CS496. Prof. Jennie Rogers. JaCoya Thompson, Lianke Qin, and Victor Bursztyn.

Data Science Seminar — Project Report

1. Introduction

Minority residents of Chicago are up to 14 times more likely to be targeted for excessive and deadly force, according to the latest statistics [1]. Seven black people have been killed by police in 2019, and the fight for justice for people of color does not end after Jason Van Dyke. Our group attempts to investigate how socio-demographic features such as race and ethnicity could play a role in the outcomes of police misconduct investigations, rather than focusing on the occurrences of misconduct in isolation.

To start, we retrieved some descriptive statistics as baselines: 8.8% of allegations made against cops are sustained complaints. This means the complaint, resulted in an investigation and the allegation is supported by sufficient factual evidence as well as a violation of policy. The average investigation time of allegations from the initial complaint date to the end of the investigation date is 267 days, or 8.7 months. Once an allegation has been made, it is reviewed by an investigator to ensure accuracy and consistency, which may include interviews with the parties involved, a review of police department records, and domain-specific analyses.

For that 8.8% of cases, the officer is given some type of disciplinary action. The top 3 most common types of disciplinary actions are "Unknown", "Reprimanded", and "1-Day Suspension." "Unknown" is the highest outcome which is often linked to settlements made between the police department and complainant. When a settlement is given, the average payment to victims is $195,054.33.

These numbers motivate our overarching research question: When a case of police misconduct victimizes groups that are historically marginalized, how different is its investigation?

To answer it, we analyze four hypothetical effects: (1) Is the allegation made by a victim of color less likely of being sustained? (2) Does victim race seem to influence investigation time? (3) Does victim race seem to influence the disciplinary action ultimately taken? and (4) Does victim race seem to influence the compensations granted by the state? In short, we are examining the presence of significant racial disparities in the outcomes of investigations involving people of color.

2. Theoretical Background & Methodology

It's challenging to use observational data to study causality. For instance, we could find that a particular socio-demographic feature is highly correlated to more severe cases of police misconduct because — the police would argue — "they are more exposed to violence." However, if we include certain lurking variables in our analyses, we can show that investigation outcomes vary significantly even after controlling for these factors, and raise a discussion. For this reason, we think it's valuable to include allegation type in our analyses besides the victim race and, of course, the decision-maker who can be proven racially biased. Therefore, the following variables form the basis of our analyses:

- Three independent variables: *victim race* (what we are testing as discriminant variable), *allegation type* (our lurking variable), *a decision-maker* who can be proven biased (either an investigator or a judge, depending on the effect being predicted); and
- Three target variables: *investigation time*, *disciplinary action*, and *settlement amount*.

Our analyses and results, in turn, are organized along three technical axes:

- First, we delve into the distributions of these variables, starting with simple histograms of the target variables and moving to more dynamic visualizations where victim race can be used to interact with the data.
- Second, we try to predict each target variable as a function of victim race, allegation type, and the decision-maker (either an investigator or a judge). In this context, our predictive analysis has two objectives: (1) to fit models that are better than naive baselines; and (2) to compare features w.r.t. their predictive power.

  Because of this dual-objective and the simple design of our experiment, we opted to use Decision/Regression Trees and interpret feature importance by looking at their information gains (i.e., the levels of the trees). Fig. 1 shows how our models consider the independent variables

victim race, allegation type, and the decision-maker, and then our effects as the dependent variable being predicted (either investigation time, disciplinary action or settlement amount).

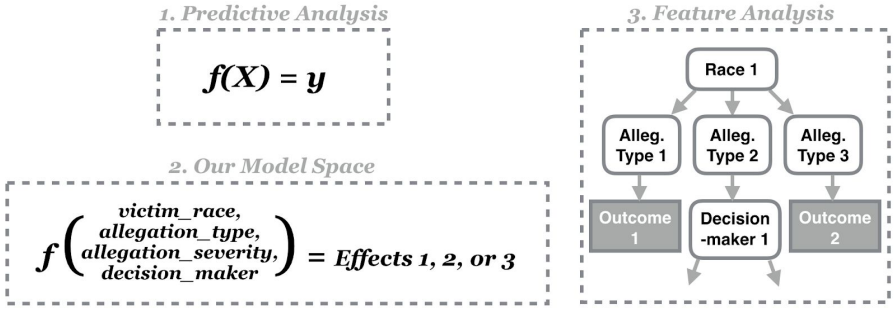


Fig. 1 — How our models will consider three independent variables: victim race, allegation type, and a decision-maker and the effects as the dependent variable, either investigation time, disciplinary action and settlement amount.

