Understanding and Predicting ”Bad Apples” and Wages Through Police Misconduct Complaints

Tiffany Huang
Princeton University
twhuang@princeton.edu

Alexis Sursock
Princeton University
asursock@princeton.edu

Abstract

How can we understand and predict policy brutality? This work is an exploration of patterns in a dataset of misconduct complaints from the New York Police Department (NYPD). We create a graph-based model representation in order to investigate the effect of ”bad apples,” or officers who are involved in other officers' misconduct complaints. We find that certain types of abuse are somewhat predictive of bad apples. Through Principal Component Analysis, we find that rank and number of complaints are key features for our dataset. Finally, in investigating officer wages, we find that offensive language is linked with higher salaries. Future work should focus on further exploring the relationship between allegation type and officer rank, promotion, and wage.

1 Introduction

Recent well-publicized instances of police brutality, such as the murder of George Floyd by police officer Derek Chauvin, have raised questions of how to understand the causes and motivations behind police misconduct. Does misconduct occur simply because some officers tend to be more violent (these officers are often referred to as ”bad apples”), or is there a culture of misconduct bred among certain police precincts?

In this work, we investigate these questions to identify key factors that lead to police misconduct. To this end, we combine two large datasets on the New York Police Department (NYPD) officers: the NYPD Misconduct Complaint Database[2] and the CapStat.NYC Police database [8]. We model police misconduct complaints through a graph model to find that 30 percent of NYPD officers are linked to one another (here, links represent complaints that involve multiple officers). We additionally find that officers with the most complaints tend to be ”lone wolves” who are not connected to many other officers.

We also identify certain features more prevalent in instigators (officers that inflict their partners in complaints) and attempt to predict them using a variety of models. Additionally, we employ unsupervised learning techniques such as Principal Component Analysis (PCA) and K-Means clustering to better explore the data. We find that the top two principal components represent rank and abuse type, respectively. We also find that officer wage is positively correlated with use of offensive language, which we encourage future researchers to explore.

2 Related Work & Motivation

The two main data sets used in this paper provide information on NYPD police officers' rank, wages, and number of complaints. We analyze the NYPD Misconduct Complaint Database, a dataset of 323,911 misconduct complaints filed against 81,550 officers in the New York Police Department, as compiled by the New York City Civilian Complaint Review Board (CCRB)[2]. We also incorporate the CapStat.NYC Police database, which contains the rank, district, wage and promotion history of 12,450 NYPD officers [8].

We are inspired by existing work on community network structures to create a graph-based model of NYPD officers where officers are connected by complaints they share with other officers. Our expectation is that we find that the graph is well-connected, which would prove that police misconduct is often influenced by group dynamics, and not just the result of a single officer's wrongdoing. We are also hoping to identify certain ”bad apples” that are connected to many other officers and prove that these officers are responsible for a disproportionately high number of complaints. We also wish to explore what factors contribute to a ”bad apple”; our prediction is that rank and number of complaints will be positively correlated with number of link connections (where bad apples are officers with many connections). We will use mean squared error to evaluate this method.

We also are interested in how performing Principal Component Analysis (PCA) and clustering will identify any patterns in the data. We hope these will help us identify significant features and/or group the data by similar features. We will evaluate PCA and clustering by the metrics of explained variance and inertia, respectively (these are explained more in Section 4.3).

Finally, we are interested in how officer wages relate to misconduct. We attempt to predict a particular year's salary using data such as past wages, rank, and allegations. We will use mean squared error to evaluate this method.

Because this is an open-ended exploration of a very large data set, we cannot predict how or what findings we will receive, if any. Our end goal is to be able to make a justified conclusion about a pattern we observe in the data.

2.1 Data Processing

We undertook significant work processing the CAPstat extended officer data set [8] and uniquely matching officers with their complaints from the CCRB (Civilian Complaint Review Board) database into a clean and combined sparse matrix.

Firstly, as no unique identifier was present on the officer data set, we had to resort to matching by full name. In order to uniquely identify officers we decided to remove all duplicate names and middle initials (not present in the CCRB data). We decided against matching duplicate names using their current county assignment as it predisposes an assumption of no changes in position throughout a career, which is unrealistic given that over 50% of officers live outside of their assigned county and frequently move precinct [9].

