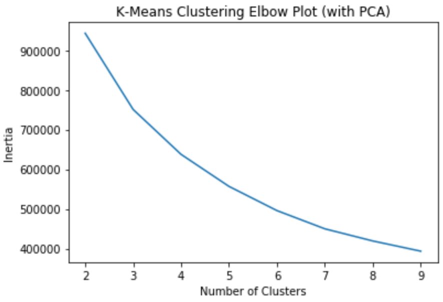


Figure 4: Elbow Curve

5 Discussion and Conclusion:

In this project we used data taken from the New York City Civilian Complaint Review Board (CCRB) to carry out network analysis and clustering in order to identify meaningful patterns within the data which may be useful in identifying ”bad apples” amongst active police officers. After running three types of centrality measures on a graph of officers and shared complaints, we found that eigenvector centrality was the most useful method for identified a large proportion of the ”bad apples” in the data set. By comparing average degrees of subgraphs of officers involved in the most race-based/physical abuse complaints, we also discovered that officers involved in racial abuse are more likely to be involved in the same complaints as physically abusive officers, suggesting that the race-based and physical abuse come hand in hand.

While examining how the distributions of allegation-based clusters of officers changed by rank and command, we found that an officer's position and command group played a large role in determining which cluster they would be placed into, indicating that officers' actions are strongly influenced by the actions of the officers surrounding them. This implied the existence of a sort of positive feedback loop within police forces. This means that it is not just the behavior of a few officers which must be fixed, but rather the behavior of all officers.

To extend our analysis, we could consider running various link prediction models on the data and checking whether the officers who we consider to be ”bad apples” would be likely to influence officers in their precincts to abuse their power more often. Our clustering analysis could be further extended by analyzing distributions on other factors, such as their salary and how it changes over time, and how frequently they experience changes in command and rank.

References

[1] Jennifer Golbeck *Analyzing the Social Web,* Morgan Kaufmann, 2013, Pages 25-44

[2] Priodyuti Pradhan, Angeliya C.U., Sarika Jalan, *Principal eigenvector localization and centrality in networks: Revisited,*. Physica A: Statistical Mechanics and its Applications, Volume 554, 2020

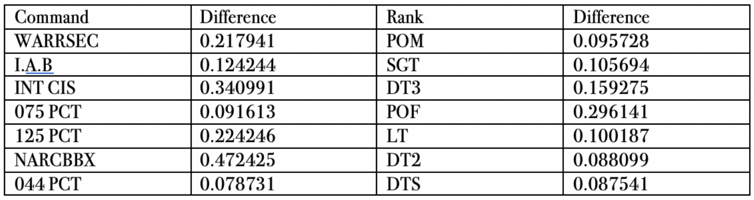
[3] K. P. Sinaga and M. Yang, *Unsupervised K-Means Clustering Algorithm,*. IEEE Access, vol. 8, 2020

[4] Chris Ding, Xiaofeng He, *K-means clustering via principal component analysis,* Proceedings of the twenty-first international conference on Machine learning (ICML '04),Association for Computing Machinery, New York, 2004

[5] D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, *The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News,*. 2018 International Seminar on Application for Technology of Information and Communication, 2018

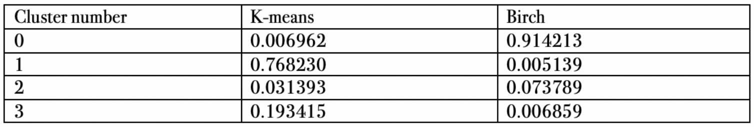
[6] Barbara E Engelhardt *Sample Homework.*

Appendices



Command	Difference	Rank	Difference
WARRSEC	0.217941	POM	0.095728
I.A.B	0.124244	SGT	0.105694
INT CIS	0.340991	DT3	0.159275
075 PCT	0.091613	POF	0.296141
125 PCT	0.224246	LT	0.100187
NARCBBX	0.472425	DT2	0.088099
044 PCT	0.078731	DTS	0.087541

Figure 5



Cluster number	K-means	Birch
0	0.006962	0.914213
1	0.768230	0.005139
2	0.031393	0.073789
3	0.193415	0.006859

Figure 6