- Third, we use graph analytics to show that allegations involving victims of color tend to exhibit more violent patterns of police misconduct, which highlights the question of whether we have just outcomes across different segments of the population.
  To do that, we construct a graph of co-occurring allegation types for the entire CPDB dataset. Nodes represent allegation types (e.g., 'Intoxicated On Duty', 'D.U.I. On Duty') and edges represent allegations committed by the same officer consecutively (e.g., 'Intoxicated On Duty—D.U.I. On Duty'). Edge weights represent the number of occurrences of a pattern for the population under analysis. *Fig. 2 below shows how to construct this graph.*
  Next, we construct two different versions of this graph: one for allegations involving black victims only and one for white victims only. To compare how allegation types co-occur for these different segments of the population, we run a community detection algorithm [4], evaluate the quality of the clustering by its modularity value [5], and analyze the resulting communities to discuss any disparities.

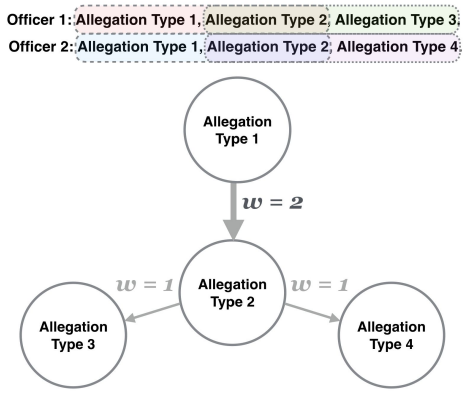


Fig. 2 — A graph where nodes are allegation types and edges represent co-occurring allegation types in an officer's career.

3. Descriptive Analysis and Data Visualization

We start with Fig. 3: a bar chart displaying the percentage of allegations sustained across victims races. We observe Asian/Pacific Islander (8%) and White (9%) to have the highest percentages of allegations sustained with Black (2%) and Hispanic (3%) the least likely to have a complaint sustained. This is our first indication of the potential influence of race in the outcomes of investigations.

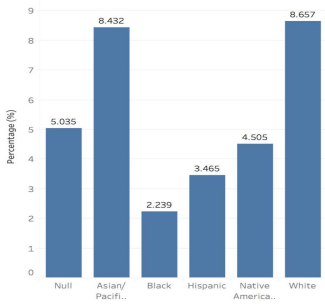
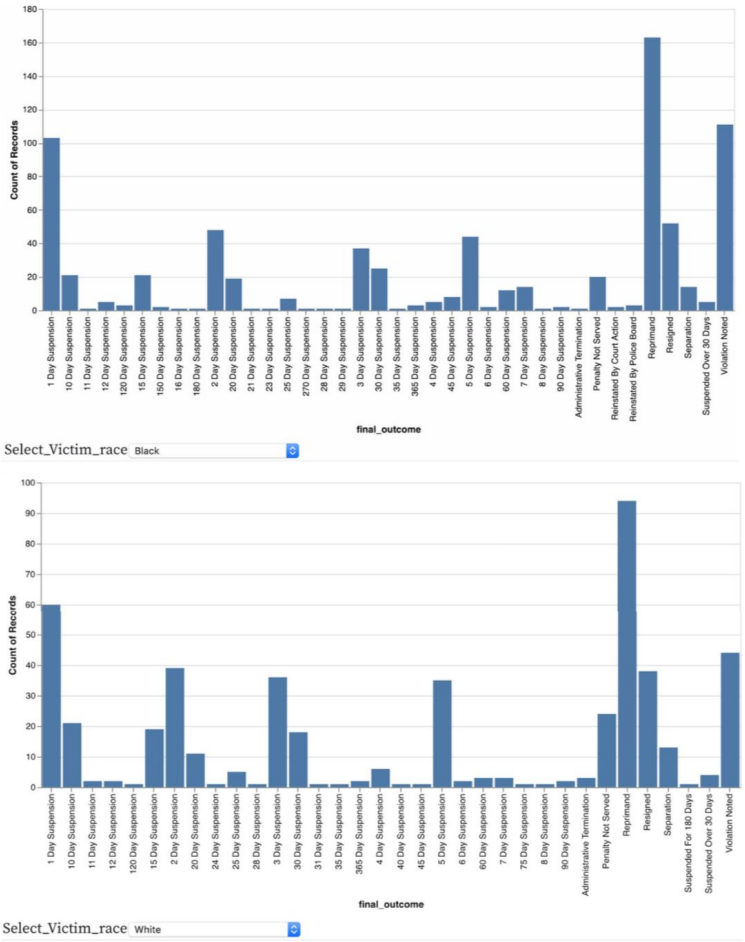


Fig. 3 — Resulting bar graph displays the percentage of allegation sustained across victims races.

