The Role of Networks in "Bad Actor" Identification: Informing Investigative Journalism

Krista Chan, Andres Nigenda, Elias Serrania

Background and Goals

Due to historically strained relationships between the residents of Chicago and the Chicago Police Department (CPD), many communities are invested in the investigation and prevention of police misconduct. The Invisible Institute is a local investigative journalism organization which has broken big stories on this issue, such as the cover-up of Jason Van Dyke's murder of Laquan McDonald.[1] We worked with journalists at the Invisible Institute to identify sources of data and a project which would support their mission to investigate and expose cases of police misconduct.

An issue within CPD which gained significant media attention after stories broke of multiple police officers involved in the coverup of Laquan McDonald's murder is the concept of a police "Code of Silence", which then Mayor Rahm Emanuel defined as "the tendency to ignore, deny, or in some cases cover up the bad actions of a colleague or colleagues".[2] The Invisible Institute's hypothesis is that this code of silence can be captured by conducting network analyses on CPD data.

Therefore, our project's main aim was not only to inform the Invisible Institute's investigative capabilities, but to test whether network features play an important role in identifying "bad actors".

Related Work

Outside of journalistic organizations, there have been police departments around the country investing in machine learning models to flag officers at risk of misconduct. One such example is the Charlotte-Mecklenburg Police Department which collaborated with the Center for Data Science and Public Policy to develop an early intervention system (EIS) to help target interventions such as counseling and training for police officers.[3]

In journalistic contexts, stories of police misconduct have typically been focused on individual "bad actors" within a force. In the field of criminology, network analyses have been used to research the proliferation of gun violence amongst civilian victims and perpetrators[4][5], but the same techniques are now being applied to police violence. Daria Roithmayr of the University of Southern California presents a theoretical framework for the spread of police misconduct through

[1] https://www.nytimes.com/2017/11/26/us/chicago-police-shooting-journalist-laquan-mcdonald.html
[2] https://theintercept.com/series/code-of-silence/
[3] https://journals.sagepub.com/doi/abs/10.1177/0887403417695380
[4] https://academic.oup.com/epirev/article/38/1/70/2754865
[5] https://link.springer.com/chapter/10.1007/978-0-387-77650-7_11

networks[6], suggesting that police officers are more likely to use excessive force after witnessing other officers do the same through observational learning and peer influence.

In a recent paper entitled “The Contagiousness of Police Violence”[7], the authors construct networks of CPD using citizen complaints and examine police misconduct as a “contagion” which spreads through the aforementioned networks. They examine the connections and clustering of officers within the networks, finding that shooting officers are more clustered together than non-shooting officers and that periods between shootings are shorter for more connected officers. This paper was foundational in helping us form our models, and is perhaps the most relevant current research being conducted in evaluating the effect of networks on police misconduct.

Problem formulation

This project’s goal is to identify which police officers will have a sustained complaint or use a firearm as a proxy for police misconduct, and to test whether network features improve model performance. To do so, we make a point of comparing feature importance across models, especially between regular models and network models, so as to isolate the added value of network features.

Data description

In order to generate our features, we use the Invisible Institute’s Citizens Police Data Project (CPDP), which includes data on police officers (gender, race, rank, birth year, salary, appointment and resignation dates), investigators (gender, race, officer id), allegations (location, date, complaint type, lat/lon, category, final finding, final outcome), complainants and victims (gender, race, birth year), and tactical response reports (TRRs). We also use the City of Chicago’s crime data portal data on reported crimes in Chicago in order to create features that adjust for the level of crime in police beats.

In the years 2010 to 2017, we have data on a total of 50,087 complaints, involving 9,953 officers. 3.5% of these complaints were sustained. In the years 2010 to 2016, we have data on 35,296 TRRs involving 7,932 unique officers. 515 TRRs involved use of a firearm by 453 unique officers (1.5%).

Analysis

Since the goal of our machine learning pipeline is to identify which police officers will have a sustained complaint or use a firearm as a proxy for police misconduct, and to test the hypothesis that networks are helpful in doing so, our analysis and evaluation reflect this.

