***Chicago Police Department - Checkpoint 3***

**Introduction**

In this checkpoint, our team focused on investigating how officers complaints are distributed in two ways. First, how are complaints clustered in groups of officers (co-accousal). Second, how do individual officers behave over time after receiving some number of complaints.

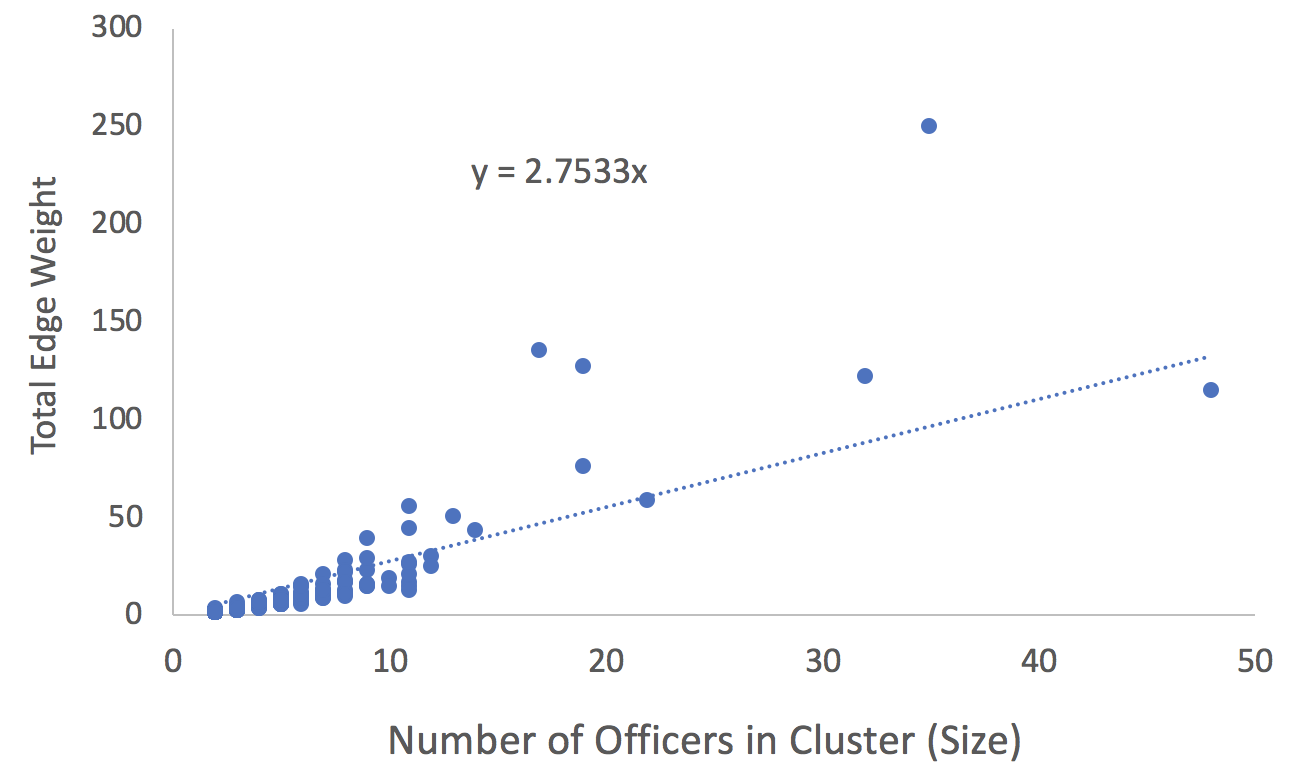
**Core Questions**

**Question #8:** ***Do complaints tend to be filed against a cluster of police officers? How do these potential clusters change over time?***

In order to explore this question - we sought to look at co-accusals of officers in the form of complaints from a graph perspective. Modeling each officer as a node, we can represent co-accusals as an undirected edge with a weight of 1, and add on additional weights if a set of officers has more than one co-accusal.