To observe the distributions of investigation time and disciplinary action based on race, we created an ObservableHQ notebook[1] powered by Vega-Lite [2]. While we can't see any remarkable differences in the global distribution of investigation time, some findings for disciplinary action stick out. Since "No Action Taken" is a majority class in this set of nominal values, our first attempts to visualize racial disparities were slightly overshadowed by this high-occurring mode. For this reason, we repeated the visualization after filtering both "No Action Taken" and "Unknown" from our dataset.

From this new standpoint, we could see distinctions more clearly (Fig. 4). Although "Reprimanded" became the new upper bound in both segments of the population, black and white victims, we can see that several mid-severity disciplinary actions are relatively more frequent among investigations of white victims. For instance: white victims and black victims have roughly the same absolute amounts of "15 Day Suspension" and "10 Day Suspension", despite the population of black victims being much larger. In general, compared to black victims, white victims seem more likely to have investigations leading to "x Day Suspension" with x > 1 (see "10 Day Suspension", "5 Day Suspension", "3 Day Suspension" and "2 Day Suspension").



[1] Fully available at https://observablehq.com/@vbursztyn/interactive-visualizations-of

Fig. 4 — Influence of victim race on the distribution of disciplinary action (for better resolution, refer to our notebook).

To conclude our visual exploration, we tap into the settlements data. After assembling a table with victim race, allegation type (allegation's primary cause), case's judge, and settlement amount, we found black victims to have the most settlements (Fig. 5). "Asian/Pacific Islander" have the highest average of settlements with $985,874.80, followed by "Hispanic" ($213,197.00), "Black" ($192,927.00), then "White" ($105,735.00). We analyzed the average compensation amount by victim race and allegation type combined and started to observe a few patterns. The highest count of allegations were false arrest of black people, with an average payment of $37,454.00.

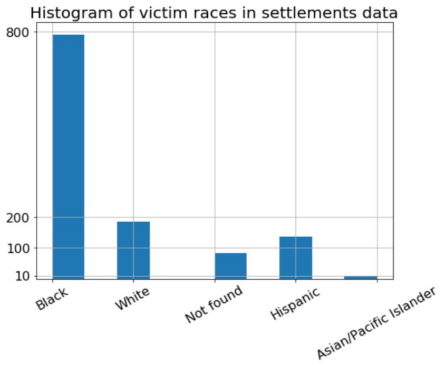


Fig. 5 — Histogram of victim races in the cases found in the settlements data.

4. Predictive Analysis and Machine Learning

Once this overview is made, we can tap into the Machine Learning toolbox and frame our research question as three relatively simple prediction tasks: Can we predict investigation time, disciplinary action, and settlement amount as a function of victim race, allegation type, and the decision-maker assigned? Furthermore, can we inspect our models to investigate signs of racial bias not only in the global level (which we partially covered with our ObservableHQ notebook) but also locally in this feature space?

4.1) Predicting **investigation time** based on victim race, allegation type and case's investigator

