Liqian Ma
Wentao Yao
Xingbang Liu

Checkpoint 4: Graph Analytics

OVERVIEW & PURPOSE

Graph analytics can be very useful in analyzing relationships between different groups of people. We can create nodes based on their income, race, neighborhood, and other attributes. After building the graph, we can analyze interactions among different nodes and even graphlets.

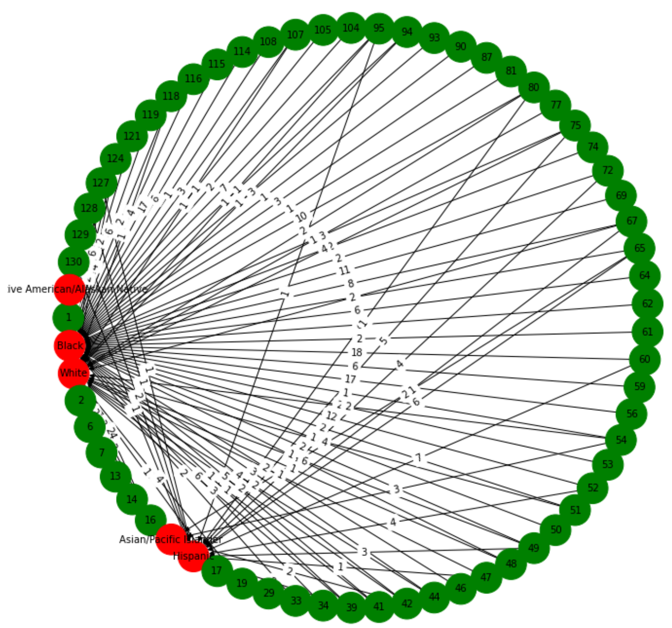
Question1

Making nodes of officers and victims by their income, race, locations, and even unsupervised machine learning models to learn the cluster and see if there is a potential connection between officers and victims.

1.1 Learn the Connection from Race

1.1.1 Graph Visualization Sample

In this section, we plot the visualized graph of the connection of the officer and the victim by race with part of the data.



Conclusion from graph

Since the graph is huge, it is not possible to plot the whole graph here. However, we still can see there is tend those officers are more likely to offense black people in the sample graph. Therefore, we may find the potential connection between the victims and the officer by the race with the whole data.

1.1.2 Graph Analysis on Race

Similarly, like the graph visualization, but we use all data now.

Graph Analysis

For this graph, ingress is the number of CRs complained by a race, and outDegrees is the number of Crs an officer received.

id	outDegree
13937	89
14442	88
32159	87
3764	86
3605	86
17613	85
21098	81
25898	81
32164	79
17647	76
8138	76
27415	75
16385	75
10152	75
32213	75
31631	74
32016	74
31872	74
31119	73
3897	72

only showing top 20 rows

id	inDegree
Black	67923
White	20519
Hispanic	12128
Asian/Pacific Isl...	768
Native American/A...	108