Using GraphX in Spark (through the GraphFrame package for Python adaptation) we were able to create such a representation. From here, we ran two core analyses. As mentioned in the queries document, we only used data from one full year for both analysis. The reason for this is explained further in the specific methods. First we ran a connected components analysis to create distinct groups of officers. In this analysis, an officer is added to a cluster if any connection is made to one or more officers in that particular cluster. This analysis benefitted from a one year time frame, because it was a sufficient lapse in time to create clusters but not so long that all officers eventually were grouped into the same cluster. From this analysis we were able to detect about 600 connected components, from approximately 2,100 officers. From these clusters we were able to create the total number of officers in each cluster, and the total edge weight for the cluster. We have plotted both metrics on the histogram below.

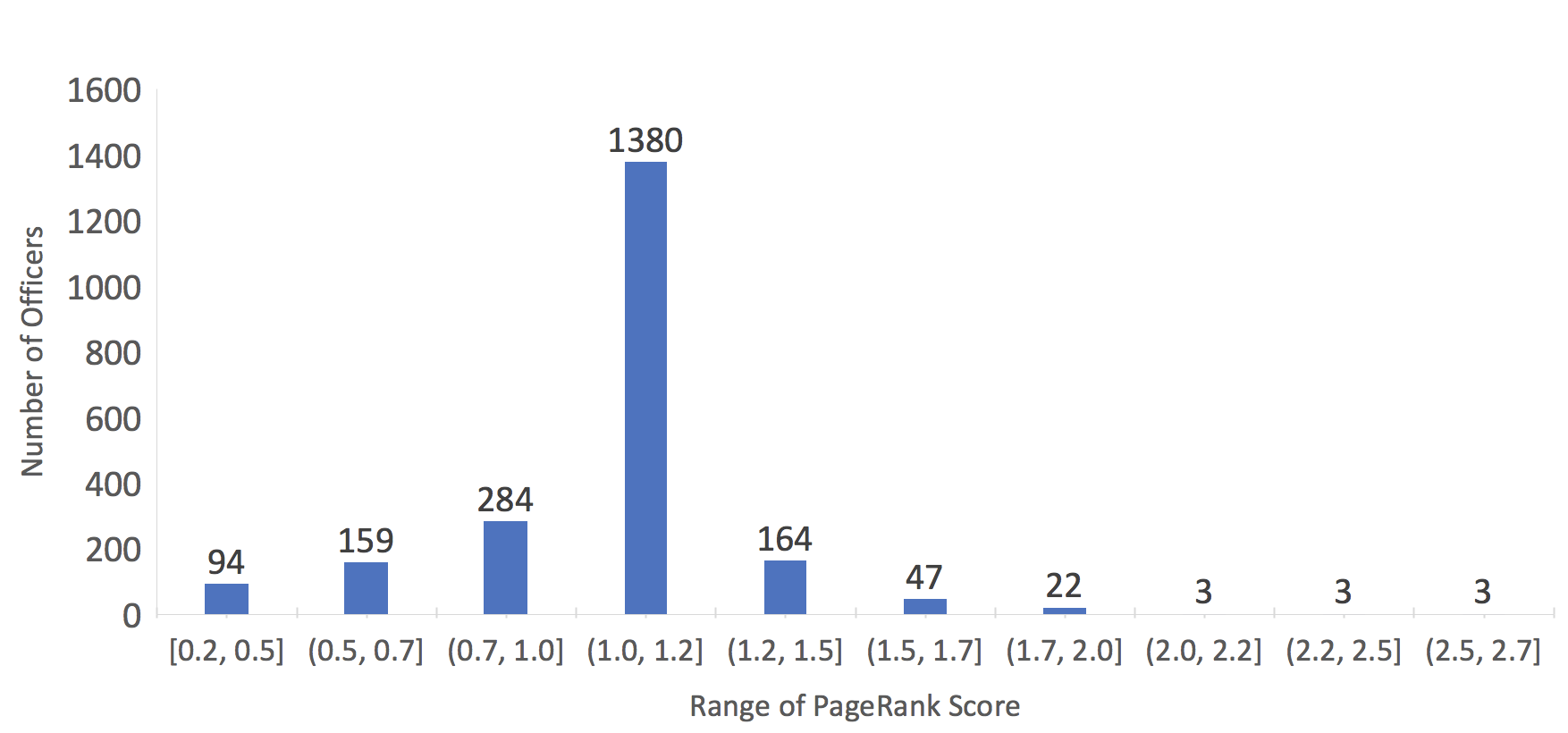
**Figure 1**: Connected Components Analysis (Scatter Plot)



Looking at this graph, we can see some interesting results. First, we observe that there are some clusters, particularly in the 15 - 25 officer range, which had a very high number of complaints for the given number of officers in that cluster. This indicates that potential repeat offenders will actually work together over time. This is further demonstrated by adding a linear regression line to get a general sense of the average number of complaints (correlated to weight) per officer in each cluster.