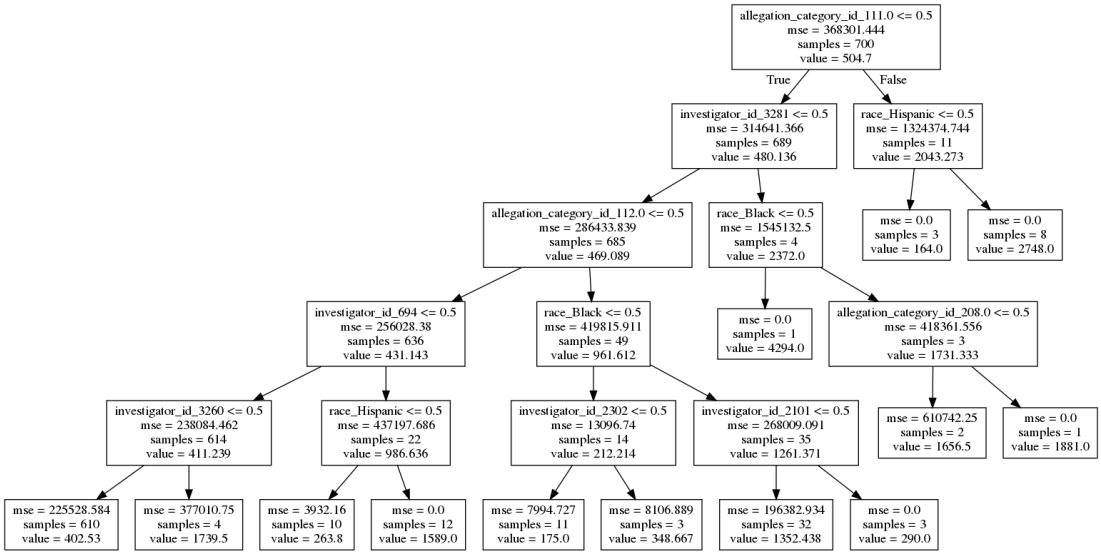


Fig. 6 — A Regression Tree for investigation time. Considering the boolean decisions in SKLearn's notation, when a decision is false it takes the left path; when a decision is true it takes the right. For instance, race_Black <= 0.5 means that race_Black is zero (False).

Model	**Training RMSE**	**Test RMSE**	**Test R^2**
Investigation Time	107	291	**0.774**
Settlement Amounts	140842	221814	**0.854**

Model	**Training accuracy**	**Test accuracy**	**Test F1 Score**
Disciplinary Action - All Classes	0.918	0.903	**0.914**
Disciplinary Action - Relevant Classes	0.895	0.586	**0.589**

Model	**Weight (W) for Race**	**W for Allegation Type**	**W for Investigator/Judge**
Investigation Time	0.09	0.38	**0.52**
Disciplinary Action	0.11	0.33	**0.54**
Settlement Amounts	0.08	**0.57**	0.34

Table 1 — A Regression Trees measured on both RMSE and R^2 (above); Decision Trees measured on both accuracy and F1-score (center); and global feature importance for each target variable: investigation time, disciplinary action, and settlement amounts.

For all Decision/Regression Trees, 70% of the data is used for training while 30% is held-out for testing. Tree height is limited to 5 to prevent overfitting which, in the case of Decision Trees, is when all leaf nodes end up with only a couple of examples (see values for "Samples" in the leaf nodes in Fig. 6 and see its caption for further details). As we can see in Table 1, our R^2 is fairly good, meaning that there's a general correlation between our feature space (the three independent variables) and the variable being predicted.

However, Table 1 also shows that, among the three features, the investigator is by far the most important in determining the pace of an investigation, followed by the allegation type. On one hand, it's an optimistic finding that race is far behind in its capacity to systematically predict the duration of investigations. On the other hand, even marginal effects should warrant further study. When we look at the lower-left of our Tree in Fig. 6, we can see "race = Hispanic" being very divisive in the general work of investigator #694. In practice, for this investigator, he had 12 investigations whose victims were Hispanics taking on average more than 1500 days (!), while the 10 cases not involving Hispanics were resolved in one sixth of the time on average. The fact that allegation types — an anticipated confounder — is included in our feature space and wasn't picked up by the model at that level of the Tree should warrant further scrutiny on investigator #694. Similarly, other more local patterns can be studied using this methodology, especially given that the model has a fairly good R^2.

4.2) Predicting **disciplinary action** based on victim race, allegation type and investigator

