Analyzing Officer-filed Allegations

Chanh Bui, Ethan Kohrt, Noah Shen

Introduction

When thinking about allegations against police officers, most people would immediately think of them as complaints from only civilians toward officers for misconduct. Surprisingly, a noticeable number of these allegations are actually filed by other officers: approximately 6.22%, totaling 13,444 internal allegations.[1] Thus, it is very interesting to dive deeper into these internal allegations to explore differences from civilian allegations and their potential implications.

For our project, we want to ask several questions about internal allegations, ranging from differences between internal and civilian allegations to the influence of these allegations on the officers. First, we want to know whether internal allegations have more impact on the officers than civilian complaints. If they are more impactful, diving deeper into internal allegations is even more important as it can give us insight into why they are so. Our next question is whether a high rate of internal complaints indicates accountability inside a police unit. This is interesting because we hypothesize that if we can see some negative correlation between civilian complaints and internal allegations, that would suggest accountability. Finally, we want to investigate the potential of predicting future civilian allegations from internal allegations. Such a predictor can help spot our potential future misconduct and help focus on those flagged officers to prevent future civilian complaints.

Features of Officers Listed in Internal Complaints

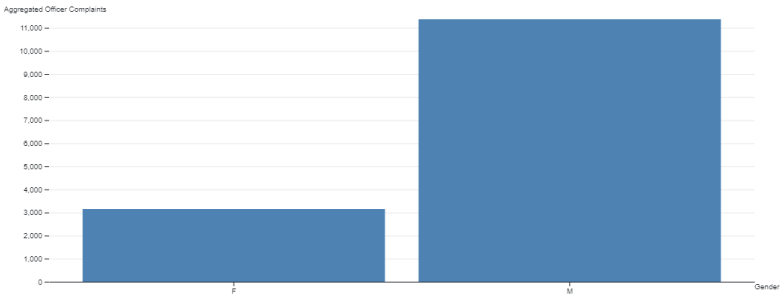
Checkpoint 3.2 was an interactive visualization[2] that enabled us to get insight into some of the distribution of demographics within the group of officers who had accumulated internal allegations. We selected a total of 6 features that we wanted to investigate including age, gender, years of service, race, number of civilian allegations, and award count. The officers were binned into their respective categories where their total number of internal allegations was aggregated and plotted. Below is an example histogram showing officers sorted by gender. From this chart, we can clearly see that males in total accumulate more internal allegations than females.

---

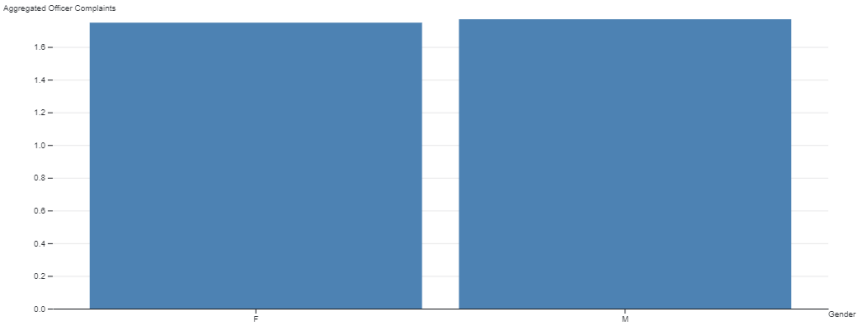
[1] We use the phrases “police-filed allegation” and “internal allegation” interchangeably throughout this report, as well as the terms “allegation” and “complaint.”

As a side note, our findings will not be expressed chronologically per checkpoint, as the narrative flows better when the results from our checkpoints are shifted around. We will make sure to cite our checkpoints so that our readers can follow along.

