Data Pirates - Tracking Effectiveness of Police Reform
Checkpoint 4 - Graph Analytics

Kavya Jaganathan, Clarissa Cheam and Ayushi Mishra

Overview:

In this checkpoint, we test the hypothesis, "Are certain officers the cause of high misconduct counts in a certain district, or is it a unit-wide issue? If so, how does this affect reform choices?". We test this hypothesis to investigate whether reform needs to be implemented unit-wide or would retraining/removing certain officers improve the situation. To this end we will be interacting with officer nodes, unit_ids and coaccusal counts. This checkpoint will expand our focus from “Tracking police reform” to “Tracking and Nudging Police reform towards efficiency”.

Question 1: What is the distribution of allegation counts and trr counts in units? Are the counts uniformly distributed, indicating a unit wide problem, or is the distribution skewed, showing the presence of certain problematic officers?

Analysis:

This question aims to plot the distribution of allegation counts and trr_counts of officers in a unit to identify whether an entire unit is corrupt, or if only certain officers are contributing to the allegation counts. If the distribution peaks for certain officers, we can safely conclude that these officers are the high contributors of allegations and should be disciplined/removed from the force. **NOTE:** There are officers who have not been assigned to Units due to either lack of data, or these could just be lone officers (a subject matter expert is required to demystify the same). It’s imperative to analyse this data to see if we have officers with high allegation counts (***assumption:*** *high allegation counts implies count> 40; we also do not factor in false allegations and increase in reporting due to COPA*).
If the distribution is uniform and on the higher range, we can conclude that the entire unit needs to be revamped as all the officers are engaging in misconduct equally.

Discussion:

To answer this question, we created dataframes for officers (nodes) and the unit id as edges. Two officers have an edge between them if they belong to the same unit.