For our model, we used features related to the officers themselves (gender, race, age, etc.), their record (number of complaints, number of victims, tactical response reports, whether or not

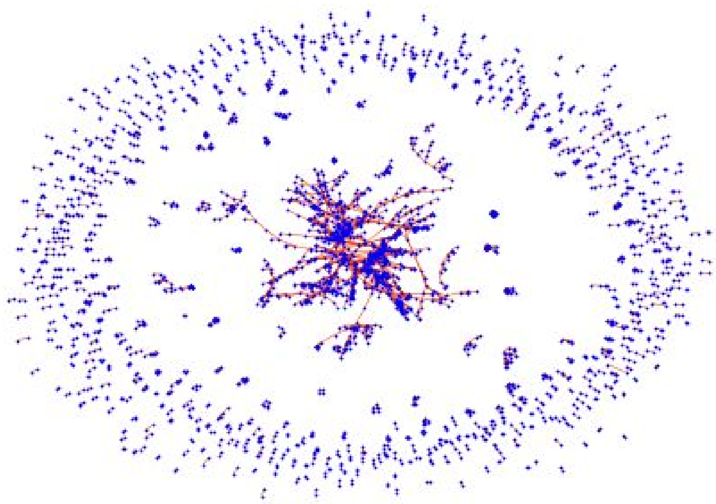
---

[6] https://legal-forum.uchicago.edu/publication/dynamics-excessive-force
[7] https://www.law.uchicago.edu/files/2018-11/chicago_contagiousness_of_violence.pdf

the officer has been previously disciplined, etc.) and network features that explore the relationships among police officers (length of shortest path to an officer who has previously discharged a firearm or had a sustained complaint, among others).

This network of police officers was built by taking into account two kinds of relationships between the officers: coaccusals and investigations. Using the networkx package, we drew a network that contained relationships between officers that had been previously co-accused in a given complaint and officers that had investigated complaints from other colleagues. The resulting network of police officers is shown in Figure 1.

Figure 1. Network of police officers



We created four temporal sets, each with four different windows: a training window, a training outcome window, and a testing window. Each set was split as follows:

1. Trained on 2010 to predict outcomes in 2011 and tested on outcomes in 2012
2. Trained on 2010 - 2011 to predict outcomes in 2012 and tested on outcomes in 2013
3. Trained on 2010 - 2012 to predict outcomes in 2013 and tested on outcomes in 2014
4. Trained on 2010 - 2013 to predict outcomes in 2014 and tested on outcomes in 2015

To prevent leakage, all continuous features were created on data only within the training period. These features were then scaled between 0 and 1 using the MinMaxScaler object from the scikit-learn package with the training period, and applying the same scaler to the remaining windows. All categorical features were "dummified" in our training data and then applied in the same way in the outcome window and test windows. If any categories did not match between the two periods, we only kept those in the training window.

In order to identify the best performing model, we ran models on four time sets, three types of features (regular, augmented and network), two outcomes (will the officer have a sustained complaint?, will the officer use a firearm?) and 7 classifiers (Random Forest, Gradient Boosting, Extra Trees, Ada Boost, Decision Tree and Logistic Regression). The combination of the aforementioned elements resulted in 4,680 models.

Evaluation

Since we framed our “policy intervention” to be journalistic investigations of police officers, we focus on our models’ precision at 5% That is, we are concerned with our model’s ability to identify police officers with the top 5% risk of engaging in police misconduct, so journalistic investigation can focus on investigating and keeping tabs on a smaller subset of police officers.

Table 1 presents the maximum precision at 5% by Train-Test set and outcome. For the “sustained” outcome, augmented models performed better than regular models on all time sets. For the “firearm” outcome, the augmented model performs at least as well as the other models. Therefore, these findings supports the hypothesis that network features help in identifying police officers that are likely to engage in misconduct, at least when misconduct is defined as a police officer’s likelihood to have a “sustained” complaint, which we think is a reasonable indicator of the police officer’s misconduct, as a low number of allegations are investigated and an even smaller number of complaints are sustained.

Table 1. Precision at 5% by Train-Test set and outcome

Set	Type	Sustained	Firearm
0	augmented	0.454	0.043
	network	0.305	0.021
	regular	0.433	0.043
1	augmented	0.519	0.019
	network	0.315	0
	regular	0.463	0.019
2	augmented	0.563	0.01
	network	0.354	0.01
	regular	0.469	0.01
3	augmented	0.533	0.013
	network	0.28	0.013
	regular	0.467	0.013

We defined the “best” model as the model that had the best precision at 5% over Train-Test sets, favoring consistency over time. That is, a model that improved or maintained high precision over time was preferred over models that loss precision with larger train sets. Figure 2 shows the performance of all models across Train-Test sets for the “sustained” outcome; the blue dot with label 63 is a Random Forest classifier with parameter set 63: {'criterion': 'entropy', 'max_depth': None, 'min_samples_split': 10, 'n_estimators': 100, 'n_jobs': -1, 'random_state': 84}. This model fit all of the criteria in order to be classified as our “best model”. This model was the fifth best model in set 0, the best in set 1 and 3, and the second best in set 2.

