Checkpoint 4: Graph Analytics

The Freedom Deer: Tianchang Li, Hualiang Qin, Qingwei Lan

November 18, 2021

We implemented our graph analytics modeling questions using GraphFrames and Apache Spark. We pulled our data from the CPDB Postgres server and performed post-processing using Python in a Jupyter Notebook environment.

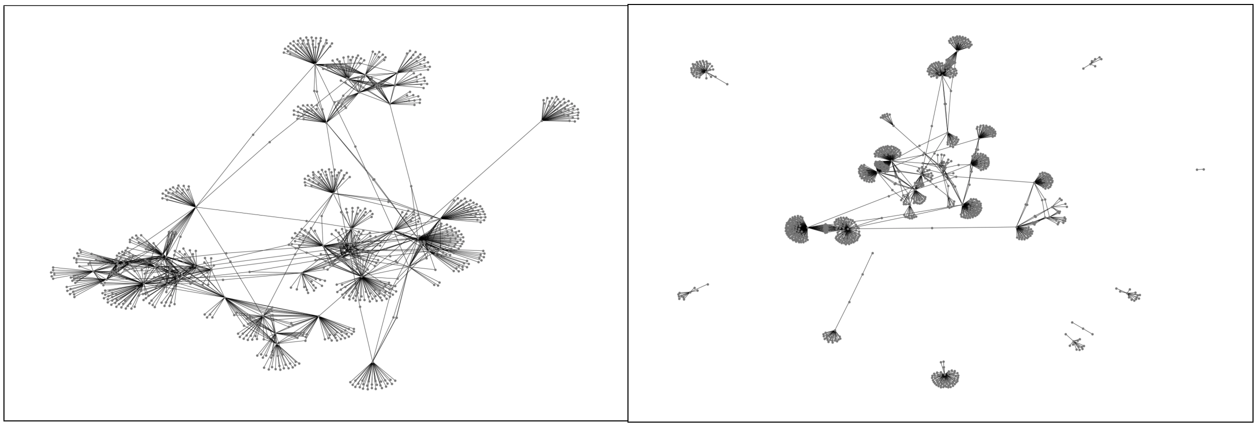
1 Analysis of Co-offending Network Graphs

We believe that officers with the worst influence are those with the most triangle counts due to the scope of influence on other officers. Therefore we plot the network graph of co-offending TRR cases of the officers with the largest triangle counts. Each node in the graph is an officer and each edge represents a co-offending pair. Furthermore we also plot the baseline network graph with each node representing an officer and each edge representing a combination of two officers working together.

The results in Figure 1a show that the TRR graph is tightly connected whereas in Figure 1b the baseline graph is rather sparse. This indicates that the officers involved in TRR cases are not those who commonly work together.

We also plot the same network graphs for TRR cases involving only cross-race use of force.

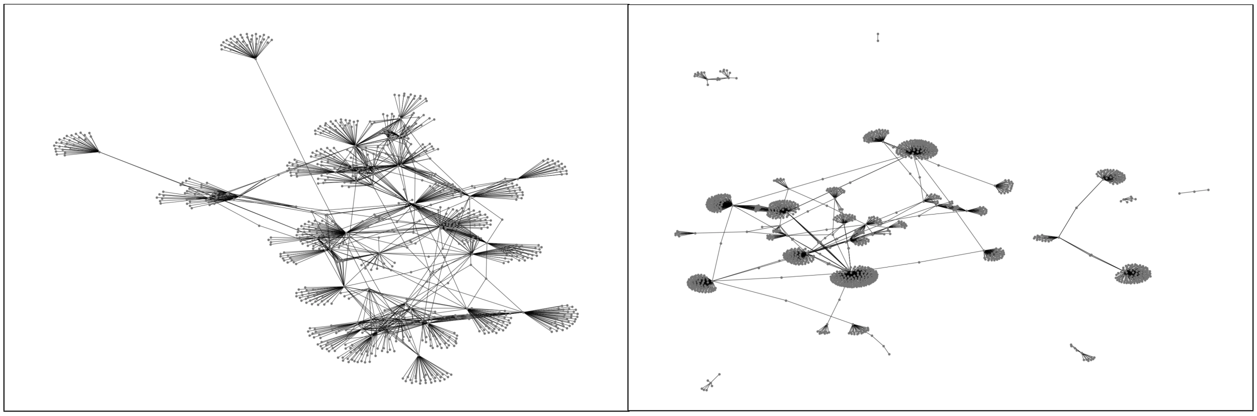
The graphs in Figure 2a and Figure 2b show the same results. This indicates that there is a group of offending officers who are responsible for the majority of the TRR cases. Furthermore, they are the one influencing other officers (the ones not commonly engaged in TRR cases). We consider these the “bad apples”.



