Checkpoint 4: Graph Analytics

COMP_SCI 396/496: Data Science Seminar
Meenakshi Kommineni, Madison McClellan, Archana Ramasubramaniam

If we construct a graph where nodes represent officers and edges signify that two officers were involved in the same allegation, can we then determine whether specific officers are more involved than others in these incidents?

To construct our graph we used data from the data_officer and data_officerallegation tables. The data for nodes was queried by selecting all of the values for officer_id from data_officer. The data for edges was queried by joining data_officerallegation with itself and selecting pairs of officers where the allegation_id matched. This match signifies that the officers were co-accused in the allegation. We then rank the PageRank algorithm over this graph to determine which nodes are highly connected to other nodes. These highly connected nodes represent officers that have been co-accused with many other officers. Looking at the PageRank results, we can identify these officers (based on their ID) that have a high PageRank value. For example, the officers identified by the IDs 8138, 9821, and 2375 have the top three highest PageRank values. Further investigation into these highly connected officers would be needed to draw any definitive conclusions, but we could potentially infer that these officers have a greater negative presence in the community.

We then wanted to determine whether these highly connected officers have any specific traits in common. To do so, we created a list of the 20 officers with the highest PageRank values. We then queried data from the data_officer table, filtering for only these 20 officers using their IDs, using the SQL statement below. All 20 officers in this group are males born in or before 1951. The racial distribution is 50% Black (10/20), 40% White (8/20), and 10% Hispanic (2/20). The average complaint percentile among the group is 92.5. Perhaps the most interesting insight out of these is that all 20 officers are male. This is consistent with what we learned from the Invisible Institute's policy recommendation of hiring more female officers.

```
SELECT * FROM data_officer
WHERE id IN (8138, 9821, 2375, 31906, 21530, 17816, 13303, 10442, 6315, 14294,
3033, 18042, 28273, 441, 5667, 8562, 16567, 29882, 11266, 9254);
```

Can we determine whether the graph described above contains many clusters (i.e. officers are highly connected in allegations) or whether officers are more isolated? How does this compare to a graph of officers involved in TRRs?

To answer this question we ran the Triangle Count algorithm over the graph constructed above. The resulting counts for each officer ID represent the number of triangles a given node is involved in. We can look at specific officer IDs and their corresponding counts to determine which officers are highly clustered. For example, we can see that the officers identified by the IDs 6315, 3033, and 3744 have the highest counts, with 32118, 32117, and 32037 triangle involvements, respectively. However, our goal is to make a conclusion about the network as a

whole rather than individual officers. As a starting point, we calculated the average count over all the nodes, which was 246.5. This value tells us that on average, a node is involved in 246.5 triangles.

We then wanted to compare the Triangle Count results of this graph to a graph of officers involved in TRRs. In this graph, nodes represent officers and edges signify that two officers were involved in the same TRR. We ran Triangle Count over this graph and found that in general, there are much fewer triangles. For example, the officers identified by the IDs 21371, 13313, and 11615 have the highest counts, with 206, 182, and 180 triangle involvements, respectively. The average count over all the nodes was only 1.8, compared to 246.5 for the allegations graph. From this comparison, we can conclude that officers are more often co-accused in allegations than they are in use of force incidents.