Next, we decided to solely focus on complaints issued between 2000 and 2018 , as CCRPB data has high duplicate and entry error rates for pre-2000 years. Additionally, as the CAPstat wage and rank promotion/demotion data is only available through 2018 we decided to focus on relations in the data only through that year, foregoing complaints posted in post-2018.

In total, we were able to uniquely identify 11,843 officers and 34,429 complaints matched to them. We encoded their ranks in order of NYPD rank structure and assigned each officer a count matrix of all allegation types against them and their complaint career total. Each officer has several clean time series data for the full 19 year time frame including promotions/demotions, rank level change, wages, and complaints.

Due to the significant effort and time devoted to efficiently and meaningfully combining these two data sets, we believe that our new data set provides outstanding opportunities for additional research and considerably facilitates the research process. We strongly recommend future researchers to utilize our data set (attached to this paper) for further analysis rather than starting with the aforementioned unprocessed databases.

3 Methods

To make a graph model, we use NetworkX [1]to create a graph where nodes are officers and edges are complaints where another officer is implicated. We then calculate the degree (number of edges) for each node and add this to our dataset of officer information. In order to predict apples, we employ LASSO feature selection on the dataset to identify which features contribute most (reasoning for using this is explained in the Results section). Additionally, we employ PCA and K-Means Clustering to explore the data. We utilize an inertia plot to identify the optimal number of clusters for K-Means Clustering. Finally, we explore wage prediction through a linear regression model (we choose this model because it is easily optimized using GridSearch, and it is a simple yet effective model for predicting continuous variables).

3.1 Spotlight Method: Principal Component Analysis

Principal Component Analysis (PCA) is a form of dimensionality reduction, or ”reducing the dimension of the feature space” through feature extraction [7]. Feature extraction works by creating linear combinations of existing attributes; this allows us to reduce the number of features we are working with without having to completely delete or eliminate any features. In addition, each of the new variables after performing PCA are independent of one another, which works well when using linear models which operate on the assumption that variables are independent of each other[7].

PCA can be useful in many situations [11]; for instance, PCA can help with data compression, so that one can store a data set in a reduced representation and save space. Additionally, it allows us to visualize high dimensional data by projecting our data set down to two or three dimensions.

We begin by choosing the number of desired components, or ”K”. PCA works by finding ”lines, planes, and hyper-planes in the K-dimensional space” that maximize the variance of the coordinates on the line or plane [4]. This intuitively makes sense; finding the directions that capture the most variability must mean that these directions encompass the most significant features in our data set (consider a feature that has the same value for every observation; this feature would have zero variance, reflecting that it is not an important feature).

We can perform PCA by calculating the covariance matrix of our data set, which represents the ”joint variability” of each feature with another [6]. We can then calculate the eigenvectors that correspond to the K largest eigenvalues of this covariance matrix. These eigenvectors correspond with the principal components of the dataset. We can think of PCA as a form of projection down to the K-dimensional space formed by the K principal components.

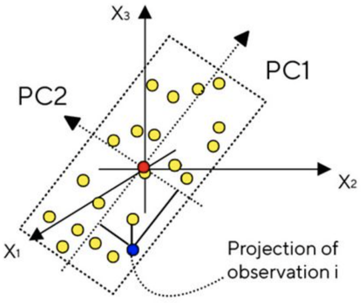

Figure 1: Visual representation of projection of observations down to a two-dimensional plane, courtesy of Sartorius [4]

One key note is that PCA just identifies the features that have the largest scale because these have a larger variance. Therefore, it is common practice to standardize the data set first so that all features are on the same scale before performing PCA analysis.

In our work, we begin with the sparse matrix created through processes outlined in the data-preprocessing section. Then, we standardize the data set by using StandardScaler(), which standardizes by removing the features' mean and scaling to unit variance [3]. We then choose our K; here, we choose K=2 as we are primarily using PCA for data visualization, and this works best when the data set is projected down to two dimensions.

4 Results

4.1 Partner Network Graph

Using CCRB data, we identified officers working together inflicted in the same complaint and call them ”partners”. We set out to evaluate the ”bad apples” hypothesis and see if certain officers increase the likelihood of their partner officers getting inflicted in a complaint. Using both NetworkX, we graph officers with their respective partners; interestingly, this results in a highly connected single large subgraph connecting over 22,900 of the about 88,000 officers in the data set, with all other clusters being smaller than 26.