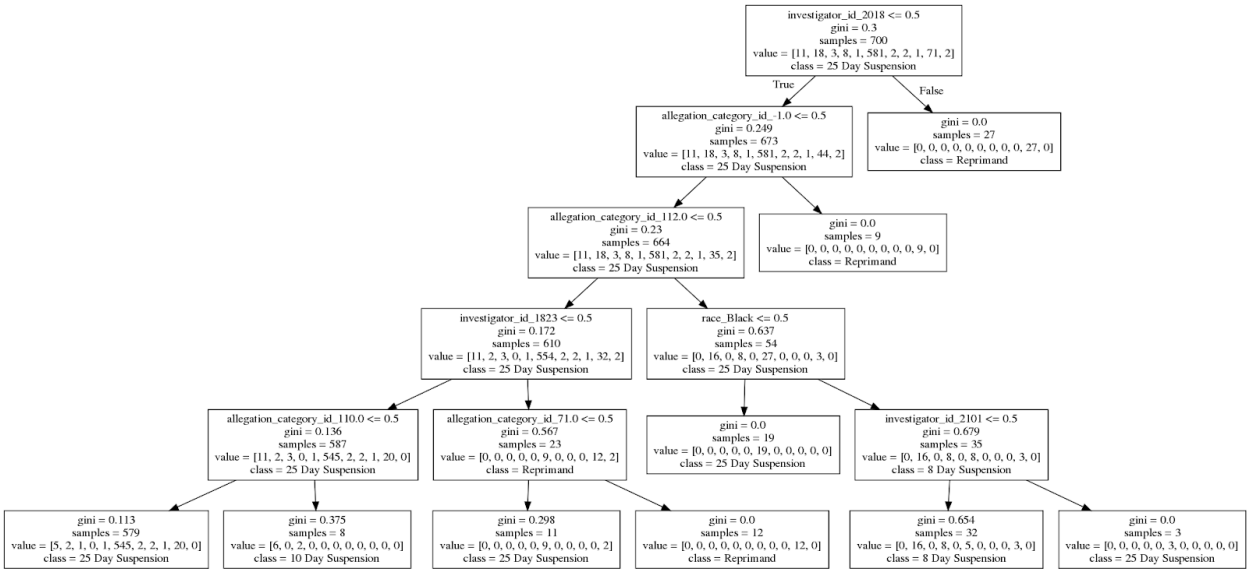


Fig. 7 — A Decision Tree for disciplinary action. Considering the boolean decisions in SKLearn's notation, when a decision is false it takes the left path; when a decision is true it takes the right. For instance, race_Black <= 0.5 means that race_Black is zero (False).

Training and testing were performed very similarly to 4.1, except for the target variable being categorical, thus the use of a Decision Tree instead of a Regression Tree. At first sight, Table 1 shows excellent F1-score for the actual phenomenon, i.e., predicting a disciplinary action; but a deeper investigation shows that this is greatly due to class imbalance, as "No Action Taken" and "Unknown" are disproportionately the most frequent outcomes. So, instead of predicting this phenomenon in its original form, we changed our definition to make sure that our model is navigating the hardest cases where disciplinary actions actually happen.

Table 1 shows a whole different picture: while the model tends to fit well the training data — as expected with Decision Trees and binary features — it doesn't seem to generalize that well, as seen in the test error being significantly higher. However, it's worth noting that, for the number of classes for this problem (i.e., all possible disciplinary actions), our Tree performs much better than a random guesser, indicating that there's signal in the data. Again, if we look into the summary of feature importance (Table 1), we can see a hierarchy that's very similar to 4.1. However, once again we can see that local patterns can still co-exist: "race = Black" at the fourth level of the Tree is somewhat divisive for the amount of days given as a suspension. But surprisingly, according to this path of the Tree, investigator #2101 seems to give 25 Day Suspensions for allegations involving black victims and 8 Day Suspensions otherwise.

4.3) Predicting **settlement amounts** based on victim race, allegation type and assigned judge