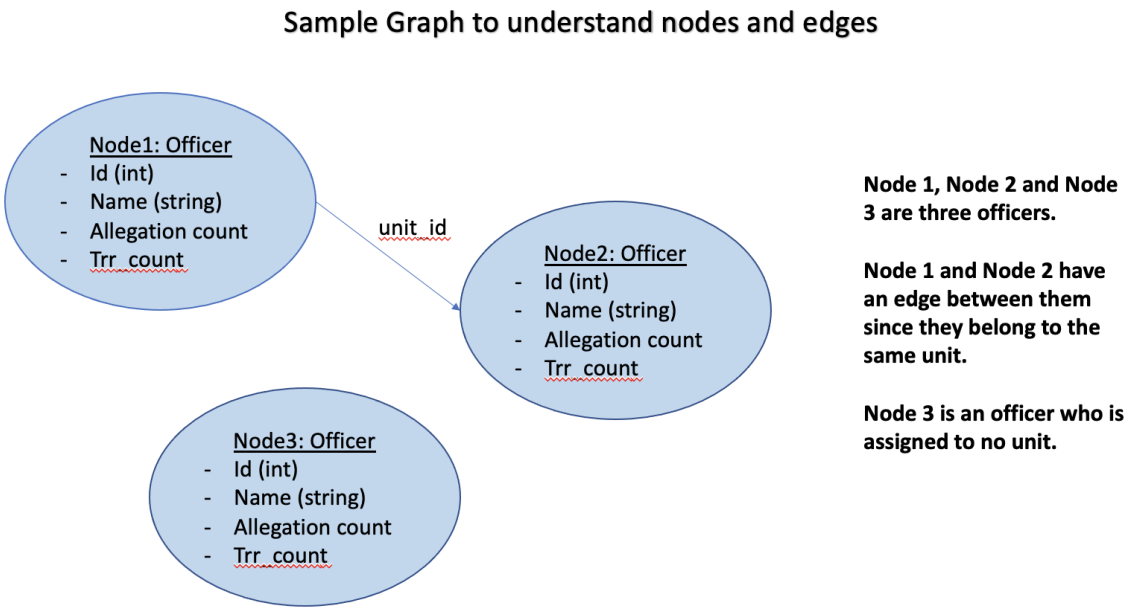

Fig 1: A sample graph illustrating node and edge definitions

id	officer_name	allegation_count	trr_count
32312	Randall Ryan	10.0	4.0
32358	Kevin Stoll	24.0	5.0
32480	David Lavin	0.0	0.0
32491	Kevin Rake	0.0	0.0
2680	Richard Brandt	0.0	0.0
29553	Hugh Versteegh	0.0	0.0
19974	Andrew Murcia	0.0	0.0
8259	Hugh Farmer	0.0	0.0
17716	Anthony Matassa	0.0	0.0
19659	John Moriarty	0.0	0.0
32506	Richard Francis	0.0	0.0
29111	Robert Tyson	0.0	0.0
1791	Chas Bennette	0.0	0.0
17146	Robert Malone	0.0	0.0
17242	Jerry Manning	0.0	0.0
17289	Michael Maratto	0.0	0.0
32467	Jonathan Boyd	0.0	0.0
33755	Phyllis Butcher	1.0	0.0
333	Leroy Alexander	0.0	0.0
18597	James Mcglynn	0.0	0.0

only showing top 20 rows

Fig 2: This is a list of sample vertices of the main graph

```
+-----+-----+------------+
|  src|  dst|relationship|
+-----+-----+------------+
|23825|31702|           2|
|23825|30455|           2|
|23825|30910|           2|
|23825|25998|           2|
|23825|25514|           2|
|23825|28065|           2|
|23825|28560|           2|
|23825|25496|           2|
|23825|31211|           2|
|23825|26045|           2|
|23825|24306|           2|
|23825|26331|           2|
|23825|24570|           2|
|23825|27486|           2|
|23825|24938|           2|
|23825|30245|           2|
|23825|27995|           2|
|23825|29552|           2|
|23825|27860|           2|
|23825|27958|           2|
+-----+-----+------------+
only showing top 20 rows
```

Fig 3: This shows the sample edges in the main graph

We then use this main graph to create sub graphs based on units. To choose a unit to analyse, we execute a query to find the summed allegation counts of each unit. As detailed in the analysis, we choose the unit with the most allegations to see how the allegation is distributed. The query output to choose the unit looks as follows.

```
(10079, 9)
(8575, 5)
(8536, 19)
(8380, 7)
(8329, 23)
(8008, 3)
(7989, 212)
(7694, 2)
(7642, 208)
(7210, 8)
(7166, 17)
(6978, 6)
(6949, 149)
(6869, 20)
(6864, 4)
```

Fig 4: This is a sample of all the units present in the CPD in decreasing order of allegation counts

From this query we choose unit 9 - the unit with the most number of allegation counts and check its distribution.

The subgraph of this unit has vertices as follows:

id	officer_name	allegation_count	trr_count
17816	Edward May	136.0	9.0
2015	Eugene Bikulcius	102.0	15.0
3897	Thomas Carey	100.0	17.0
3764	Sean Campbell	90.0	33.0
11190	Mahmoud Haleem	87.0	28.0
19591	Richard Moravec	70.0	8.0
19407	David Montes	70.0	3.0
18205	John Mc Govern	67.0	7.0
25039	Kevin Ryan	65.0	8.0
22828	Michael Poppish	64.0	2.0
29833	Corey Walker	61.0	19.0
4758	John Clark	60.0	32.0

Fig 5: This is a sample of vertices in the subgraph

We then plot the distribution of allegations and trr against officers that belong to the generated subgraph.