(a) TRR network graph

(b) Baseline network graph

Figure 1: TRR network graph and baseline network graph for all TRR cases.



(a) TRR network graph

(b) Baseline network graph

Figure 2: TRR network graph and baseline network graph for TRR cases involving only **cross-race use of force**.

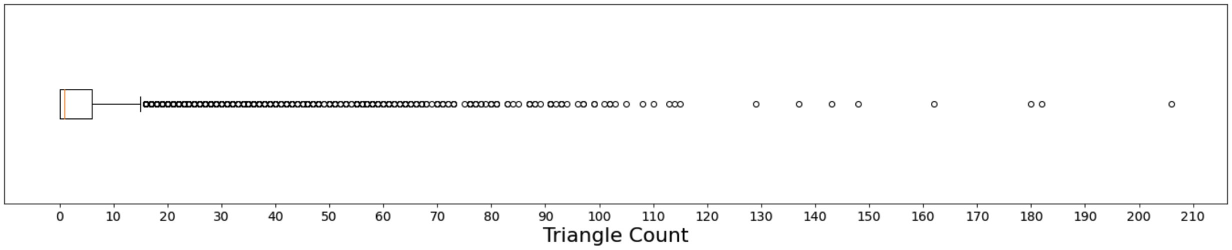


Figure 3: Triangle count boxplot for all TRR cases.