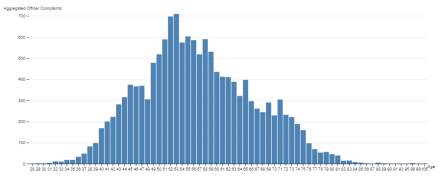
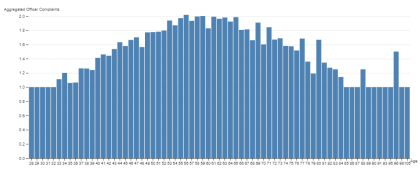
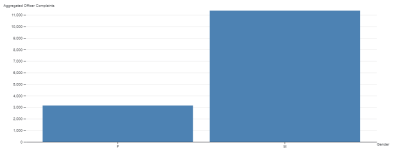
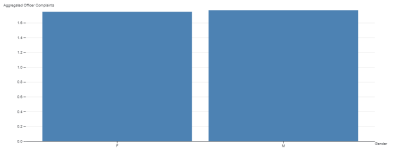
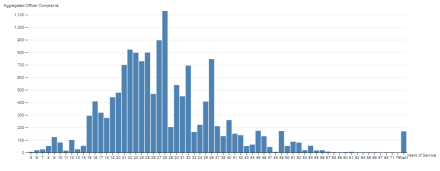
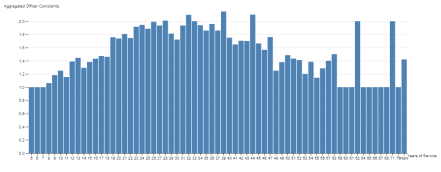
[2] https://observablehq.com/d/691e08eae899d1e5

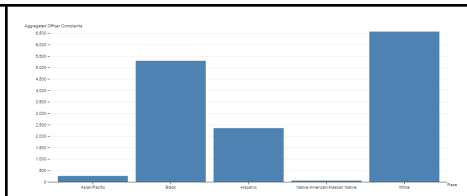
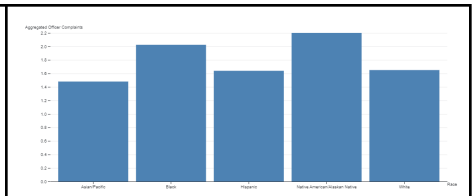
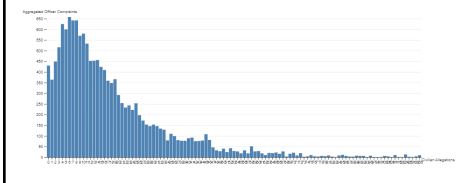
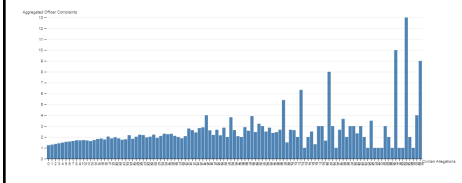
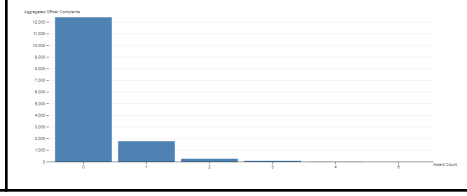
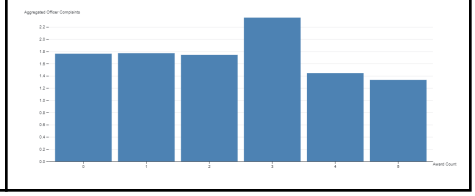


What we soon realized was that it was not fair to solely aggregate over categories. In the example above, while males make up the majority of internal allegations, males also make up the majority of the Chicago Police Department. To get a better idea of the average internal allegations of a specific category it was necessary to normalize based on the number of police officers. This gives us our new graph below which shows that on average, male and female officers accumulate the same number of internal allegations.



For the rest of the categories, we can display them below with unnormalized on the left:

Categories	Unnormalized	Normalized
Age - The birth year of the officer subtracted from 2022. The dataset did not include specific birthdays so we had to normalize age.		
Gender - The gender of the officer		
Years of Service - Some officers did not have a recorded appointment date, so they were marked as NaN		

Race - The race of the officer. This feature was highly exclusive in that no officers could identify as multiple races, for example, Asian and white.		
Civilian Allegations - How many civilian allegations a police officer has accumulated up to the present day		
Award Count - How many awards the office has received up to the present day		

From these charts we can see that only age and years of service maintain their approximate peak across both their unnormalized and normalized versions. We will use these visualizations to support our next checkpoint in machine learning.