Figure 2. Precision by augmented model, sustained outcome

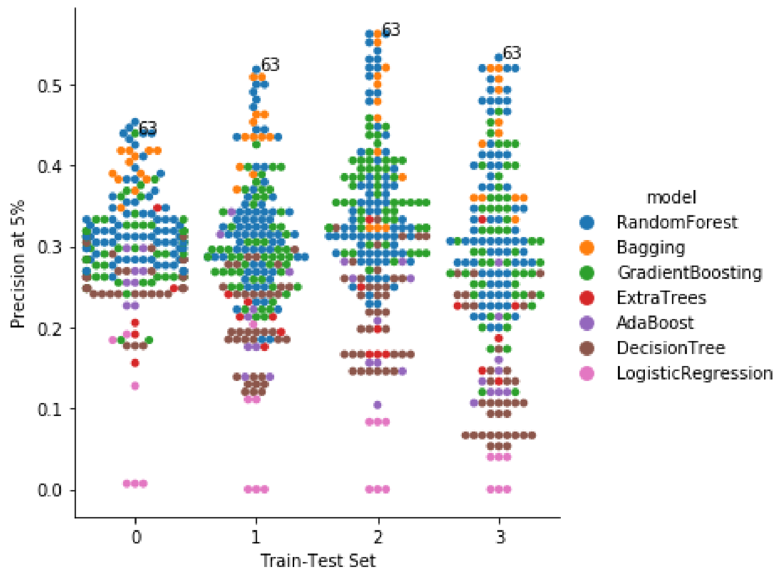
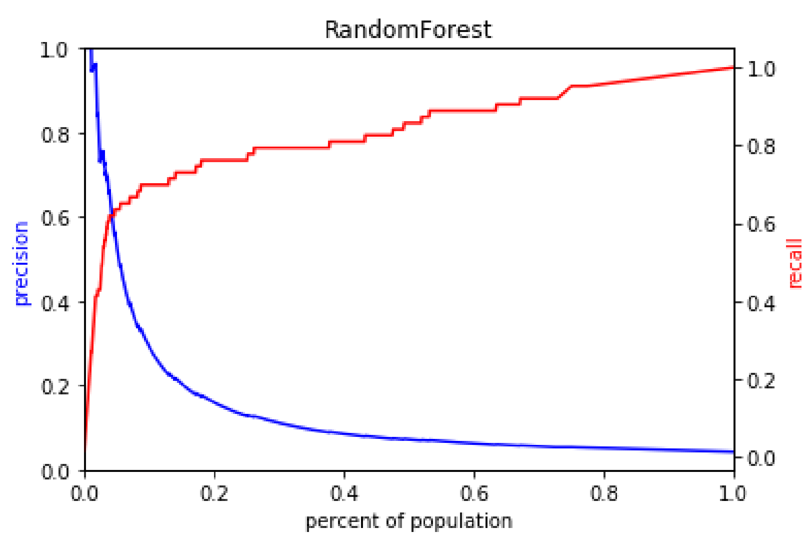


Figure 3 shows the precision-recall plot for our “best” model on Train-Test set 3. The “best” model is very precise at identifying “bad actors”, and performs well at the top 5%, having a relatively high recall too.

Figure 3. Precision-recall for “best” model, set 3



Discussion

Table 2 presents the feature importance for the top 10 most important features of the “best” augmented model for each type of specification in Train-Test set 3. Column (2) of Table 2 shows that the officer’s age, tenure, shortest path to an officer with a sustained complaint, their number of victims and number of complaints are the top 5 important features that help in predicting an officer’s likelihood of having a sustained complaint against them. Columns (3) and (4) show the importance of the features when the models are run without network features and with network only features respectively. Overall, two network features are in our model’s top 10 features (out of 184 features total).

