PART II: PUTSAFIRST
COMMUNITY DETECTION

Community detection in network theory is a graph partitioning problem where a group of nodes is identified such that the connections a node has within a group are more than the connections a node has outside a group.

Data Source

#PutSAFirst Reply Network - Largest Component

Data Transformation

Community detection techniques are typically applied to undirected networks. The direction in which a tie is sent or received is irrelevant; a connection is established as long as there is an interaction between users. The current network is therefore converted to an undirected network for the purposes of this analysis.

Method

Using pythons iGraph library, I apply the **Leiden** method to this community detection problem.

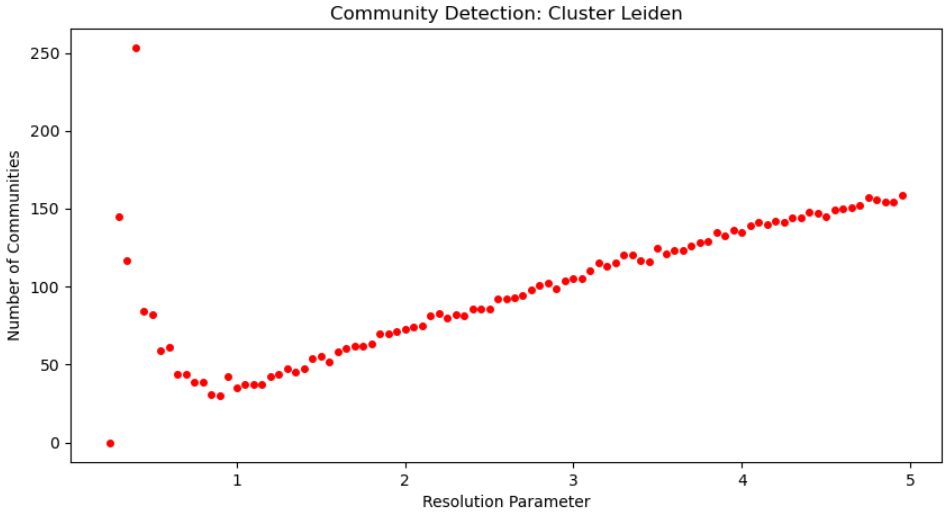
`Undirected Graph:`

```
IGRAPH U-W- 7606 19003
```

---

1 OPTIMISATION

To choose an optimal resolution parameter (gamma), I estimate the leiden algorithm and explore community outcomes across different values of gamma. Letting the data tell me which points are more optimal than others, I select gamma at the points in the graph where the number of communities detected reach a constant plateau or flatness.



---

2 DETECTION

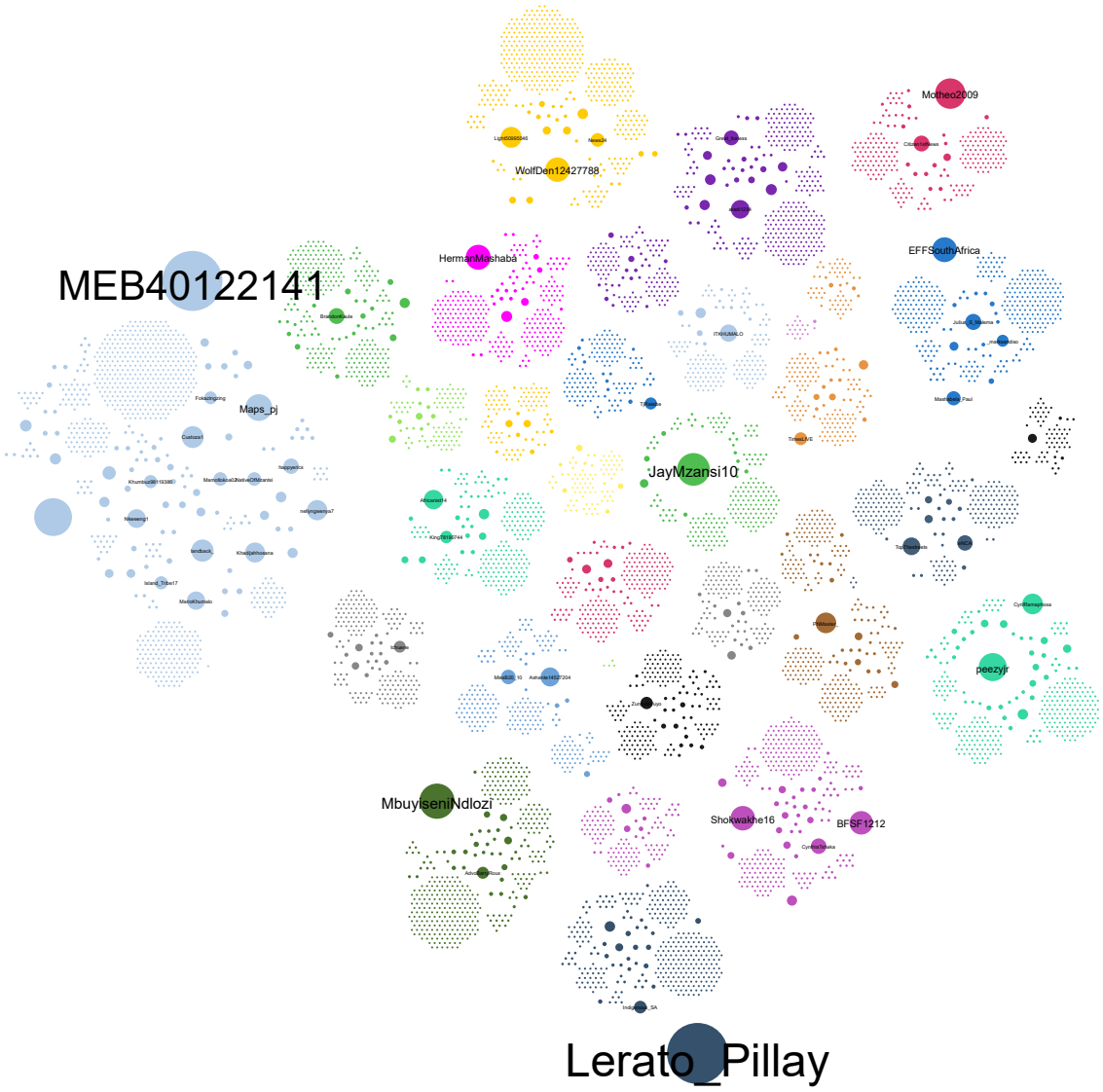
The point of optimised modularity is less clear in this case. Setting gamma at the default resolution parameter of gamma=1, I detect 35 communities with a modularity of 0.47. The detected communities are visualised below along with descriptions of the top 3 largest communities (by weighted degree).

```
--------

Communinities:
35

Modularity:
0.4629748237652841


--------
PutSAFirst Communities:
```