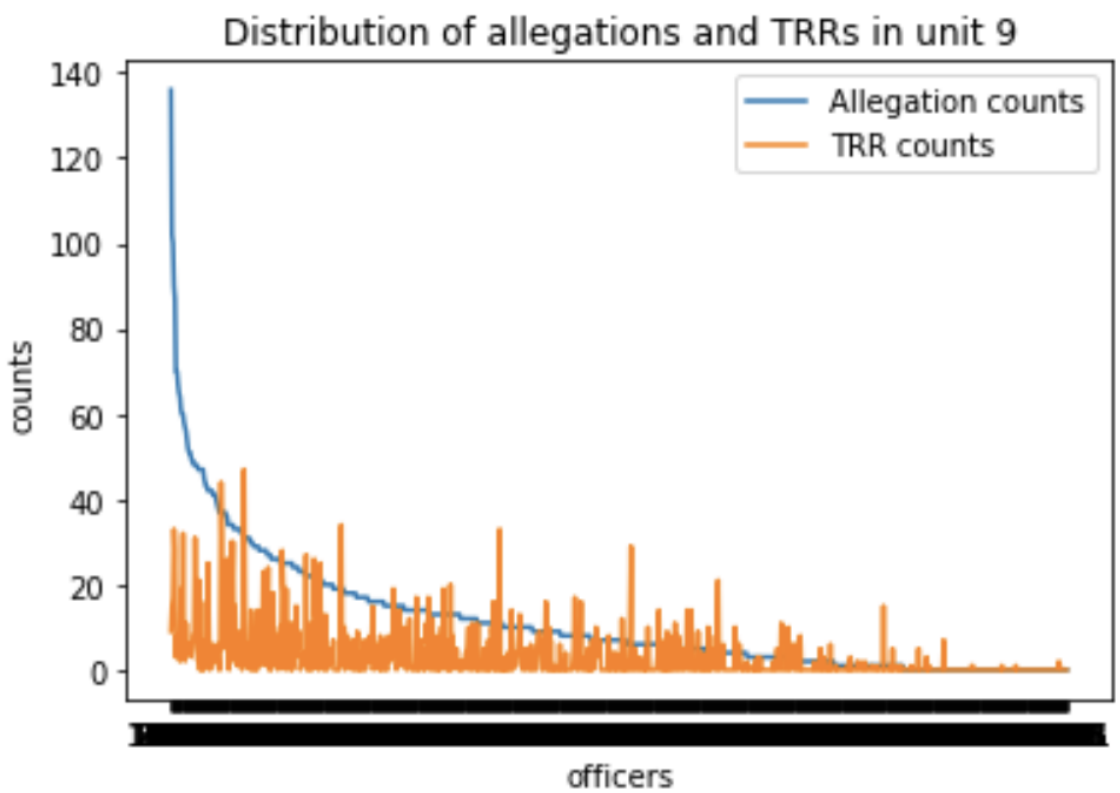

Fig 6

We repeat the same process for unassigned officers and the plot distribution is as follows:

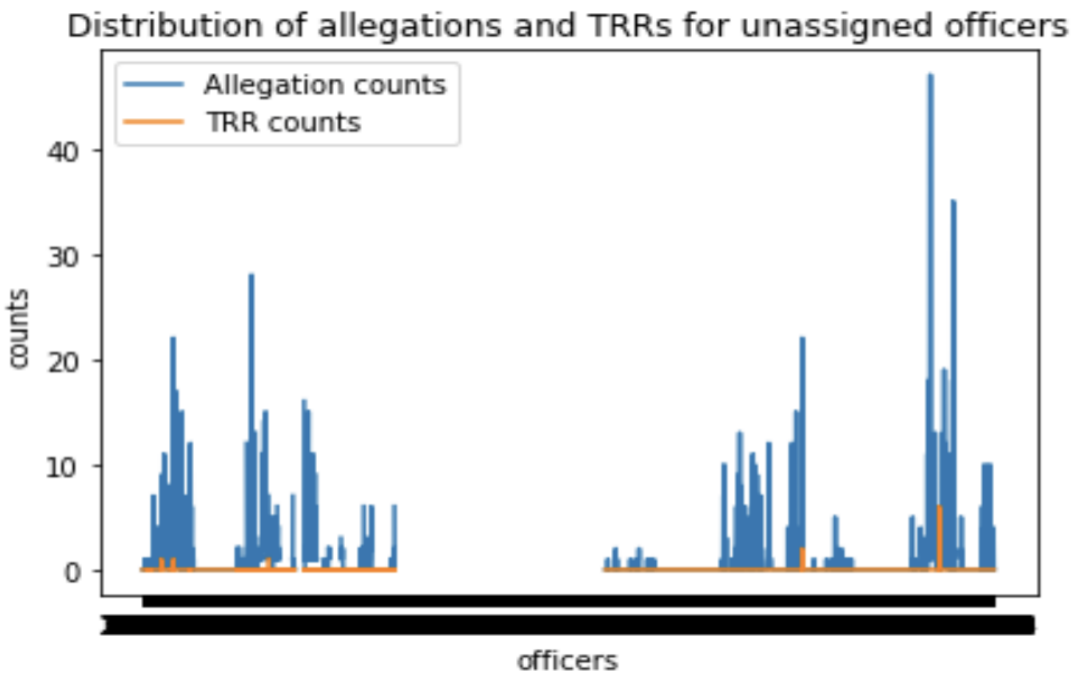

Fig 7

Conclusion:

We first choose unit id 9 that operates in the 8th district - Chicago Lawn. As a unit they have contributed to 10079 allegations in the total of Chicago allegations. This unit has a total of 895 officers. This unit makes up 2% of the officers in Chicago. Their allegation counts make up 4% of all the allegations in the Chicago Police Database.