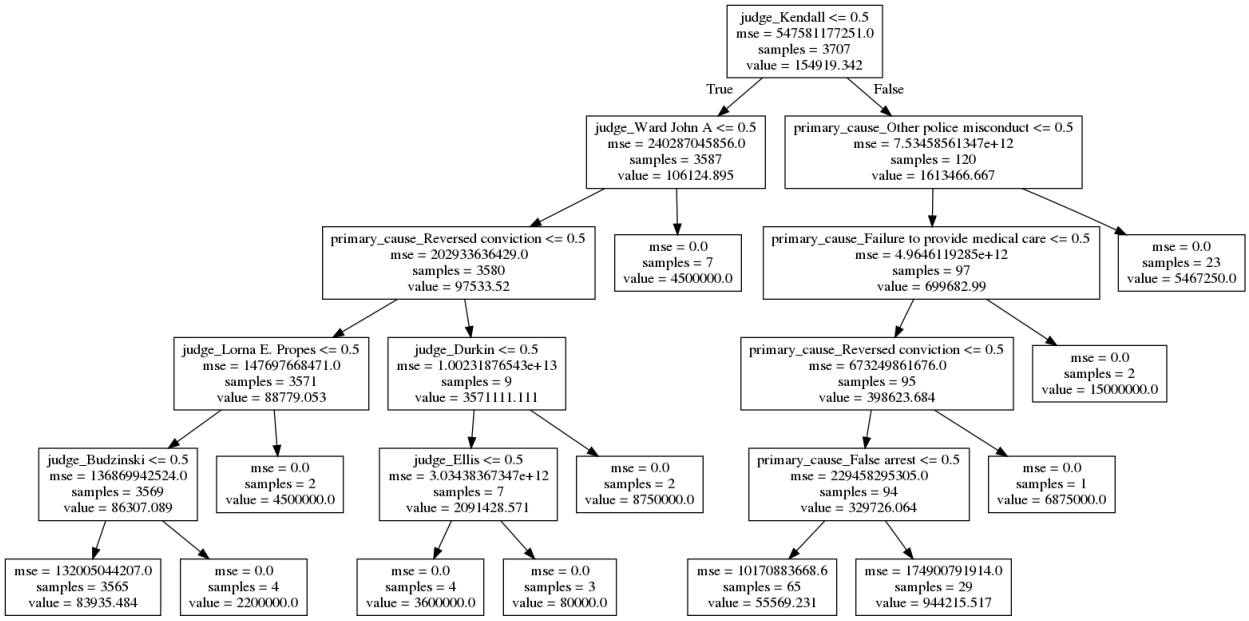


Fig. 8 — A Regression Tree for settlements. Considering the boolean decisions in SKLearn's notation, when a decision is false it takes the left path; when a decision is true it takes the right. For instance, race_Black <= 0.5 means that race_Black is zero (False).

Training and testing were performed very similarly to 4.1. However, in comparison to 4.1, settlements seem to have a higher variance than investigation duration, which makes modeling it even more challenging. Like 4.2, we do have some imbalance due to the fact that many settlements are denied (equal to zero), but not to the point that we need to refactor the problem.

Given all the difficulties, having an R^2 of 0.854 is surprisingly good. In an error analysis, we note that predictably the highest absolute errors stem from the cases with highest settlement amounts. We don't have enough examples to reliably model settlement amounts. The high-variance tied to the sample size would, in practice, make this problem very uncertain and any model error-prone in practice.

Even then, we can see in Table 1 a reversal in the hierarchy of feature importance: for the first time, allegation types are more important than decision-makers (now judges, instead of investigators). We consider this an important general finding from 4.3, which can be supported by the Tree render in Fig. 8. Unlike previous renders (Fig. 6 and Fig. 7 above), we can't spot any decisions involving victim race.

5. Relational Analysis and Graph Analytics

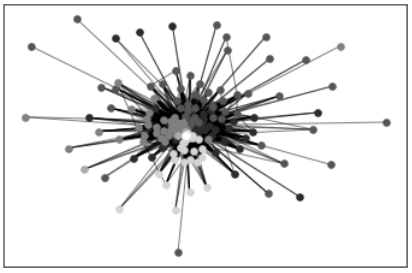
After showing that two realities coexist — while global views don't show clear disparities related to race, we could detect specific patterns that are worrisome — we can tap into a highly relational representation of the CPDB dataset to unearth how police misconduct may be affecting different segments of the population. In particular, we can see how violent patterns are much more connected to black victims.

5.1) Analyzing **co-occurring allegation types** for the entire population

Communities of Co-Occurring Allegation Types — Entire CPDB Population
'Proper Care - Injury / Death', 'Escape', 'Excessive Detention - After Arrest', 'Prisoners Property', 'Illegal Arrest / False Arrest', 'Bonding/Booking/Processing', 'Search - Person / Property', 'Search Of Premise Without Warrant', 'Bonding/Booking/Processing', 'Injury / Death (Under Color Of Law)', 'Improper Search Of Vehicle', 'Telephone / Attorney / Relative Privileges', 'Racial Profiling', 'Arrest, Improper Procedures', 'Improper Detention', 'Failure To Ensure Civil Rights', 'Unlawful / Excessive Investigative Detention (Witness)'

'Sexual Orientation', 'Miscellaneous', 'Use Of Profanity', 'Improper Search Of Person', 'Impairment .04 Or Greater - On Duty', 'Possession / Drinking Alcohol - On Duty', 'D.U.I. - Off Duty', 'Intoxicated Off Duty', 'First Amendment', 'Racial / Ethnic, Etc.', 'Intoxicated On Duty', 'Impairment .04 Or Greater - Off Duty', 'D.U.I. - On Duty'
'Inadequate / Failure To Provide Service', 'Neglect Of Duty', 'Misuse Of Department Records', 'Eeo Investigations', 'Insubordination', 'Misuse Of Department Equipment / Supplies', 'Late - Roll Call / Assignment / Court', 'Traffic Pursuit', 'Firearm Discharge With Hits - Rifle / Assault Weapon', 'Lunch / Personal Violations', 'Motor Vehicle Fatality - On Duty', 'Take Down (Thrown To Ground)', 'Firearm Used As An Impact Weapon', 'No Injury', 'D.U.I., Drugs / Controlled Substance - On Duty', 'Drug/Substance Abuse'

Table 2 — Co-occurring allegation types for the entire CPDB dataset.