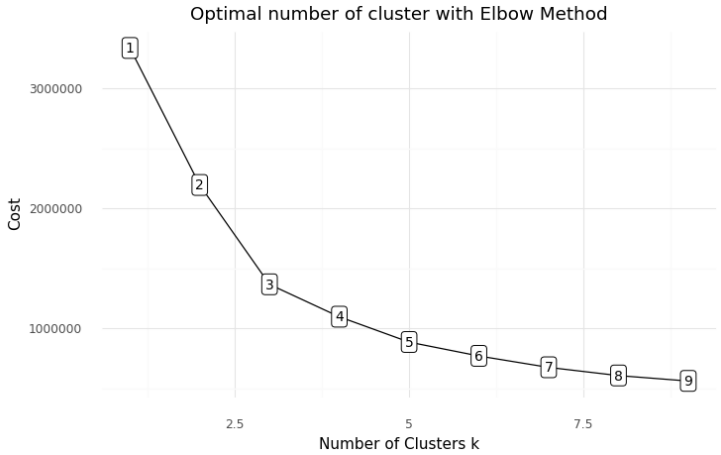
We also would like to acknowledge some potential pitfalls with our visualization and data collection. One of which was that we classified officers based on their current-day statistics. What this means is that for example, an officer may have accumulated all their internal allegations when they were in their 20's but may be in their 50's as of 2022. It could be argued that they should be put into the 20's group since that is when they were written up, and would also remove the need for normalization within these time-sensitive categories. However, we found that it was very hard to classify officers to allegations as some incidents did not provide an exact date, thus creating bias once again. For this reason, we choose to go with a more generalized approach.

Checkpoint 4.2 served as a supplement for Checkpoint 3.2. In our interactive visualization, the user was able to explore the different features themselves to create their own analysis. In this checkpoint, we apply those same features to a clustering ML model in order to gain further insight into how officers can be classified. We used k-Means clustering which is an unsupervised learning model. This means that our model attempts to learn from untagged data while being fed in features to determine their labels. k-Means works by taking in data points and a number of groups and, in short, takes the distance between points as a measure of similarity in order to identify clusters.

One caveat to k-Means that more experienced ML users may question is that k-Means is typically only applied to continuous data. However, with the addition of features such as gender and race, the base k-Means algorithm fails to work. For this purpose, we used a Python library

called “k-Modes” which movies the k-Means algorithm to accommodate for categorical variables (https://github.com/nicodv/kmodes).