Insights from the plot distribution for unit 9 (Fig 6):

1. It is evident from the plot distribution for unit 9 that the counts are extremely high for some of the officers whereas it is almost 0 for others.
2. We can infer that the elimination of these officers from this dataset will reduce the allegation contribution of this unit drastically.

Insights from the plot distribution of unassigned officers (Fig 7) :

1. This plot is pretty erratic with one extreme peak with an allegation count>40.
2. We cannot really draw any significant conclusions from this data separately but 40 is still a large number and should be addressed with relevant data to that officer. To do so, it is imperative that the dataset is improved to provide clarity on why these officers don't have a unit id.

On inspecting the rest of the units we further observed that all the units have a distribution similar to unit 9. From this we can infer that none of the units are entirely engaging in misconduct instead it is certain officers of the unit that are contributing to the allegations.

Possible reasons for this could be:

1. The officers with high counts are the "rotten apples" in that unit.
2. The officer could be a beat officer who has more engagement with the public than a supervisory agent who will consequently have a low allegation count due to the nature of the job.

These distributions are also muddled due to the assumptions we made with respect to allegation counts. To demystify these assumptions we want to know how important these officers with high allegations are, in the network of officers. Which is what we explore in the next question.

Question 2: If certain officers are the problem in their respective units we want to magnify this network and investigate which officers are the ring leaders, and at the center of co-accusals, influencing other officers?

Analysis:

As an extension to Question 1 , this question aims to identify the ring leaders (if present) within the police network. The identification of these ring leaders is purely based on the coaccusal counts and their connectivity with other officers. This removes the analysis from the assumptions made about allegation counts.

Discussion:

To answer this question, we created dataframes for officers (to be set as nodes) and coaccusals (to be set as nodes) after setting up our colab and cpdb instance. This was followed by the plotting of a graph frame with id, officer_name, allegation_count, trr_count from data_officer table set as vertices and the edges containing the relationship between two officers (called src and dst) from data_officerallegation table that pointed to the count of coaccusals performed by the two officers together.