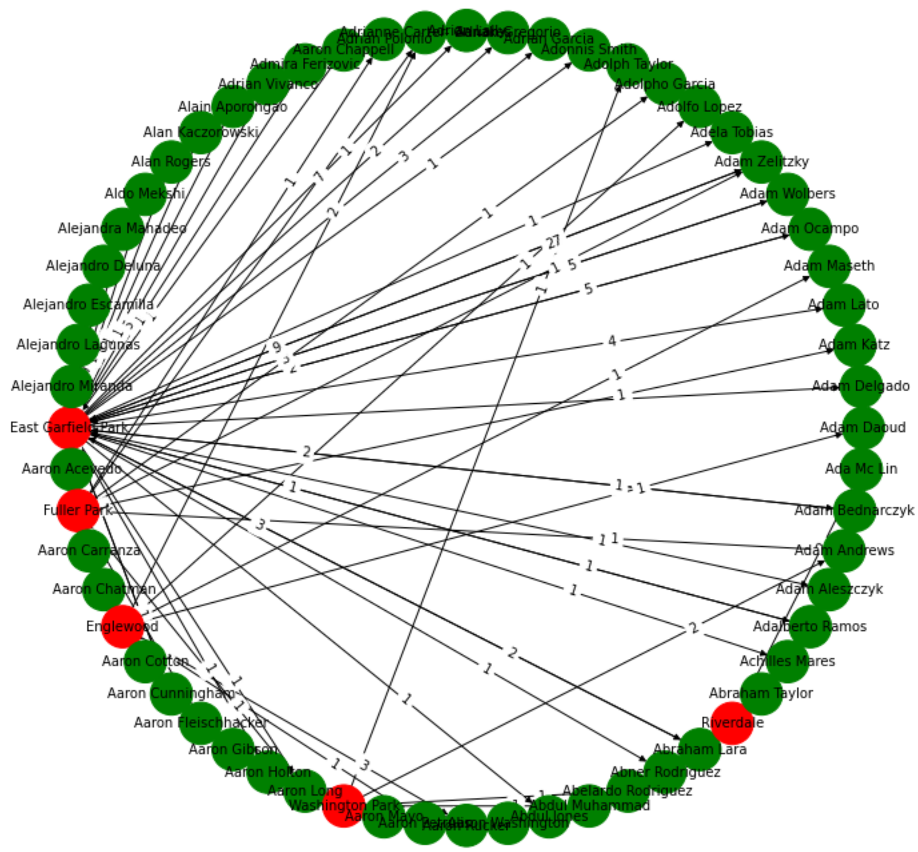
1.1.3 Conclusion on Race

We can find that there is a high volume of complaints from black people, since the indegree is 67923 which is 3 times of the second highest complaints race, white, which has 20519 complaints. So, we may assume that there is an over-policing based the race bias due to the extremely large number of complaints from a specific race. However, we are not interested in the bias, this section is only used for proving our main theme, " Is there over-policing in low socio-eco status neighborhoods? " From a different aspect. There is more discussion in the following sections.

1.2 Learn the Connection from Location

1.2.1 Graph Visualization Sample

In this section, we plot the visualized graph of the connection of the officer and the victim by the location with part of the data.



Conclusion from graph

Since the graph is huge, it is not possible to plot the whole graph here. However, we still can see there is tend for officers to have more TRRs and CRs from some communities (East Garfield Park on this graph). Therefore, we may find the potential connection between the victims and the officer by the location.

1.2.2 Graph Analysis on Location

Similarly, like the graph visualization, but we use all data now.

CRs:

src	dst	relationship
Austin	Alan Krok	CR
Englewood	Ruth Johnson	CR
Chicago Lawn	Michael Mayhew	CR
South Deering	Nora Collins	CR
Woodlawn	Tracy Quarles	CR
East Garfield Park	Gerard Murphy	CR
Near North Side	Jose Zuniga	CR
Near North Side	Frank Cool	CR
Norwood Park	Jeffrey Fronczak	CR
Garfield Ridge	George Mc Murray	CR
Near West Side	Debra Ippolito	CR
Lower West Side	Jack Dedore	CR
Pullman	Joseph Buss	CR
Belmont Cragin	Latonia Harris	CR
Lincoln Square	Gail Martin	CR
Austin	Marienne Perry	CR
Englewood	Marilyn Uldrych	CR
Auburn Gresham	Michael Devine	CR
Beverly	George Porter	CR
Ashburn	Nicola Zodo	CR

only showing top 20 rows

We can split the graph by its relationship between src and dst. For CRs, inDegress is the number of CRs an officer received, and outDegrees is the number of Crs a community complains.

id	inDegree
Joe Parker	129
Jerome Finnigan	124
Edward May	114
Charles Toussas	114
David Brown	109
Kevin Osborn	108
Maurice Clayton	107
Glenn Evans	106
Adam Zelitzky	105
Jerome Turbyville	99
Robert Smith	98
James Grubbs	93
Robert Johnson	93
John Carney	88
Gregory Jackson	87
Tyrone Jenkins	87
Broderick Jones	87
Kevin Ryan	85
Eugene Bikulcius	85
Edward Howard	83

only showing top 20 rows

id	outDegree
Austin	10470
West Englewood	7979
Loop	7927
Near West Side	7411
Near North Side	7327
Auburn Gresham	6009
Humboldt Park	5760
North Lawndale	5503
Englewood	5360
West Town	5267
South Shore	4932
East Garfield Park	4900
New City	4891
Roseland	4763
Chicago Lawn	4741
Logan Square	4368
Lake View	4114
Greater Grand Cro...	4088
Uptown	3833
Woodlawn	3752

only showing top 20 rows

TRRs:

src	dst	relationship
Michael Jacob	Rogers Park	TRR
Agustin Cervantes	Avondale	TRR
Walter Ware	North Lawndale	TRR
John Flisk	North Lawndale	TRR
David Morales	North Lawndale	TRR
Demosthen Balodimas	Belmont Cragin	TRR
Timothy Gilbert	East Garfield Park	TRR
Thomas Davey	Near West Side	TRR
Brian Ferguson	Humboldt Park	TRR
Paul Meagher	Austin	TRR
Kent Erickson	Uptown	TRR
Martin Teresi	Beverly	TRR
Raymond Wilke	Beverly	TRR
Nicolas Chapello	Irving Park	TRR
Kerry Mc Guire	Irving Park	TRR
Michael Leverett	East Garfield Park	TRR
Jeffrey Zwit	East Garfield Park	TRR
Timothy Gilbert	East Garfield Park	TRR
Joseph Simon	Humboldt Park	TRR
Slawomir Plewa	Humboldt Park	TRR

only showing top 20 rows

We can split the graph by its relationship between src and dst. For TRRs, inDgress is the number of TTRs happen in the community, and outDegrees is the number of TRRs an officer has.

id	inDegree
Austin	5721
Humboldt Park	2848
West Garfield Park	2622
South Lawndale	2230
North Lawndale	2092
Near North Side	1721
Near West Side	1648
West Town	1607
East Garfield Park	1502
Belmont Cragin	1064
Lake View	1033
Rogers Park	928
North Park	771
Lincoln Park	765
Logan Square	760
West Ridge	757
Norwood Park	747
Uptown	703
Edgewater	576
Albany Park	520

only showing top 20 rows

id	outDegree
Cesar Kuri	67
George Granias	67
Richard Pellerano	66
Michael Walsh	64
Patrick Josephs	60
Peter Chambers	59
Robert Roth	56
Matthew Bouch	56
David Kleinfelder	55
Patrick Altwasser	54
John Dalcason	53
Bartholom Murphy	52
Lucas Wise	51
Christoph Cannata	51
Aaron Acevedo	51
Tomasz Zatora	51
Daniel Kolodziejski	50
Samuel Truesdale	49
Michael Tews	48
Erick Seng	48

only showing top 20 rows

1.2.3 Conclusion on Location

We can conclude that communities like Austin, West Englewood, and Loop have a high volume of complaint report to officers, and Austin, Humboldt Park, and West Garfield Park have a large amount of TRRs. From this result we can find in the high-income community, people are more likely to complain about the behavior of the police. People from low-income communities receive more "threats" of tactical response. One possible explanation is that people who live in high-income communities have time to report the misbehavior of over-policing officers. But in the low-income community, people have no power to against the over-policing. Anyway, a high amount of reports of tactical response shows that there is potential over-policing behavior in those areas. Combining with the result we find in Checkpoint 1, a community like West Garfield Park is a low-income area. Therefore, we can assume that there is over-policing in the socio-economy status community.

Question2

Network dynamics of co-accused in each cohort can be interesting. The analytics can be done with the following:

1. Make use of Triangle Count Algorithms for each cohort.
2. Make use of the Page Rank Algorithm to find the most connected officer in all cohorts.
3. How many CRs that officers have and how many co-accused for each cohort.
4. Compare the top k largest cohort of police officers in high and low socio-economy status.

And we will answer the following questions:

1. Who among the officers has the most triangle counts?
2. Who has the most page rank score?
3. Are there any communities in the officers?
4. What are the allegation reports number for those officers inside a cluster?
5. What are the top large cohort of police officers in high and low socio-economy status?

2.1 Prepare the Data

These queries are to draw co-accused officers from the allegation database. The basic logic is to join the allegation table with itself on the condition of the same allegation id and unequal officerid.

Nodes can be generated with data_officer table or allegation id by counting the number of allegation id. Here we chose data_officer table by removing Nan or 0s on allegation_count.