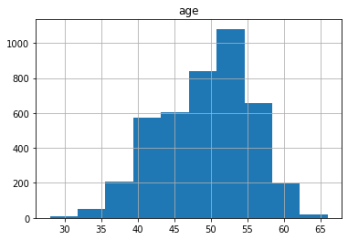
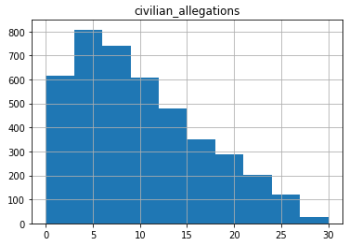
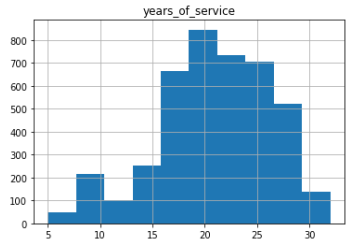
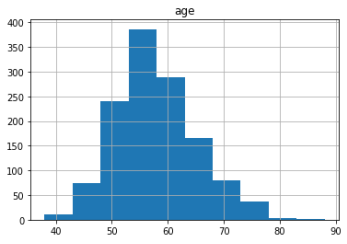
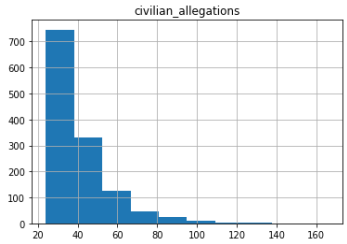
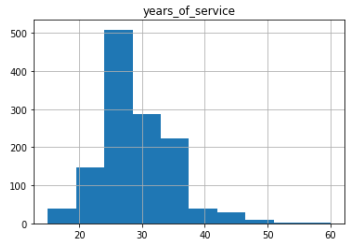
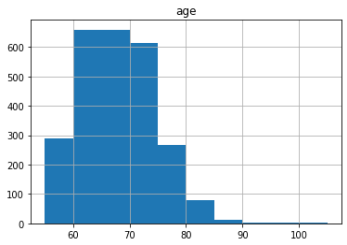
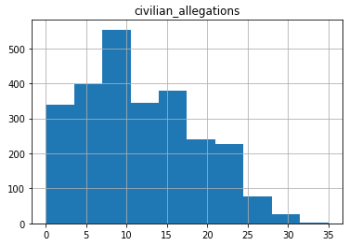
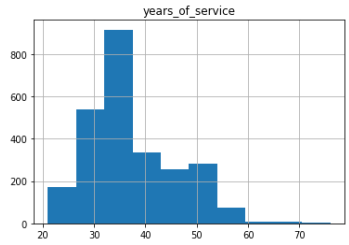
The most important part of the k-Means algorithm is arguably selecting a good k, cluster, value. To do this we tried k-Modes on our 6 features for all officers with an internal allegation for k values between 1 and 9. Using the elbow method, we found that the optimal value for k was 3. In short, the elbow method is a heuristic that is used to determine the number of clusters within a set by noticing where the cost function tapers out. From the photo below, we can notice how the stark change in cost dramatically slows down after k=3, thus we settle on that value.



With the value for k chosen, we can now apply the k-Modes algorithm to our dataset and begin clustering the officers. We can apply chi-squared contingency to our result in order to determine the importance of features in determining the clusters. Chi-squared contingency works by testing if there is a relationship between a set of variables. In other words, it tests for the independence of variables. A higher chi-squared value indicates that a feature was more important for clustering. These results are shown below:

	chi2	p	df
featureage	6350.184718	0.000000e+00	136.0
featurecivilian_allegations	7357.797890	0.000000e+00	212.0
featureyears_of_service	5939.654431	0.000000e+00	126.0
featureofficer_allegations	369.019143	6.971831e-58	34.0
featuregender	231.554412	5.231164e-51	2.0
featurerace	255.823765	1.003186e-50	8.0

We can notice that the k-Modes algorithm found that age, number of civilian allegations, and years of service were the most important features when clustering. This is interesting because we can relate these results back to our interactive visualization and notice that these are the same features that maintain the same trends with and without normalization. This is in contrast to gender and race which we saw were more or less equivalent after normalization despite the majority of officers being white and male. We plotted the three important features for each of the clusters below for further analysis.

Cluster	Age	Civilian Allegations	Years of Service
1			
2			
3			

Looking closer at these histograms, we can notice that all three clusters have radically different distributions of officer demographics. Age, being the most important feature per the chi-squared testing, has clear distinguishability between young, middle-aged, and old officers. From there a viewer can scroll right and see how their civilian allegations and years of service differ as well.

Internal Allegations as a Predictive Tool

One question we want to answer relating to our theme is whether we can use these internal allegations as a “canary in the coal mine” for future civilian complaints. If these internal allegations are predictive of future misconduct, we can create a predictor using this information to identify potential officers who might receive future civilian complaints and focus more attention on them. Such practice can help prevent misconduct and help fewer complaints toward the office force.

Specifically, we want to know whether the officer will get a civilian complaint in 2 years using their current internal allegation and general information about them. In order to do this, we want to train several classification models to find the best one and see if it is a good model. If we

can find such a good model, we will have supporting evidence for our “canary in a coal mine” hypothesis, and vice versa.

Our training data has 21 features. Categorical features are turned into one hot encoding as preprocessing. These features can be divided into 2 categories:

- Allegation-related features: recommended outcome, final outcome, disciplined, category, days experience at time of incident, and age at time of incident.
- Officer-related features: gender, race, percentile, salary and count features such as complaint percentile, internal allegation percentile, trr counts, etc.