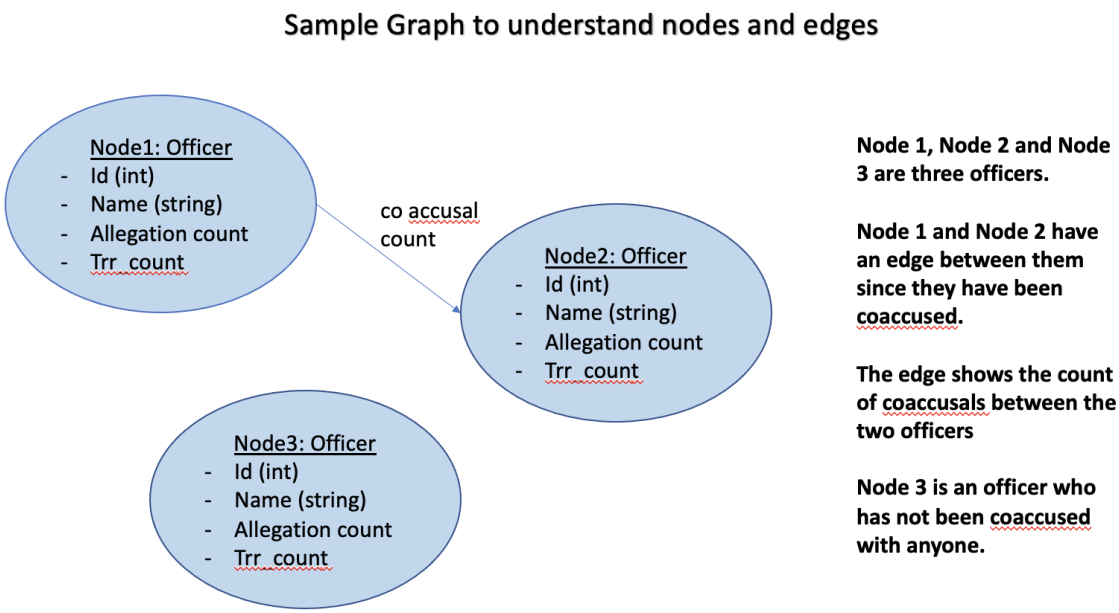

Fig 8: A sample graph illustrating node and edge definitions

id	officer_name	allegation_count	trr_count
32312	Randall Ryan	10.0	4.0
32358	Kevin Stoll	24.0	5.0
32480	David Lavin	0.0	0.0
32491	Kevin Rake	0.0	0.0
2680	Richard Brandt	0.0	0.0
29553	Hugh Versteegh	0.0	0.0
19974	Andrew Murcia	0.0	0.0
8259	Hugh Farmer	0.0	0.0
17716	Anthony Matassa	0.0	0.0
19659	John Moriarty	0.0	0.0
32506	Richard Francis	0.0	0.0
29111	Robert Tyson	0.0	0.0
1791	Chas Bennette	0.0	0.0
17146	Robert Malone	0.0	0.0
17242	Jerry Manning	0.0	0.0
17289	Michael Maratto	0.0	0.0
32467	Jonathan Boyd	0.0	0.0
33755	Phyllis Butcher	1.0	0.0
333	Leroy Alexander	0.0	0.0
18597	James Mcglynn	0.0	0.0

only showing top 20 rows

Fig 9: Sample output of vertices in our graph frame

```
+-----+-----+------------+
|  src|  dst|relationship|
+-----+-----+------------+
|12478|32166|          53|
| 8562|27778|          47|
| 1553|10724|          43|
| 2725|21703|          41|
| 3605|14442|          41|
|12479|20713|          40|
| 8658|13788|          38|
|14045|15502|          37|
| 2369| 7015|          36|
| 1553|16699|          35|
|12074|12825|          35|
| 8852|31119|          35|
|15360|23239|          34|
|13361|20150|          34|
| 8562|23841|          32|
| 8562|18206|          32|
|32265|32347|          32|
|31882|32401|          32|
|31119|32336|          32|
|27415|32172|          31|
+-----+-----+------------+
only showing top 20 rows
```

Fig 10: Sample output of edges in our graph frame

We then used the Pagerank algorithm available in GraphX that uses the Pregel interface and runs PageRank until convergence to identify the most important officers (ring leaders) in terms of coaccusals (refer Fig 11 for output).

id	officer_name	allegation_count	trr_count	pagerank
32442	John Zinchuk	23.0	7.0	165.10729586041256
32440	Mark Zawila	34.0	0.0	116.94160976162037
32425	Perry Williams	27.0	2.0	98.30895865468763
32350	Robert Spiegel	20.0	31.0	93.89434875347713
32410	Joseph Watson	29.0	2.0	93.08113520836221
32430	Michael Wrobel	22.0	29.0	91.40650266023003
32074	Ronald Jenkins	46.0	3.0	90.96964645853048
32284	Mark Reno	76.0	6.0	88.61011443215773
32351	Boonserm Srisuth	25.0	4.0	85.74844965857213
32433	Kenneth Yakes	29.0	0.0	82.5344126116758
32419	Eric Wier	18.0	0.0	78.00667585859934
32384	Edwin Utreras	47.0	8.0	77.30836139098388
32435	Mohammad Yusuf	22.0	11.0	76.78881509893944
32413	Carl Weatherspoon	69.0	5.0	75.1520442159433
32337	Louis Silva	21.0	19.0	75.00181054882047
32431	Albert Wyroba	15.0	6.0	74.79734661303651
32289	John Rivera	44.0	2.0	73.23421927205004
32401	Joshua Wallace	45.0	10.0	72.4657128499791
32375	James Triantafillo	31.0	32.0	65.51924756708055
32436	Edmund Zablocki	28.0	7.0	62.94909260087443
32406	Reginald Ward	17.0	1.0	58.778841638641104
32324	Kathleen Schmidt	15.0	0.0	58.452209975107635
32334	Terrence Shields	52.0	0.0	36.56263056790429
32397	James Vins	39.0	0.0	35.39424295705865
32388	Marc Vanek	12.0	5.0	35.329469344829285
32412	Adam Wazny	6.0	10.0	35.032761213464745
32255	Gerardo Perez	76.0	23.0	34.497479224060534
32022	John Hamilton	26.0	1.0	33.703068335912995

