The Spectacular Sailors
Justin Chae
Milan McGraw
Pengyi Shi

Modeling Risk of Police Misconduct

Checkpoint 1 (Updated 18 October 2020)

OVERVIEW

In Checkpoint 1 we establish a base of knowledge to characterize officers by their membership in three distinct groups. We base our initial categorization based on the results of previous topic modeling efforts that identified communities of officers and sub-communities as crews. The third community of officers are those that are neither in a crew or in a community, i.e. all other officers.

In this analysis, we define three groups as “Cohorts” where Cohort 1 are those officers in crews, Cohort 2 are those officers in communities and Cohort 3 are all other officers. Given these groupings, we investigate what the average police officer looks like in each group as compared to their counterpart in the other group. At a high-level, the questions we investigate below are (1) who are the members of each group, (2) how often are they accused or disciplined for misconduct, (3) how does each group compare to each other in terms of misconduct, and (4) what does the average officer look like in each group? The findings are outlined below; further data and details are available on GitHub.

Findings - Question 1

Question 1: Who are the police officers according to their membership across three distinct cohorts (“the Cohorts”): (1) in a crew, (2) in a community, (3) not in a crew and not in a community?

Analysis 1

In a previous document analysis effort, officers are classified into two distinct groups, crew or community. The premise behind grouping officers in this manner is that some tight-knit groups of police officers account for an outsized portion of police misconduct. If we can model problematic groups of police officers, then it may be possible to inform policy and management interventions to prevent groups from forming or from sustaining misconduct. As a result, the first step of Checkpoint 1 is to identify specific data about member in all three Cohorts with a few base queries.

Discussion 1

To answer Question 1, we start with a series of queries based on data_officercrew, data_crew, and data_officer to determine who and how many officers are in a crew, in a community, or neither in a community or a crew. The result is a working table called ***officers_cohorts_data*** that flags each officer's Cohort and joins data about the officer and data about the officers' misconduct. In our cp1_crews_comm_other.sql from line 0 to approximately line 228, we build the ***officers_cohorts_data*** working table; key fields include:

- **officer_id**: the officer ID
- **crew_id**: the crew ID, else 0 if not in a community
- **community_id**: the community ID, else 0 if not in a community
- **cohort**: 1 is crew; 2 is community, 3 is everyone else
- **Also includes**: Officer Demographics, Complaint Data, Disciplinary Data

Query Summary for Question 1 `SELECT * FROM officers_cohorts_countstotal;`	
cohort	total_officers
1	1156
2	10071
3	23780

Conclusion 1

When summarized, we start to get a picture of each Cohort. For example, in the first table below, we have a representation of all police officers, segmented by Cohort affiliation. The topic modeling work of prior groups is a critical starting point for this analysis as it allows us to create our version of a "Rosetta Stone" to go between Cohorts, Officers, and Misconduct Data. In the following sections, we slice, aggregate, count, and summarize officers when grouped by Cohort.

Findings - Question 2

Question 2: Within each Cohort, what is the average number of co-accusals per individual complaint? Where the average is given by the sum of co-accusals in a Cohort divided by the total number of complaints (where a complaint is a unique CRID)

Analysis 2

WIth our rosetta stone table in place, we continue joining and querying tables in the schema to add to our knowledge base. In Question 2, we attempt to summarize high-level statistics that may provide insight into how each Cohort is different. A particular challenge that we continue to work through is to suss out the unique complaint data for each officer, denoted by unique CRID, and correctly average coaccusals for each Cohort.