Our target is whether the alleged officer of the given internal allegation received another civilian allegation in the next 2 years after the incident date of the allegation.

We look for our best model by training 4 different models with its best-tuned hyperparameter using a cross-validation grid search. The 4 models are logistic regression, decision tree, random forest classifier, and gradient boosting classifier. Here are our best models.

Model	Accuracy	Hyperparameter
Logistic Regression	78.562%	C = 6235, Penalty = l2
Decision Tree	81.025%	Max depth = 5, min sample leaf = 0.001, min sample split = 0.067
Random Forest Classifier	80.122%	Max depth = 21, min sample leaf = 0.01, min sample split = 0.01, n estimator = 100
Gradient Boosting Classifier	**81.173%**	Max depth = 1, min sample leaf = 0.01, min sample split = 0.01, n estimator = 200

From our training, our best model is gradient boosting classifier with approximately 81.173% accuracy and F1 score of 0.44. For our confusion matrix below, we name Class 1 as Having a civilian complaint in 2 years, and Class 2 as not having a civilian complaint in 2 years. Looking at our recall statistics, the recall for both class 1 and class 2 are decent (68% and 83%). This is a good statistic as if we used this as a predictor to pay attention to potential misconduct officers, this suggests that more than two-thirds of officers with future civilian complaints will get flagged by the predictor. We also have significantly high precision in class 2 (95%). This suggests that if we do not do anything to the non-flagged officers by our predictor, only 5% of them would receive civilian complaints in 2 years. The only low statistic is the precision on class 1 (33%). However, we would consider this acceptable and less important if we want to use this system to flag officers with potential future misconduct. This is because we would rather flag less risked officers than not flag those that are likely to receive civilian complaints. This is similar to the case where we would rather flag people without cancer rather not flag people with cancer in a cancer test. Overall, with the general accuracy of 81% and decent class recall, we are confident that this is a good model for predicting future civilian complaints for the next 2 years, suggesting that internal complaints can be predictive of future civilian complaints and can potentially be used for flagging future misconduct.

Truth data

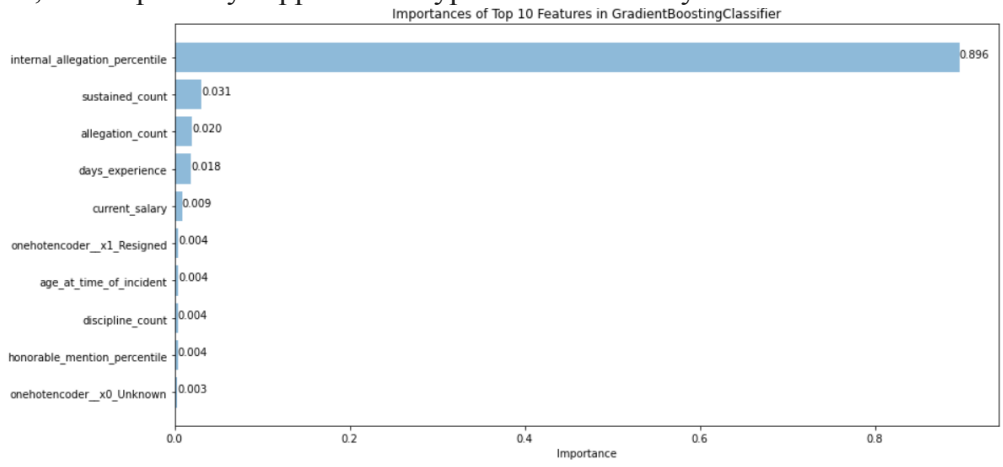
Classifier results	Class 1	Class 2	Classification overall	User's accuracy (Precision)
Class 1	356	716	1072	33.209%
Class 2	167	3450	3617	95.383%
Truth overall	523	4166	4689	
Producer's accuracy (Recall)	68.069%	82.813%		