Table 2. Feature Importance by Model, top 10

Feature (1)	**Augmented (2)**	**Regular (3)**	**Network (4)**
age	11.6	17.94	
tenure	9.47	14.42	
shortest_path_below_four_sustained	7.02		44.34
victim_count	4.9	6.65	
number_complaints	4.11	6.16	
shortest_path_below_four_shooting	3.89		20.27
trr_report_count	3.54	5.75	
pct_black_victims	3.44	5.62	
pct_sustained_complaints	2.47	2	
pct_white_victims	2.05	3.31	

These results identify the main features that help in identifying “bad actors” and also make a case for the use of networks in order to improve targeting.

Policy Recommendations

The Invisible Institute should investigate how specific police officers’ networks have influenced the sustained complaints and use of firearms outcome. The identified features are important in prioritizing the Invisible Institute’s investigations.

With our model, the Invisible institute can target their research efforts on a specific subset of officers and characteristics. This model should be trained with more data as it becomes available, and the appropriateness of maintaining the “best model” should be evaluated by comparing its performance to other models that result from our machine learning pipeline. Furthermore, taking our fairness and bias results into consideration, the Invisible institute might want to run separate models for different sets of officers (ie. race specific models), and validate if the feature importance remains similar or if it differs greatly.

Ethical Issues, Bias and Fairness

Auditing issues of bias and fairness are essential to our project since officers of different identities, based on categories like gender or race, may have different experiences in police networks. This may lead to disparate outcomes in false positive rates or false omission rates. The consequences of these disparities may lead to one group being unfairly investigated more frequently or less frequently than others.

Figure 4. False Positive Rate

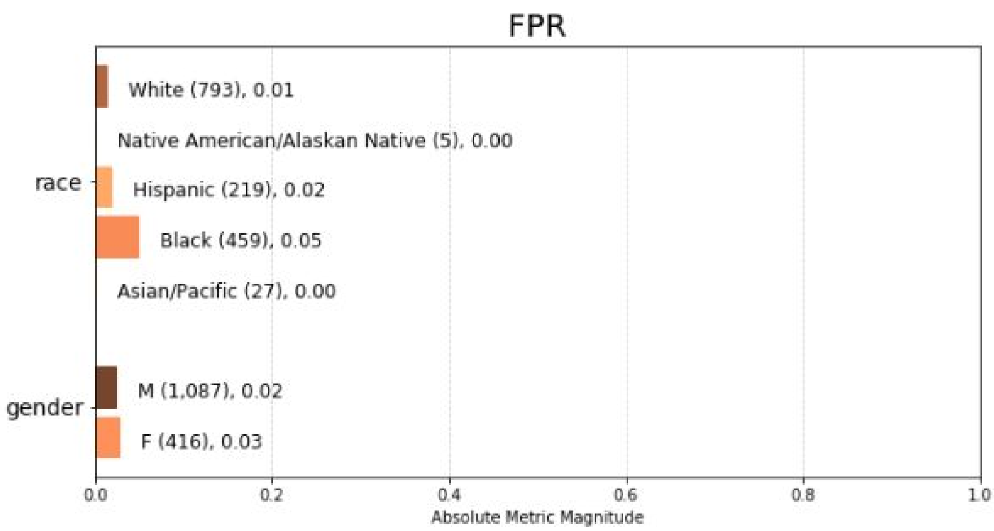
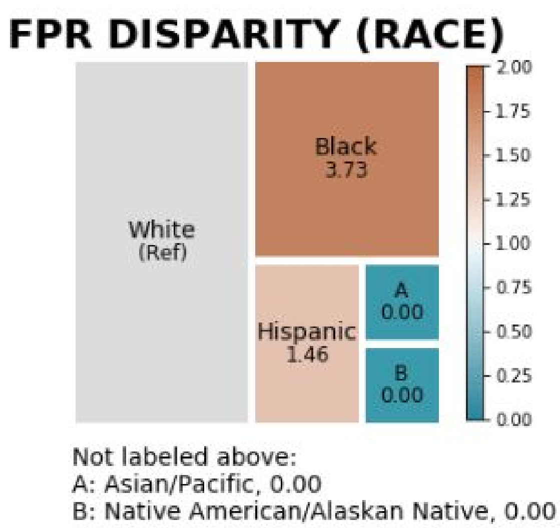


Figure 5. False Positive Rate Disparity



Based on results from our "best" Random Forest model, Black officers had the highest false positive rate, which can lead to disparate outcomes compared to their white peers.