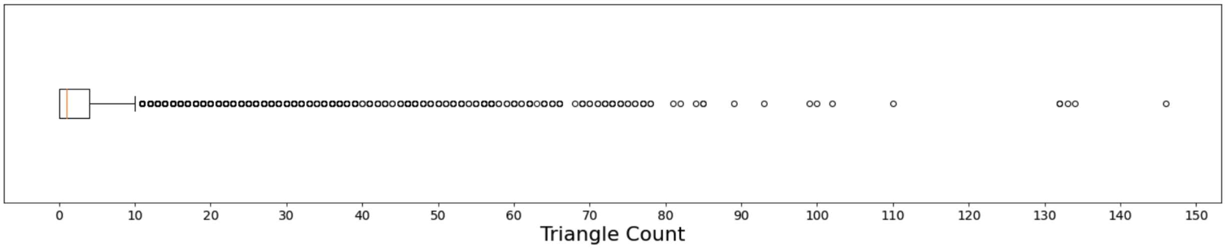


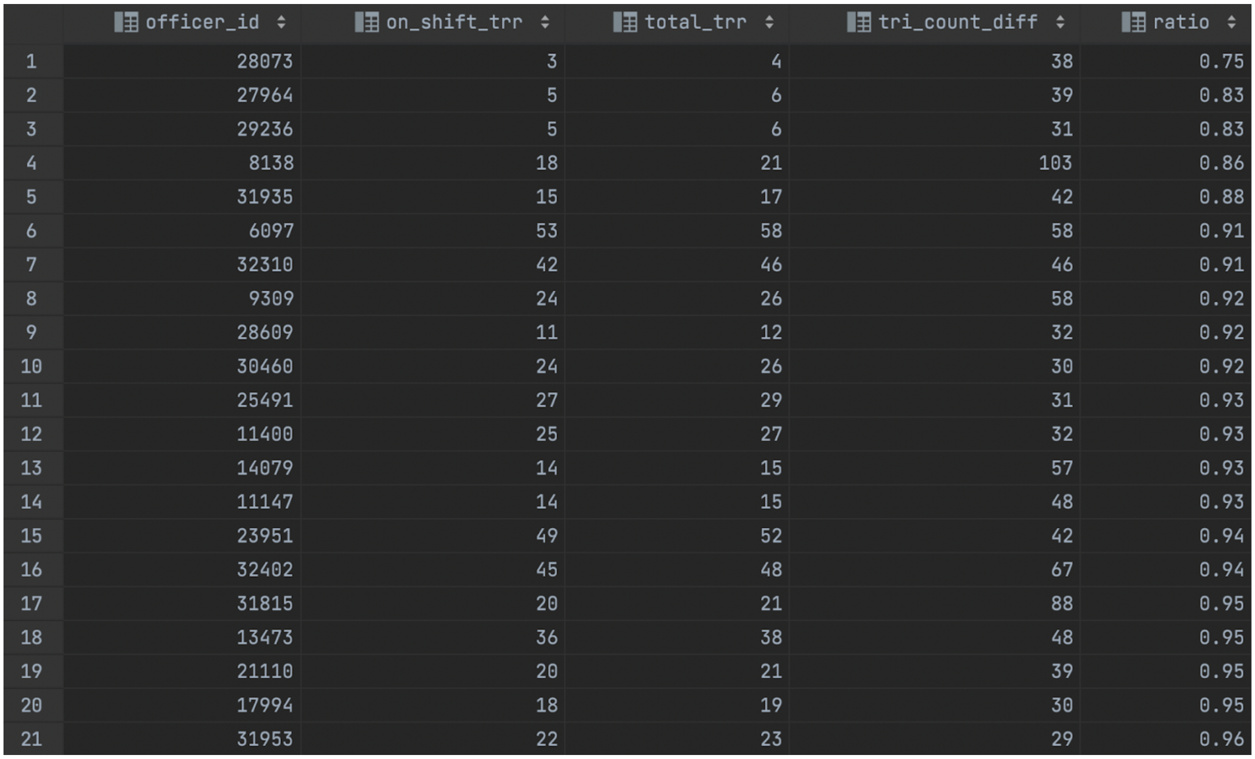
Figure 4: Triangle count boxplot for TRR cases involving only **cross-race use of force**.

To further support this hypothesis, we plot boxplots for the triangle counts for TRR cases. Figure 3 and Figure 4 show that most of the officers with a large number of triangle counts are considered outliers.

2 Off Shift TRR Analysis

We hypothesize that officers with relatively more off-shift TRR cases given their total number of TRR cases are surely breaking the rules and actively influencing others with excessive force usage.

To investigate this, we extracted the top 200 “bad guys” (defined by the officers with the largest difference in their triangle count between TRR event graph and baseline shift graph). Then we extracted the ratio of their number of off-shift TRR over their total number of TRR, as below (“ratio”). Our thought is that officers with the most involvement in TRR community might be the ones who engage in the undesired activities.



	officer_id	on_shift_trr	total_trr	tri_count_diff	ratio
1	28073	3	4	38	0.75
2	27964	5	6	39	0.83
3	29236	5	6	31	0.83
4	8138	18	21	103	0.86
5	31935	15	17	42	0.88
6	6097	53	58	58	0.91
7	32310	42	46	46	0.91
8	9309	24	26	58	0.92
9	28609	11	12	32	0.92
10	30460	24	26	30	0.92
11	25491	27	29	31	0.93
12	11400	25	27	32	0.93
13	14079	14	15	57	0.93
14	11147	14	15	48	0.93
15	23951	49	52	42	0.94
16	32402	45	48	67	0.94
17	31815	20	21	88	0.95
18	13473	36	38	48	0.95
19	21110	20	21	39	0.95
20	17994	18	19	30	0.95
21	31953	22	23	29	0.96

Figure 5: Table of officers with highest percentage of off-shift TRR.