Overall accuracy (OA):	81.169%
Kappa[1]:	0.349

-

Confusion Matrix (Class 1 = Have a civilian complaint in 2 years,
Class 2 = not have a civilian complaint in 2 years)

We are also interested in the importance of features in our model. Many of the features are not very significant so we decided to graph the top 10. Because our model is a gradient boosting model, the size difference between feature importance is just an artifact of the calculation method, but the relative order of importance still stands. Since we have a variety of features that are important here, our predictions are not entirely dependent on a few features. Finally, we also notice that the internal allegation percentile is the feature with the highest importance, suggesting that internal allegations are generally predictive toward future civilian complaints, which partially supports our hypothesis of "canary in a coal mine."



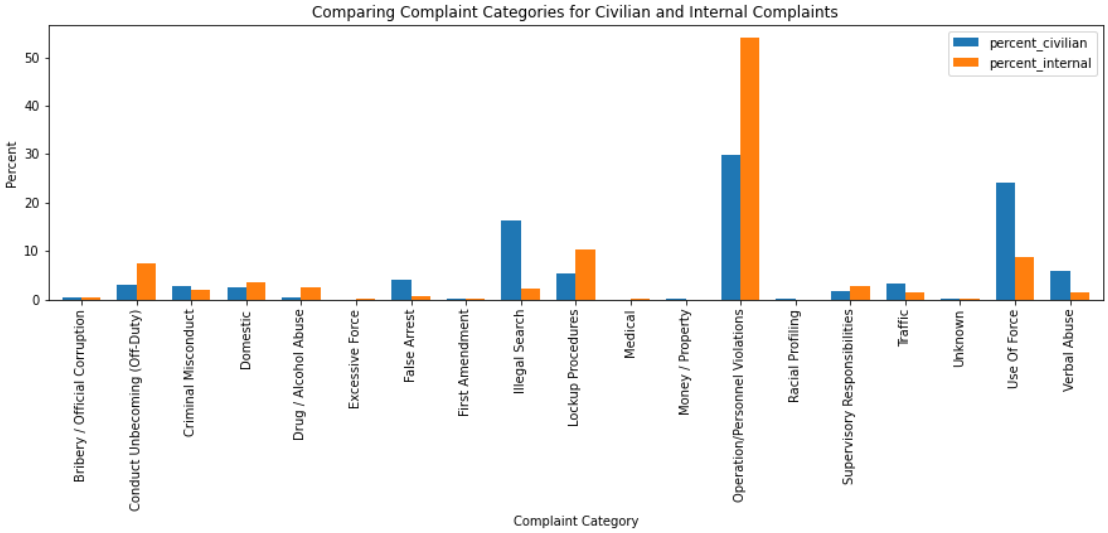
Comparing Civilian and internal Complaints

We then investigate the differences between misconduct complaints filed by civilians and those filed internally by police officers. Certainly, some civilian and internal complaints refer to the same instance of misconduct, but 94% of all recorded complaints are filed by civilians, indicating that civilians file complaints far more often than officers do. If internal complaints are

meant to be a mechanism for police units to hold themselves accountable for misconduct, this discrepancy is already a piece of evidence that it is not very effective, as an ideal system of accountability would show an internal complaint for every civilian complaint, plus more for instances of misconduct not seen by the public.

Perhaps internal complaints are reserved for only the most egregious instances of misconduct. If this were the case, we might see that the complaint categories for internal complaints would skew toward certain offenses that come up often in civilian complaints, such as use of force or illegal search (keeping in mind that police interaction of any kind can greatly disrupt someone's life, whether or not violence is involved).