We examine if officers in the main cluster can be considered ’bad apples’ and find that they (representing 28% of all officers) are responsible for only over 29.7% of all complaints. This shows that we are not able to uniquely cluster a majority of complaints with a small minority of officers and contradicts the controversial ’a few bad apples’ theory, which states that most police complaints can be traced back to a limited few, undermining the legitimacy of systemic issues [10].

We further set out to examine the top instigators in the dataset, by ranking by officers’ graph degree. We find that the 22 officers with the highest degrees are involved in 1.02% of all complaints (or over 3,000) across the board (as either a defendant or a partner). None of the aforementioned officers were terminated and a majority of them received promotions even after years of complaints against them.

Interestingly, in examining the top 10 officers with the most complaints, we find that only three of them have degrees of greater than 10. In addition, the top 3 offenders have degrees of less than 5. This suggests that the top offenders may be more likely to be ”lone wolves” who operate mostly alone, though this needs to be studied further.

4.2 Bad Apples Predictor

Due to the interesting nature of the NetworkX officer degree findings, we decided to add a ”degree” column to our officer data set. We define ”bad apples” as officers involved in 5 or more complaints with partners and aim to predict if an officer is a bad apple or not given their rank, wages, and allegation types over time. To avoid overfitting or ”leaking data”, we remove the total complaint count column and refrain from using any graph information in the prediction models.

We utilize Lasso feature selection due to its advantages over step-wise selection, mainly its penalization of the l1-norm which forces increased sparsity during variable selection. Our feature selection produced extremely high prediction values for the FADO allegation type features and the resulting model boasts an average 94.8% accuracy on our cross-validated testing set (with a randomized baseline of 92.1%). Notably, excessive ”force” and ”abuse of authority” were significant features for instigator prediction, whereas ”offensive language” and ”discourtesy” were less so. Further research is encouraged into why a harmful action might be more instigating than a negative demeanor.

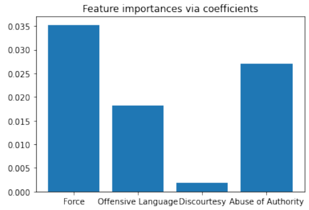

Figure 2: FADO Allegation type feature importances for instigator prediction (LASSO Feature Selection

Other features, including rank, wages, and promotions were not significantly more predictive than the random state. Several other attempts were made using a recurrent neural network, linear regression, and naive Bayes models; none were found to be especially effective. Our hypothesis that there might be a correlation between an officers’ negative influence over others and their rank and wages was thus falsified.

4.3 PCA and K-Means Clustering

In order to better understand and visualize our data, we use PCA through the processes outlined in the Spotlight Method section and identify top two principal components (PCs);here, we choose K=2 to allow for ease of visualization. Investigation into the top five largest values of these principal component vectors (meaning: which features contribute most significantly to the components) reveals that PC1 is all composed of features relating to an officer's rank in a given year, while PC2 is largely of features such as an officer's total number of complaints and the number of complaints they receive in each complaint type (the types being: Force, Abuse of Authority, Discourtesy, and Offensive Language).

One key note is that our top two principal components explain 17.3% and 7.9% of the total variance (in sum they explain about 25% of the total variance). This is a very low portion of the total variance, which suggests that we need many more components in order to accurately model the data. Further exploration shows that we need 81 principal components in order to explain 80% of the total variance; this implies that PCA with just two components is not an accurate representation of the data. Future work should explore the results of PCA when choosing more than two principal components. It is probable that PCA is not the best method of unsupervised learning for this dataset.

After reducing our dataset through PCA, we perform K-Means clustering as a means of better understanding how we can find patterns in the data. In order to identify the number of clusters to find through K-Means clustering, we plot the number of clusters (from 1 to 10) vs the inertia obtained through clustering with that amount of clusters. Inertia is defined as a measure that ”tells how far away the points within a cluster are” [5]. In order to minimize inertia while also minimizing number of clusters, we look at the ”elbow” of the inertia plot (appendix: 3)and find that it is at K=4 - this is our optimal number of clusters. We then plot these clusters (appendix: 4); further explanation of the clusters can be found in the appendix. An final note is that the inertia of our clustering is still quite high (around 100,000), which suggests that the data may not be easily cluster-able (this makes sense because we could not visually identify any clear clusters).

4.4 Time Series Evaluation

Using our own data set, we set out to evaluate any correlations between wages, promotions/demotions, and number of complaints.