From Figure 5, we found that, ordering by "ratio", there are many of them with off-shift TRR. However, the officers with the lowest ratio are those with fewer TRR cases. Most of these officers have a ratio over 0.9, which means off-shift TRR is not quite relevant to active use of force. This disagrees with our original hypothesis. Therefore, a taken-away message could be it's not going to be very helpful for regulating use of force by monitoring officers' off-shift TRR.

3 PageRank Analysis

To further analyze the prominent influencers in this network, we compute the PageRank of each officer in the network graph. Since our edges are originally directed, we need to first convert them to undirected edges by cloning each edge and reversing the source and destination. Then we concatenate the reversed edges with the original edges to produce an undirected graph.

PageRank measures the relative importance of a node within a connected graph. In our case, officers with a higher PageRank value mean they are influencing more officers by being in TRR co-offending cases with more different officers. This measure is similar to the triangle count measure but accounts for influence on a global scale whereas the triangle count is more limited to local influence within tight-knit groups.

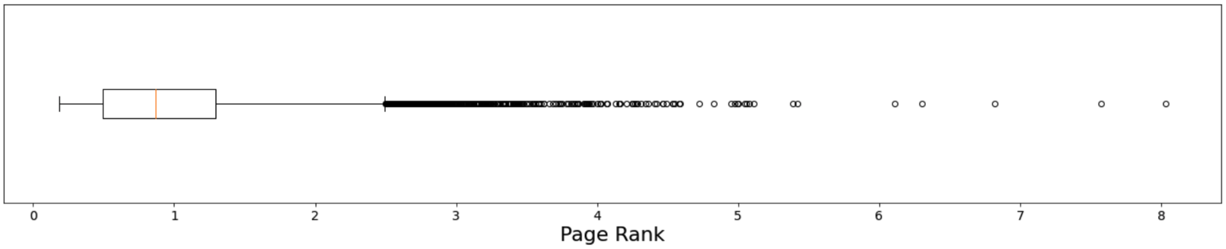


Figure 6: PageRank value boxplot for all TRR cases.

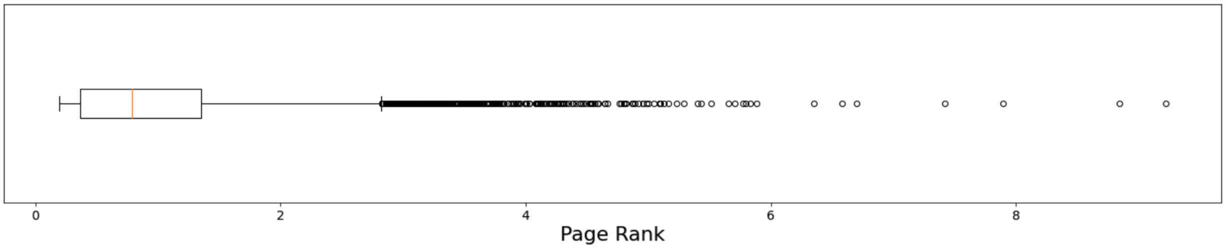


Figure 7: PageRank value boxplot for TRR cases involving only **cross-race use of force**.

From Figure 6 and Figure 7, we can quickly verify that the most influential officers are outliers in the graph. However, by comparing the 200 most influential officers according to PageRank values, we can see that they are fundamentally different from the influential officers determined by triangle count (100 out of 200 officers are different). This supports our reasoning of the local vs. global influence.

4 Conclusion and Future Research

From our graph analysis, we can conclude that there is a group of officers that account for the majority of the TRR cases. These officers are also influencing other officers based on our analysis on co-offending TRR cases. This analysis allows us to dig deeper into the specific, individual cases to discover reasons on why this is the case.

Unfortunately, we didn't find any evidence of cross-race use of force playing a significant role in this system. For future research, we would look into the individual reports and apply NLP techniques to determine whether racial discrimination plays a role in police use of force.