To further examine the connectedness of the officers, we also sought to run the PageRank algorithm. We were able to run the algorithm using Graphframe - and given that the process is very computationally expensive, the smaller time frame of a year allowed us to complete the process in a reasonable amount of time. The same graph representation was used above, and instead of weighting edges, each edge represented a weight 1, and multiple edges from one node to another were allowed.

**Figure 2**: PageRank Analysis (Histogram)



Looking at the figure of PageRank scores, we can see a similar behavior. Here we see there is an inflection point past about 1.2 PR score where officers are highly connected - i.e. they have many different claims against them with several different officers.

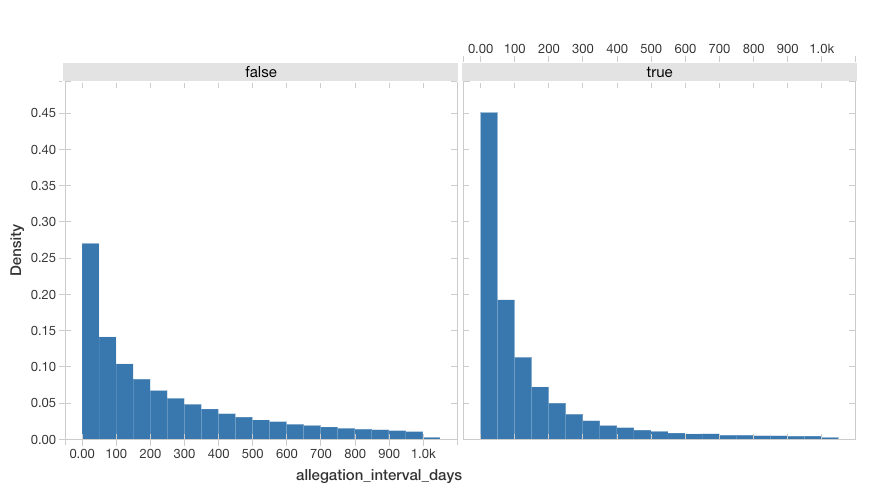
As we move forward with the project, we would like to explore this analysis for different timeframes, and observe how officers with a very high pagerank fit into our previous normalization analysis.

**Question #9:** ***Are we able to see if once police officers potentially "get away" with certain behaviors, they are much more likely to keep committing the same acts repeatedly?***

To analyze how the officers react to complaints, we examined the time between complaints for all of the officers. To see how being a repeat offender affects this analysis, we flagged the officers that we identified as ‘repeat offenders’ from our previous work completed in checkpoint 2. We took the officers that were in the top 7% of officers with the highest number of complaints per year normalized by their district average. We chose this number given that this was an inflection point in the distribution. These officers are the ones in the right-side of the following plots with ‘repeat\_officers’ attribute being ‘true’.

We calculated the time interval in days between complaints for each officer, and tracked whether that officer was flagged as a ‘repeat officer’ or not. We aggregated the complaint intervals for these two buckets of officers, and created histograms to show how they are distributed.

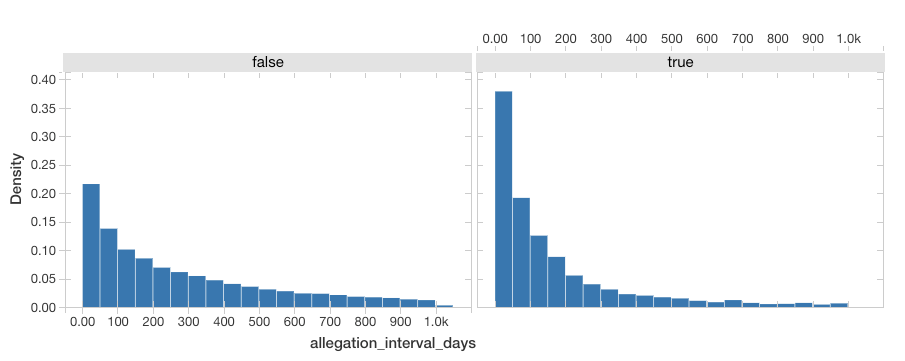
**Figure 3**: Histogram of Interval B/W all complaints (Density vs. Allegation Interval Days)