Fig11: Output of Pagerank calculated based on coaccusal counts of an officer in descending order

The pagerank calculated above was then further analyzed and understood by visualizing the graph using NetworkX based on motif finding for a particular officer. For brevity purposes we chose to showcase in our findings, the officer with the highest pagerank score of 165.07 (Officer John Zinchuk with ID 32442- refer fig. 12) and an officer with a lower mid range pagerank score of 35.032 (Officer Adam Wazny with ID 32412- ref fig.13) for a contrasting effect.

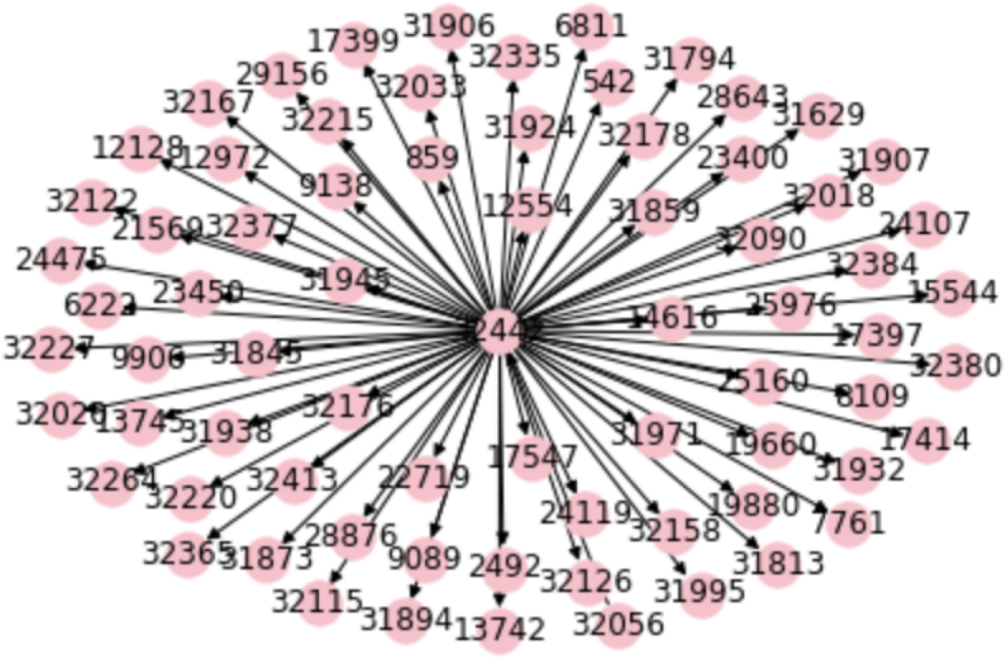

Fig12: Officer id '32442' (John Zinchuk) coaccusal network graph

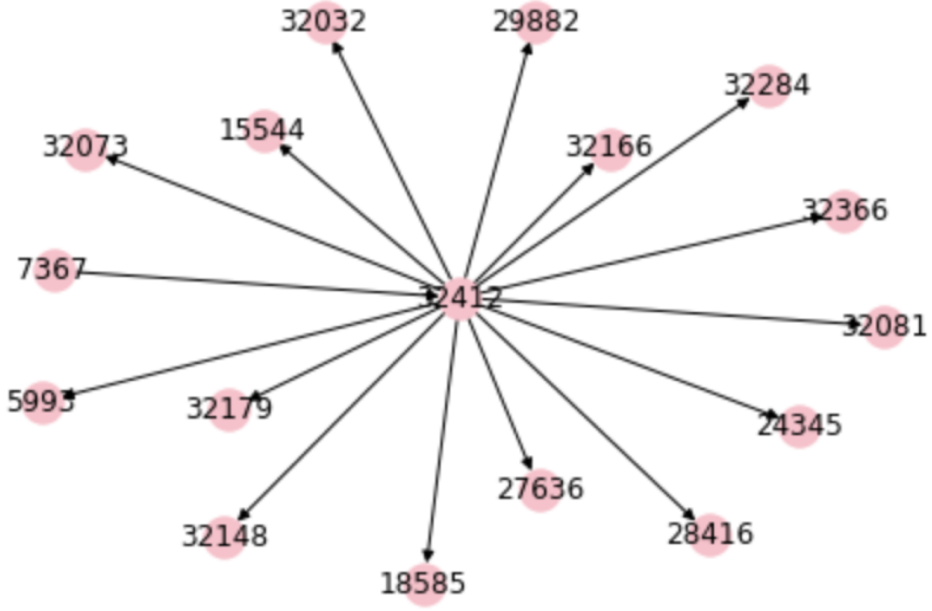

Fig13: Officer id '32412' (Adam Wazny) coaccusal network graph

From Fig 12 and 13 we can clearly see the difference between the coaccusal network density of both the officers. This helps us understand how well connected a ring leader (for eg-32442) is in comparison to other officers (for eg-32412) and how focusing on ring leaders in an officer's network in terms of reform is a potential avenue of improvising the reform implementation.

So to restrict the analysis we see the top 20 officers that are the most connected (refer Fig14). This means officers that have pagerank greater than 60.

	id	officer_name	allegation_count	trr_count	pagerank
804	32442	John Zinchuk	23.0	7.0	165.107296
3816	32284	Mark Reno	76.0	6.0	88.610114
3946	32384	Edwin Utreras	47.0	8.0	77.308361
4919	32074	Ronald Jenkins	46.0	3.0	90.969646
5326	32410	Joseph Watson	29.0	2.0	93.081135
8485	32350	Robert Spiegel	20.0	31.0	93.894349
12615	32440	Mark Zawila	34.0	0.0	116.941610
13980	32430	Michael Wrobel	22.0	29.0	91.406503
16681	32436	Edmund Zablocki	28.0	7.0	62.949093
18458	32431	Albert Wyroba	15.0	6.0	74.797347
18944	32433	Kenneth Yakes	29.0	0.0	82.534413
19004	32337	Louis Silva	21.0	19.0	75.001811
19523	32425	Perry Williams	27.0	2.0	98.308959
21261	32289	John Rivera	44.0	2.0	73.234219
26420	32435	Mohammad Yusuf	22.0	11.0	76.788815
27414	32351	Boonserm Srisuth	25.0	4.0	85.748450
27907	32401	Joshua Wallace	45.0	10.0	72.465713
29791	32419	Eric Wier	18.0	0.0	78.006676
30840	32413	Carl Weatherspoon	69.0	5.0	75.152044