As seen in Fig. 9, it's hard to analyze large graphs using conventional visualizations. And although important, global distributions of topological features such as node degrees and edges types can only inform us of global characteristics of a graph. Fig. 10, for instance, shows that edges in this graph of co-occurring allegation types follow a power-law distribution, that is, relatively few combinations of allegation types are much more frequent than all the others. This is, however, a common trait of many large graphs [3].

Fig. 9 — A render of our first graph.

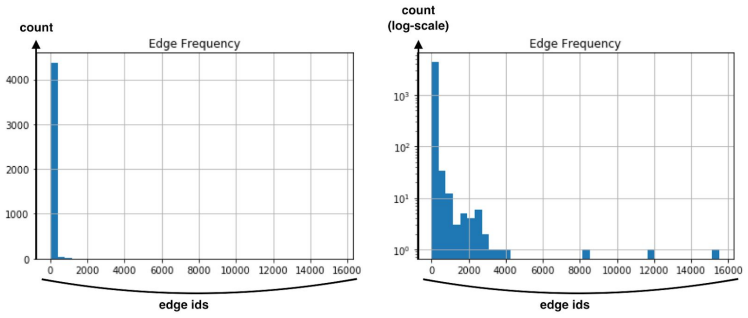


Fig. 10 — Distribution of edges shows how some patterns are much more prevalent than others, following a power-law. Left: absolute differences in edge counts. Right: differences in log-scale.

For these reasons, we resort to a community detection algorithm to analyze our graph [4] and we evaluate its performance by using modularity, which measures the extent to which nodes belonging to a community tend to form connections with each other in comparison with the rest of the graph [5]. For the entire population, community detection returned a modularity of **Q = 0.55**, which indicates *very meaningful communities* from a topological perspective [5].

Looking more closely at each community, Table 2 highlights our most interesting findings. The first community has "Racial Profiling" and many other violent allegation types tending to co-occur with it, including fatal misconducts and detention of witnesses. The second one has patterns of intoxication and DUI ('Possession / Drinking Alcohol - On Duty', 'D.U.I. - Off Duty', 'Intoxicated Off Duty', 'Intoxicated On Duty', 'D.U.I. - On Duty'), and more racially motivated misconduct ('Racial / Ethnic, Etc.'). The third is another one mixing patterns of discrimination ('Eeo Investigations'), intoxication, and a few other violent allegation types. EEO investigations refer to formal complaints of discrimination. Other communities were omitted because they are either too large (catch-all communities) or too small (isolated nodes).

5.2) Analyzing **co-occurring allegation types** for black *vs.* white victims

Communities for Black Victims	Communities for White Victims
'Racial / Ethnic, Etc.', 'Stomped / Stepped On', 'Taser (Probe Discharge)'	**'Excessive Force / On Duty - No Injury', 'Verbal Abuse'**
'Closed Hand Strike (Punch)', 'Shots Fired - No Hits'	**'Domestic Incident - Not Physical', 'Racial Profiling', 'D.U.I. - On Duty', 'Closed Hand Strike (Punch)'**
'Unnecessary Physical Contact / Off Duty - No Injury', 'Conduct An Improper / Inadequate Investigation'	**'Gang Affiliation', 'Illegal Arrest / False Arrest', 'Use / Abuse Drugs / Controlled Substance - Off Duty', 'Take Down (Thrown To Ground)'**
'Excessive Force - Use Of Firearm / Off Duty - Injury', 'Theft', 'Eeo Investigations', 'Kicked'	**'Sexual Orientation', 'Unnecessary Physical Contact / Off Duty - No Injury'**
'Excessive Detention - After Arrest', 'Prisoners Property', 'Dragged', 'Firearm Discharge With Hits / On Duty', 'Failure To Ensure Civil Rights', 'Assault / Battery, Etc.'	**'Unnecessary Display Of Weapon / Off Duty', 'Misconduct During Issuance Of Citation', 'Use Of Profanity'**
'Unnecessary Display Of Weapon / Off Duty', 'D.U.I., Drugs / Controlled Substance - Off Duty', 'Choked', 'Intoxicated On Duty'	**'Racial / Ethnic, Etc.', 'Harassment'**
	'Neglect Of Duty', 'Criminal Sexual Assault', 'Failure To Ensure - Civil Rights', 'Violation (Other Than D.U.I.) - On Duty', 'Other Felony'
	'Excessive Force / Off Duty - Injury', 'Sexual Misconduct', 'Burglary', 'Initiate Proper Action'
	'Excessive Force - Use Of Firearm / On Duty - No Injury', 'Theft', 'Use / Abuse Drugs / Controlled Substance - On Duty', 'Drugs / Controlled Substance, Possession Or Sale'