Looking at the this first plot, we see what is largely to be expected, and the intervals for the ‘repeat offenders’ is more heavily aggregated towards the left. The density of the first bucket (interval of 0 to 50 days) for ‘repeat offenders’ is around 45%, whereas the first bucket for ‘non-repeat offenders’ is only around 27%. Note - the x axis for both of these plots is the same. This means the repeat offenders have a higher proportion of short intervals between complaints against them.

To dig a bit deeper, we wanted to see if getting disciplined has an affect on the amount of time that passes before the officer gets another complaint. We repeated the previous analysis, but this time only captured the intervals where the officers’ were disciplined for the first complaint.

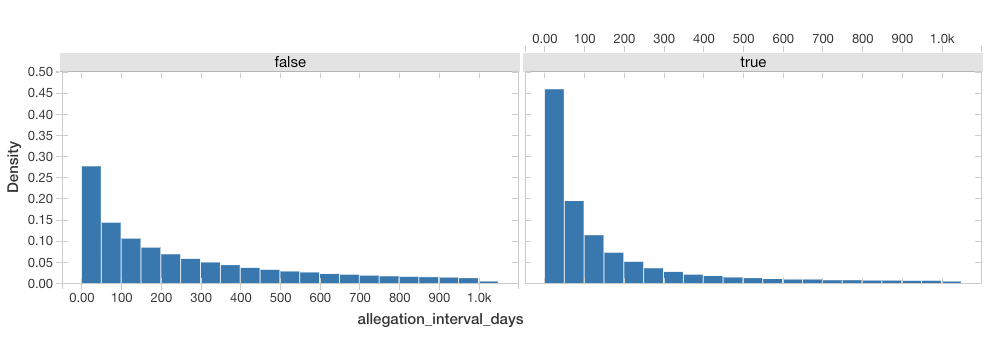
**Figure 4:** Histogram of intervals between complaints after Disciplined complaints (days)



As expected, the density of the first bucket (0-50 days) drops for both ‘Repeat offenders’ and ‘non-Repeat Offenders’, but not by much. This means the officers are reacting towards the discipline and more time goes by before they act out again. However, we expected the effect to be much larger on the Repeat Offenders given that they simply have a higher volume of complaints. This is something to dig into further as we expand upon this segmentation.

We then looked at the counter-set of intervals between complaints if the officer was not disciplined for the first complaint.

**Figure 5:** Histogram of intervals between complaints after Non-Disciplined complaints (days)



Again, as expected we see an increase in density for the lower-bound intervals for both groups of officers, meaning less time passes between them getting another complaint after getting away without punishment for a complaint.

We summarized these findings by finding the overall averages for these three analyses and got the following:

**Figure 6:** Average Intervals Between Complaints

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Overall Average Interval (Days)** | **Average Interval After Disciplined Complaint (Days)** | **Average Interval After Non-Disciplined Complaint (Days)** |
| **Repeat Offenders** | 145 | 177 | 142 |
| **Non-Repeat Offenders** | 398 | 481 | 393 |

Overall, this data supports the assertion that disciplinary action does in fact have an effect on keeping the officers in line. After being disciplined, on average, more time passes before they receive another complaint.

**Conclusion**

During this checkpoint, we were able to answer the two questions in our proposal well. Overall, our intuition was largely correct for both cases. Officers who have many complaints tend to form clusters, and subsequent complaints happen faster after not being disciplined, and slower after being disciplined. Moving forward, as we gather more information from following checkpoints, we will be able to come back and refine our analysis as mentioned above. We would also like to consider using velocity vectors to add additional support for question 9 - and we think having a more robust idea of what to look for from further checkpoints.