31966	32375	James Triantafillo	31.0	32.0	65.519248

Fig14: 20 ring leaders identified based on pagerank

<u>Conclusion:</u>

Despite the possibility that post reform interventions like COPA there could be an increase in count of misconduct in various categories due to more reporting by citizens and inadvertent increase in false allegations made against the officers; having observed the findings of pagerank algorithm we can conclude that if we were to eliminate the identified ring leaders or focus on reforming them particularly it would positively impact the police units and the neighbourhood

misconducts in these cases. We say this on the basis that just these 20 officers are involved in a total of 2581 coaccusal cases which is a significant number.

Final Conclusion:

Answers to questions we asked:

1. Are certain officers the cause of high misconduct counts in a certain district, or is it a unit-wide issue?

a. We know from research, and inferences from previous checkpoints that all the reforms have a broad target with respect to implementation and focus. We also know that disciplinary actions against officers are weak in comparison to their sustained allegations. This checkpoint drives in the fact that the contributors to allegation counts are certain officers in a unit and not unit wide. We further also identify certain officers who are ring leaders in the CPD network. Reform targeted towards these officers can dismantle the vicious network of engaging in bad behaviour.

2. How does this affect reform?

The conclusion from the previous question should translate into reformation activities. The CPD advised by data should focus on these certain officers and reprogram the CPD network to affect a positive change in the allegation counts.

From Checkpoints 1-4 we have deeply examined reforms and officer allegation count relationships. Subsequently, we hope to explore the tonality of accusals/allegation summaries before and after reformations to see how the tone or wording of the officer's responses may have changed since these reformations. With the insight we hope to retrieve from the aforementioned analysis, we can strengthen reformation plans to ensure the reforms are not met with resistance.