Note: These queries are copied and modified from the GraphX demo class, which shares a similar analysis goal as ours.

```
+----+----------------+---------------+-----+
|  id|    officer_name|allegation_count|label|
+----+----------------+---------------+-----+
|  29|    Henry Abrams|              6| 6534|
| 474|Ignacio Alvarado|              7|28838|
| 964|  Colleen Austin|              6| 3744|
|1677|    Chad Behrend|             25|17372|
|1950|    Thomas Beyna|             22|  442|
|2214|    Calvin Blunt|             21|28273|
|2250|Kathleen Boehmer|              2|17372|
|2453|   Joseph Boston|             59|28838|
|2509|  Rosalind Bowie|             14|32382|
|2529|     Emmett Boyd|             11|12644|
|3091|  Michael Browne|              9|32041|
|3506|John Butterfield|              1| 3506|
|3764|   Sean Campbell|             90|28838|
|4894|Danyelle Cochran|              1| 4894|
|5385|  Gerald Corless|              2|27851|
|5409|  Rodolfo Corona|              4|17372|
|5556| Ramon Covington|              6|11980|
|7225|     Judy Dotson|              2| 7225|
|7279| Terrence Downes|              6|17372|
|7747|     Donald Eddy|              4| 7747|
+----+----------------+---------------+-----+
only showing top 20 rows

There are 2809 communities in this sample graph.
```

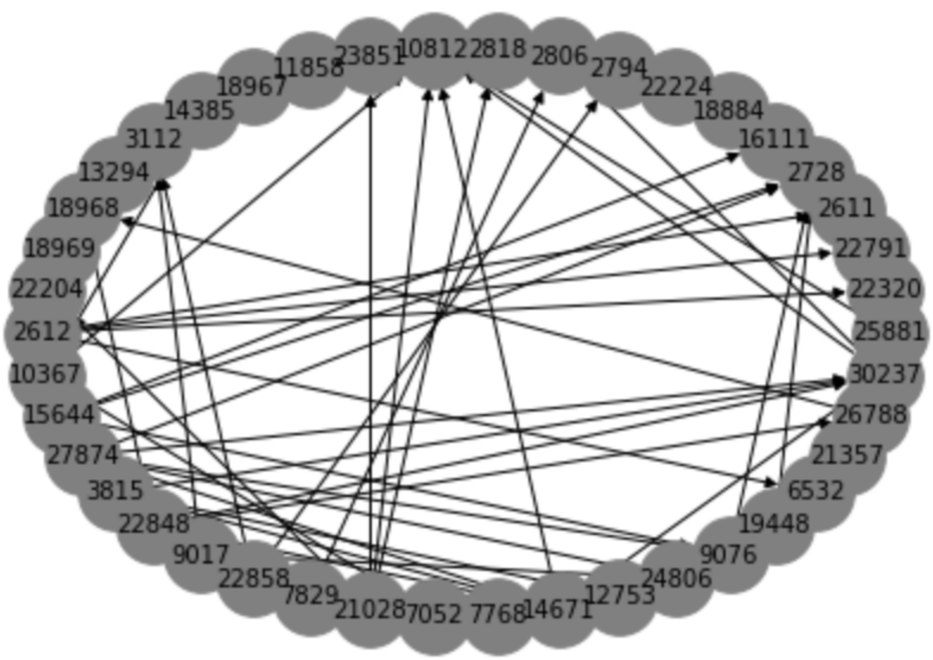
Recognizing the largest comminutes is important. So, we ranked the label propagation algorithm result by sorting descending the number of members in the community.

label	count
17372	8316
3744	1636
29511	1224
11980	652
28273	596
28838	450
32014	364
32068	323
32382	257
26622	257
13631	256
14106	243
32274	211
32041	207
6534	187
18915	186
23787	173
2981	162
21912	155
23033	115

only showing top 20 rows

After identifying those top big communities, we are also interested in how the community is constructed and its internal architecture.

We plotted the 22809 community which is consisted of over 50 nodes. It is clear to us those officers 2612, 30237, and 21028 are among those "leading" nodes with multiple indegrees and outdegrees inside the clique.



2.2 Triangle Count analysis

The triangle counting algorithm is to count the triangle-like relationship among 3 nodes that have connected in pairs. We want to find out those outstanding nodes in the graph which have a lot more triangle counts.

id	count
33748	0
33751	0
33724	0
33798	0
33755	0
33746	0
33749	0
33737	0
33725	0
33738	0
33728	0
33752	0
33711	0
33723	0
33750	0
32312	37
32358	109
33753	0
33758	0
33709	0

only showing top 20 rows

count	id	officer_name	allegation_count
32118	6315	Terence Davis	38
32117	3033	Raimondo Brown	17
32073	3744	Derek Campbell	8
27855	18042	Donald Mc Coy	22
27823	441	Fernando Alonzo	16
23900	21530	Michael Overstreet	56
23518	27349	Charles Stanton	11
23499	5180	Stephen Conner	9
23487	5667	Jerry Crawley	30
23477	16747	Evetta Lundin	7
23475	8844	Thomas Flynn	19
23472	23654	Lloyd Reid	4
23472	14750	William Kissane	23
20185	19856	Ronald Muhammad	11
19322	8138	Glenn Evans	132
18773	29882	Fred Waller	49
18648	28273	James Taylor	36
18602	28459	Curtis Thomas	36
18539	5577	Michael Cox	20
18502	30841	Teresa Williams	37

only showing top 20 rows

In this part, we sorted all the nodes according to their triangle counts. We can see over 20 nodes appearing in over 18,000 triangle relationships, which indicates strong community leadership potential like officers 6315 and 3033.