Figure 6. False Omission Rate

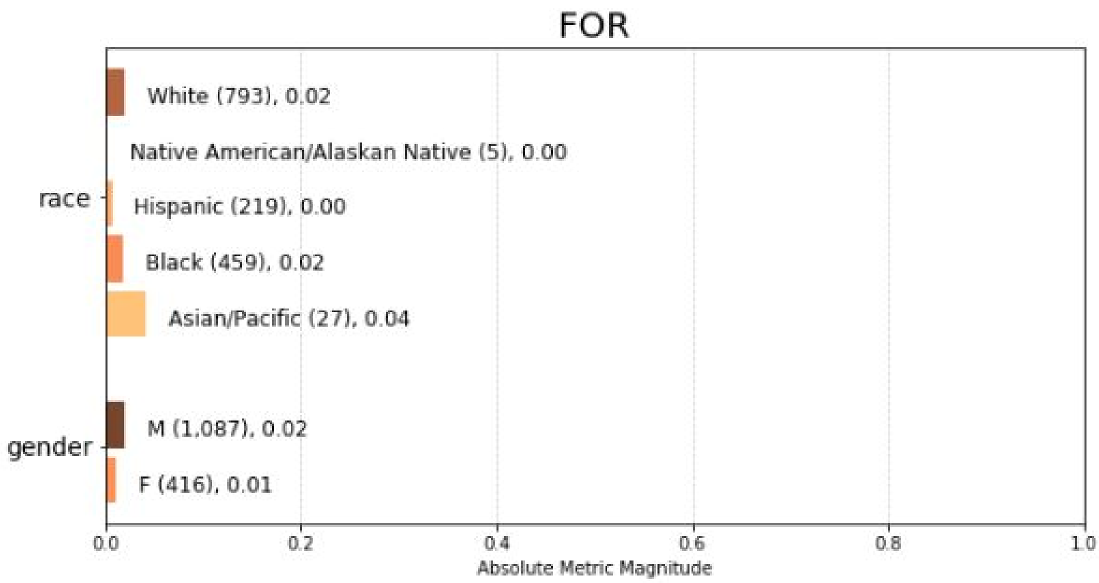
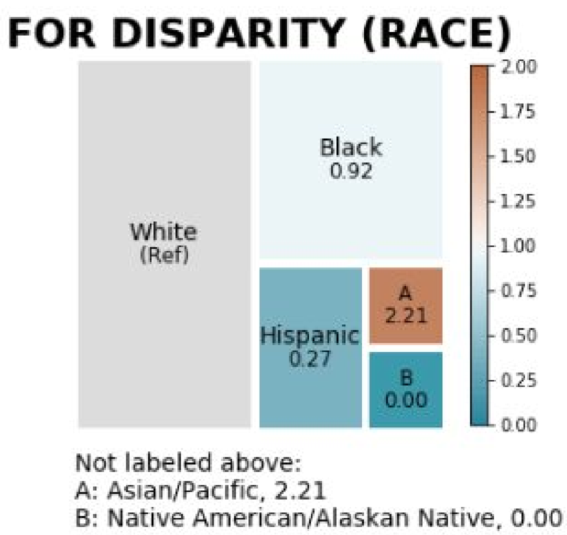


Figure 7: False Omission Rate Disparity



From the same model, we see higher false omission rates for Asian/Pacific officers, meaning that among predicted negatives, Asian/Pacific officers were more likely to be false negatives. This may be driven, in part, by the relatively small sample size of Asian/Pacific officers in our data.

Our results may warrant exploring other models with lower false positive rates for black officers or lower false omission rates for Asian/Pacific officers in order to avoid unfairly targeting officers for investigation. It could also be valuable to examine feature importance for different categories of officers in order to gain qualitative insight into how officers are impacted differently as a medium for better informing how officers are investigated.

Limitations

As in any model stemming from a machine learning approach, the user of this model should make it a point of feeding the model new data as it becomes available in order to timely capture changes in the data that might end up making another model the “best” model over time. In addition, our current labels are only proxies for “police misconduct”, which should call to contrast results from more than one proxy for police misconduct.

In order to carry another robustness check on our results is to run our pipeline with an even bigger parameter grid. Our model also has room for improvement, particularly in the addition of other features that might be considered important for classification of “bad actors” (ie. complaints that would have been filed if not because of the affidavit policy, officer’s history of stops and arrests).

Finally, the results stemming from our “best” model should incentivize a qualitative exploration of the reasons behind the observed feature importance, so the interpretation of the model can be further grounded in reality.