Table 3 — Co-occurring allegation types for black (left) *vs*. white victims (right).

Focusing on victims of color, community detection returned a modularity of **Q = 0.67**, which indicates *extraordinarily meaningful communities* from a topological perspective [5].

Again, we highlighted our most interesting findings in Table 3. Communities with "'Racial / Ethnic, Etc.', 'Stomped / Stepped On', 'Taser (Probe Discharge)'" and "'Racial Profiling', 'Firearm Discharge With Hits - Handgun'" are particularly worrisome, linking violent behavior to racism. The other communities show more exceptionally violent behavior tending to co-occur in allegations involving victims of color.

Focusing on white victims, community detection returned a modularity of **Q = 0.74**, which also indicates *extraordinarily meaningful communities* from a topological perspective [5].

Again, we highlighted in our most interesting findings in Table 3. While we do find patterns of intoxication, we see "'Excessive Force / On Duty - No Injury', 'Verbal Abuse'" and "'Unnecessary Display Of Weapon / Off Duty', 'Misconduct During Issuance Of Citation', 'Use Of Profanity'" illustrating how misconduct patterns tend to be less violent when compared to our previous analysis. However, we can see patterns of misconduct due to sexual orientation or sexual offense, and still other patterns of racial and ethnic discrimination.

It's also worth noting that we have a larger range of allegation types involving people of color compared to white victims (125 vs 108). Similarly, we have many more combinations of allegation types (1292 vs 725).

6. Conclusion

In conclusion, our work uncovered a number of findings indicating racial disparities both in the occurrence and in the treatment of police misconduct in the city of Chicago. We started our work by showing that police misconduct allegations involving victims of color are less likely of being sustained.

Next, we showed how disciplinary actions can also be influenced by victim race and how settlement amounts, although harder to analyze, are not always equitable.

We tapped into our Machine Learning toolbox to show how very simple Decision Trees — with only three features! — can capture remarkable signal. For investigation time and disciplinary action, we saw that the investigator is the most important feature for determining the outcome, followed by the allegation type. Although victim race has only a marginal effect when evaluated globally, we could see that under specific conditions (i.e., in specific Tree paths) victim race can be among the features with highest information gains. Conversely, we saw that settlement amounts are much more dependent on the allegation type than anything else, and we couldn't spot any racial profiling associated to a judge. Overall, by tying good predictive performance and interpretability, we believe that Decision Trees are a valuable tool for navigating this data and pushing for accountability of specific investigators.

Finally, we constructed a graph of co-occurring allegation types to show how different patterns of police misconduct may affect different segments of the population. We ran community detection algorithms for the entire CPDB dataset, allegations involving black victims, and allegations involving white victims, to show how different patterns tend to co-occur for different segments. We consistently got good clustering results according to a gold-standard in the field. As a result, we could look into each community and find different patterns of violence strongly associated with race, ethnicity, and sexuality, which not only confirms our initial hypothesis but potentially extends it towards other segments of the population. Looking at our results for black victims, in particular, we found that violent behavior is much more prevalent among this population and is manifested in a number of different patterns.

Overall, this project has provided quantitative information to reason about race and police misconduct in the city of Chicago. We believe that our results support the hypothesis that race and ethnicity may play a role in the outcomes of police misconduct investigations, even if under specific circumstances, and that the use of force seems to be racially discriminatory in a more systematic manner.

7. References

[1] "Chicago Police are 14 times more likely to use force against black men than against whites." — https://theintercept.com/2018/08/16/chicago-police-misconduct-racial-disparity/

[2] Satyanarayan, A., Moritz, D., Wongsuphasawat, K., & Heer, J. (2016). Vega-lite: A grammar of interactive graphics. IEEE transactions on visualization and computer graphics, 23(1), 341-350.

[3] Barabási, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509-512.

[4] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

[5] Newman, M. E. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.