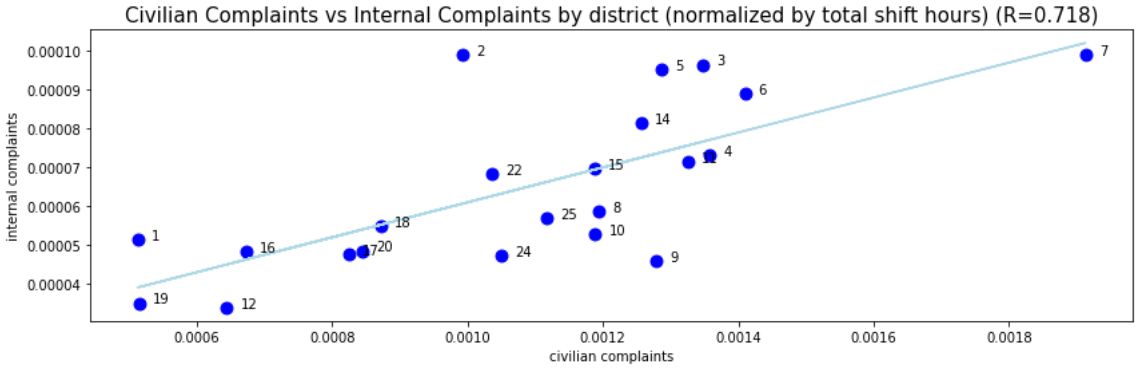
However, we do not see this relationship in the data; instead, more than half of internal complaints (54%) fall under the category 'Operational/Personnel Violations,' a broad category that includes all kinds of workplace misconduct such as lateness, failure to submit forms, insubordination, and sexual harassment. The overall mismatch in category distributions is striking; as the figure below shows, internal complaints tend to focus more on workplace violations and procedural failures, while civilian complaints tend toward more conspicuous types of misconduct, like use of force, illegal search, false arrest and verbal abuse. It is clear that the two mechanisms of complaint, civilian and internal, are used in vastly different ways. This mismatch may reveal that police officers have different priorities when it comes to reporting misconduct, or differing views on what constitutes misconduct.



To further explore the differences between civilian and internal complaints, we created an interactive visualization[3] that breaks down the data for each police unit in Chicago. In it, we can see that the complaint category distributions of individual units largely reflect the overall trend. We can also compare the rates of civilian and internal complaints by unit, and see how each rate evolves over time. Here there is no visible relationship between rates of internal complaints and civilian complaints over time, again casting doubt on whether internal complaints indicate accountability in a unit. If problematic officers were being disciplined internally at high rates, we would expect to see civilian complaints decline, but this does not appear to be the case.

[3] https://observablehq.com/d/664d92cd192aa9f1

Similarly, plotting each unit's rate of complaints shows that rates of civilian and internal complaints are related positively, even when each police unit is normalized by the number of shift-hours worked. This is the opposite of what we would expect if high rates of internal complaints leads to lower civilian complaints. There are likely other confounding factors involved in this relationship, like the fact that some instances of misconduct produce both a civilian complaint and an internal one, so this would require further investigation. But given that these rates are normalized by total shift hours, we leave it as an open question whether units with low rates of each type can be considered 'better' than the other units in terms of misconduct.



It is interesting to note that internal complaints appear to be somewhat more impactful on an officer's career than civilian complaints. Looking at whether an officer resigns within two years of a complaint, 9.59% resign in two years after an internal complaint, while 6.48% resign after a civilian complaint.

Conclusion

We have shown that officer-filed complaints are innately different from civilian complaints. Of all complaints, 94% are filed by civilians, and the categories or reasons for the complaints differ dramatically. Furthermore, no evidence was found that higher rates of internal complaints correspond with lower civilian complaints. This suggests that internal complaints alone are not a sufficient mechanism for police accountability. Officers who receive high numbers of civilian complaints also receive high numbers of internal complaints, and information about an internal complaint can be used to predict future civilian complaints with some accuracy. Internal complaints also have a higher 2-year resignation rate compared to civilian complaints, suggesting that internal complaints have a greater impact on an officer's career.

However, some open questions remain; for example, why would internal complaints be more impactful on an officer's behavior? If we had information about the officers who actually file each complaint, we could discover deeper insights about the dynamics within police units. And does a high rate of internal complaints in a police unit indicate high accountability or general dysfunction? Perhaps a truly accountable unit should have low rates of both types of complaints, but this depends on how one defines "accountability." We hope that internal complaints continue to be studied in the future for their valuable insights.