Query Results for Question 1 and 2

```
SELECT * FROM officers_cohorts_coaccused_counts;
```

cohort	unique_crid_count	total_coaccusals	avg_coaccused_count
1	20174	56294	2.790423317
2	96651	196844	2.036647319
3	47137	91000	1.930542886

Discussion 2

To answer Question 2, we sliced a working table for each Cohort, grouped, and counted each Cohort for unique CRID records, and then combined them with counts to produce **officers_cohorts_coaccused_counts**. The working table called **officers_cohorts_coaccused_counts** provides the average coaccusals for each Cohort that accounts for officers that have at least one accusal record. In our cp1_crews_comm_other.sql from line 228 to approximately line 358, we build the working table; key fields include:

- **cohort**: 1 is crew; 2 is community, 3 is everyone else
- **Unique_crid_count**: given by distinct CRID (working issue: some CRID records have a leading “C”; resulting in approximately 1,100 unresolved duplicates as of 18 OCT 2020).
- **Total_coaccusals**: a sum of coaccusal column data
- **Ave_coaccused_count**: when segmented by officers in a cohort with at least one allegation, the average number of coaccusas.

Conclusion 2

The topic modeling work of prior groups is a critical starting point for this analysis as it allows us to create our version of a “Rosetta Stone” to go between Cohorts, Officers, and Misconduct Data. In the following sections, we slice, aggregate, count, and summarize officers when grouped by Cohort.

Findings - Question 3

Question 3: Within each Cohort, what percentage of allegations results in disciplinary action? Where the percentage is calculated by total allegations in cohort / total disciplined in cohort.

Analysis 3

With Question 3, we explore how officers in each Cohort compare as given by their disciplinary percentage. For instance, given a number of allegations or coaccusals, how often is an officer in each cohort disciplined? Of note, in Question 3, we start to see how nearly every officer in Cohorts 1 and 2 have at least one allegation, while just over half of all officers in the third population, Cohort 3, have any allegations at all.

Query Results for Question 3

```
SELECT * FROM officers_cohorts_countdisciplines;
```

cohort	officers_with_allegations	is_disciplined
1	1156	1144
2	10071	9921
3	12217	7880

Discussion 3

Initially we counted the numbers of officers with allegations to compare against whether they are disciplined. However, we instead divided the number of coaccuasls by the number of disciplinary actions to paint a relatable picture. As provided in the following conclusion table in Question 4 the discipline ratios, by Cohort, are as follows:

- Cohort 1: 5.67%
- Cohort 2: 10.26%
- Cohort 3: 16.17%

Conclusion 3

Officers in Cohort 1, crews, are infrequently disciplined as compared to other cohorts by as much as three times. Given this insight, we wonder what factors explain this discrepancy. For instance, are ‘bad cops’ bad because they are not disciplined, or are good cops good because they are?

Findings - Question 4

Question 4: For each Cohort, describe the average police officer in terms of demographics, accusals, and payout data.

Analysis 4

While considering how each group behaves, we join and summarize all of our findings into a single table and 5as we consider next steps for visualizations and machine learning. While some fields are already explained, we perform several new joins and queries that include a sum of total costs which are given by a combination of payouts and lawsuit costs. Moreover, we continue to refine the queries and tables to include metrics such as age, years on force, and demographic breakdowns of each officer.

Query Results for Question 4 and Summary of Checkpoint 1

```
SELECT * FROM officers_cohorts_counts;
```

cohort	total_officers	officers_with_allegations	unique_crid_count	is_disciplined	total_coaccusals	avg_coaccused_count	discplined_rate	total_cost
1	1156	1156	20174	1144	56294	2.790423	0.056706	6138736998
2	10071	10071	96651	9921	196844	2.036647	0.10264	20729755018
3	23780	12217	47137	7880	91000	1.930542	0.167172	2410155781

Discussion 4

At this time, we are initially intrigued with data that affirms our intuitive bias. For instance, crews have a relatively higher coaccusal percentage and lower discipline percentage that the other Cohorts. However, in terms of total cost, Cohort 2, *communities*, accounts for nearly 70% of the monetary cost of misconduct. This last point about cost may be worth further investigation. For instance, although it may be alluring to focus on crews, the most widespread economic gain may be to focus on how to reform the 'middle third' of officers that are in communities.

Conclusion 4

As of Checkpoint 1, we have answered some preliminary questions about the differences between Cohorts, or crews, communities, and all others in the police force. However, as we continue to investigate, we now have more questions about lower-level attributes of each officer to answer. We hope to build on this effort with visualization and machine learning next.