We predicted wages using past wages, promotions, rank, and number and type of allegations. Our linear regression model, with optimized hyper-parameters using GridSearch cross validation, had a coefficient of determination of 96% with an RMSE of $4,400. Interestingly, previous year wages alone were only 67% indicative of current year wages. Using feature selection, given our data sets size and sparsity, we found that the FADO allegation type ”Offensive Language” was a strong positive indicator of wage (with a coefficient of $1,500 per past complaint), whereas ”Abuse of Authority” was a negative one. This indicates that officers with a ”harsher tone” might be rewarded by the department and poses ethical questions regarding systemic issues in police promotions.

5 Discussion and Conclusion

This paper's main work relates to creating a new and improved data set and experimenting with several different potential approaches to it. Our data set of over 11,000 police officers and nearly 200 features relating to them provides future researchers with extensive opportunities to explore hidden patterns in NYPD police complaints.

We find that the ”degree” component of instigators is especially promising and our attempt at predicting it has shown certain allegation types indicative of an officer's instigating nature. We found that PCA was not an extremely effective learning method for our dataset, but future work might focus on the benefits of PCA for a specific subsection of the data. We strongly encourage future researchers to evaluate, contrast, or forecast the time series data (wages, promotions, and complaint dates) found in our data set, to better understand its relationship with allegation type.

Acknowledgments

As always, we want to thank Dr. Li for supporting us on Ed and via email. We also want to acknowledge Deniz Oktay, Siena Dumas Ang, Yaniv Ovadia, and Sulin Liu for helping us in Office Hours (and often staying overtime!) and on Ed.

References

[1] Networkx. https://networkx.org/.

[2] Nypd misconduct complaint database. https://www.nyclu.org/en/campaigns/nypd-misconduct-database.

[3] sklearn.preprocessing.standardscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[4] What is principal component analysis (pca) and how it is used? https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186.

[5] Azika Amelia. K-means clustering: From a to z. https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a.

[6] Moshe Binieli. An overview of principal component analysis. https://www.freecodecamp.org/news/an-overview-of-principal-component-analysis-6340e3bc4073/.

[7] Matt Brems. A one-stop shop for principal component analysis. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c.

[8] CAPstat. Capstat.nyc police database. https://www.capstat.nyc/.

[9] David Cruz. A majority of nypd officers don't live in new york city, new figures show. https://gothamist.com/news/majority-nypd-officers-dont-live-new-york-city-new-figures-show.

[10] Malorie Cunningham. 'a few bad apples': Phrase describing rotten police officers used to have different meaning. https://abcnews.go.com/US/bad-apples-phrase-describing-rotten-police-officers-meaning/story?id=71201096.

[11] Barbara Englehardt. Lecture 16: Principal components analysis(pca).

6 Appendix

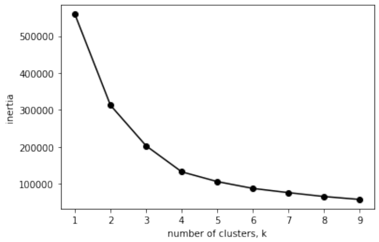

Figure 3: A plot of inertia vs number of clusters when performing K-Means Clustering. We identify the "elbow" to be where K=4.

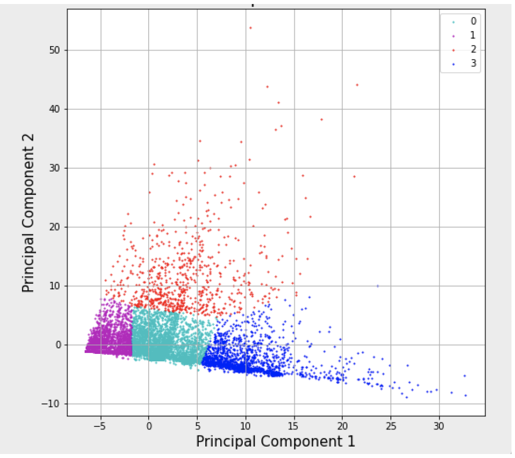

Figure 4: K-means clustering performed on our dataset after PCA. The axes represent the top two principal components. Here, cluster 1 (magenta) seems to represent low rank officers with a small number of complaints. Cluster 0 (cyan) represents medium rank officers with a small number of complaints. Cluster 2(red) represents low and medium rank officers with a high number of complaints. Cluster 3(blue) represents high rank officers, who mostly seem to have a very low number of complaints.