2.3 Page Rank analysis to find key nodes

Page rank algorithm is developed to find out important nodes inside a graph by iterations of calculations of the possibilities to get to the node by starting randomly.

id	officer_name	allegation_count	pagerank
32442	John Zinchuk	23	127.52903862900281
32440	Mark Zawila	34	90.32581504596747
32425	Perry Williams	27	75.93393690155354
32350	Robert Spiegel	20	72.52408784740014
32410	Joseph Watson	29	71.8959609008098
32430	Michael Wrobel	22	70.6024730642657
32074	Ronald Jenkins	46	70.26504490198167
32284	Mark Reno	76	68.44254003101547
32351	Boonserm Srisuth	25	66.23218732944623
32433	Kenneth Yakes	29	63.74966193544296
32419	Eric Wier	18	60.25243358901534
32384	Edwin Utreras	47	59.71305480353141
32435	Mohammad Yusuf	22	59.31175673367685
32413	Carl Weatherspoon	69	58.047513284732524
32337	Louis Silva	21	57.93147265165182
32431	Albert Wyroba	15	57.773544505418506
32289	John Rivera	44	56.566183401162725
32401	Joshua Wallace	45	55.97258828063104
32375	James Triantafillo	31	50.60713162542214
32436	Edmund Zablocki	28	48.62194138740303

only showing top 20 rows

From the above calculations, we can identify officers with significant impact in the graph. For example, officers 32442 and 32440 are a major part of the clique and maybe the "bad apple" in the organization.

2.4 The Correlations Between Police Cohort and CRs/TRRs

In this section, each police are counted for the time they had the same allegation with other police officers. The counted number will then be compared with the CRs and TRRs they gave and received to find the correlation between them. The goal of the correlation is to find whether police officers are more likely to misconduct when working as a group.

Graph Analysis

	id	inDegree	outDegree	officer_id	cohart count
0	Joe Parker	129	0.0	21837.0	None
1	Jerome Finnigan	124	1.0	8562.0	None
2	Edward May	114	2.0	17816.0	None
3	Charles Toussas	114	0.0	NaN	None
4	David Brown	109	0.0	3005.0	None
...	...	...	...	...	...
21811	Ronald Truhlar	1	0.0	NaN	None
21812	Gregory Czyznik	1	0.0	NaN	None
21813	Anthony Alviani	1	0.0	NaN	None
21814	C Ahern	1	0.0	NaN	None
21815	Brittni Martinez	1	0.0	NaN	None

21816 rows × 5 columns

	member1	member2	co-case count
0	12478	32166	53
1	8562	27778	47
2	1553	10724	43
3	2725	21703	41
4	3605	14442	41

	id	inDegree	outDegree	officer_id	cohart count
0	Joe Parker	129	0.0	21837.0	27
1	Jerome Finnigan	124	1.0	8562.0	119
2	Edward May	114	2.0	17816.0	86
3	Charles Toussas	114	0.0	NaN	0
4	David Brown	109	0.0	3005.0	9
...	...	...	...	...	...
21811	Ronald Truhlar	1	0.0	NaN	0
21812	Gregory Czyznik	1	0.0	NaN	0
21813	Anthony Alviani	1	0.0	NaN	0
21814	C Ahern	1	0.0	NaN	0
21815	Brittni Martinez	1	0.0	NaN	0

21816 rows × 5 columns

For the above chart, we can get the correlation of CRs is 0.4151983851569962 and correlation of TRRs is 0.21054139647875614

Conclusion

The correlation between group allegation and complaint reports is positively correlated, and the group allegation is less correlated to tactical response reports. It is possible that when police officers are co-accused, they are more likely to have actual misconduct activity. It is because if they gave more tactical responses than receive complaints, or if they have an equal number of tactical responses and complaints, they would be less likely to have misconduct.