Top 3 Weighted Degree Communities:

Community 1:

Size:

1422

Density:

0.0019092752771121544

Average Degree:
4.35

Total Degree:
6182

Members(10*):
0 MEB40122141
1 nan
2 Maps_pj
3 landback_
4 Custoza1
5 nellyngwenya7
6 Khadijahhosana
7 Nkweeng1
8 MarioKhumalo
9 happyerics

Community 2:

Size:
866

Density:
0.0018288857146671294

Average Degree:
1.98

Total Degree:
1712

Members(10*):
0 WolfDen12427788
1 Light50995046
2 News24
3 MbalulaFikile
4 siza_mhayise
5 Pheaga1
6 VictoriaAfrica9
7 ssoshaah
8 LvovoSA
9 alfred_cabonena

```
--------------------------------

--------
Community 3:
--------

Size:
530

Density:
0.00363804972001284

Average Degree:
2.77

Total Degree:
1468

Members(10*):
0 peezyjr
1 CyrilRamaphosa
2 MYANC
3 PresidencyZA
4 SAPoliceService
5 tito_mboweni
6 ewnupdates
7 AlomNyama
8 GovernmentZA
9 Sibusisok16614
--------------------------------
*Top 10 highest degree in community.
```

---

3 EVALUATION

Using various pair counting scores, I compare community outcomes of the leiden model at gamma=1 and gamma=2.5, and compare the leiden model at gamma=1 and the louvain model.

3.1 Leiden (Gamma=1) vs Leiden (Gamma=2.5)

```
--------
Leiden (Gamma=2.5)
--------------------------------

Communinities:
90
```

Modularity:
0.4388048823643757

Comparison Scores:

Variation of Information:
0.5960006524376178

Adjusted Rand:
0.2732328442840795

Normalized Mutual Information:
0.5960006524376178

3.2 Leiden (Gamma=0.8) vs Louvain

Louvain

Communinities:
36

Modularity:
0.4555789102607928

Comparison Scores:

Variation of Information:
0.5419488041481584

Adjusted Rand:
0.39619142893310055

Normalized Mutual Information:
0.5419488041481584

---

4 SUMMARY

In this section, I use community detection methods to identify densely connected groups and communities present in the #PutSAFirst twitter reply network.

When searching for optimum modularity, I find the resolution parameter robust around gamma=1 and gamma=2.5. I detect 35 and 90 communities with a modularity score of 0.46 and 0.44 at these points, respectively. Similarly, using the louvain method, 37 communities are detected with a modularity score of 0.843. The coincidence scores from the evaluation and the modularity scores, which are below the 0.7 threshold, are all a sign of a poorly partitioned network. Furthermore, given the large number of communities detected, a reasonable or meaningful interpretation is difficult. In other words, attemping to characterise all 35 communities or predict each of their behaviour and influence in the discourse based on their memberships will unliklely result in an intuitive outcome.

To atleast get some sense of the communities detected, I examine the sizes, densities, and degrees of the top 3 largest communities as well as the top 10 members (sorted by *weighted degree*) assigned to each of the 3 communities. Going by the users and tweet content, I find **Community 1** to contain users in strong support of the #PutSAFirst movement, with user **MEB40122141** leading the charge. The second largest contributer to this community is listed as **nan**, which means *none* or missing. For some reason this account's username has been left out of the data set, entirely. If it were not for their username being mentioned in other users tweets as **mudzu_thabe**, it would not have been possible to identify them by name. There is also no record of them having sent any tweets. In fact, their significant contribution to the discourse is only a result of large numbers of users replying to their (ommitted) tweets. Currently the user's page has been suspended, which may be the reason for the missing data (especially if they were suspended prior to the data being extracted).

For interest, this detection outcome puts the **Lerato_Pillay** account at the top of list in the 4th largest community by total weighted degree.

In the next section, I classify users by the positions or roles they play within the discourse based on their patterns of interactions and other network features.