PART III: PUTSAFIRST
POSITIONAL ANALYSIS

Positional analysis is a subfield of network science concerned with the identification of actors or groups of actors who occupy similar positions or roles based on some feature or structure of the network.

Data Source

#PutSAFirst Reply Network - Largest Component

Data Transformation

I exclude observations with any missing data from the feature data set as they are not well handled in this clustering algorithm. Additionally, I use a data scaling method called "MinMaxScaler" to standardise the data.

Method

Using pythons scikit-learn library, I present feature based techniques that detect optimal **k-means algorithm** clusters.

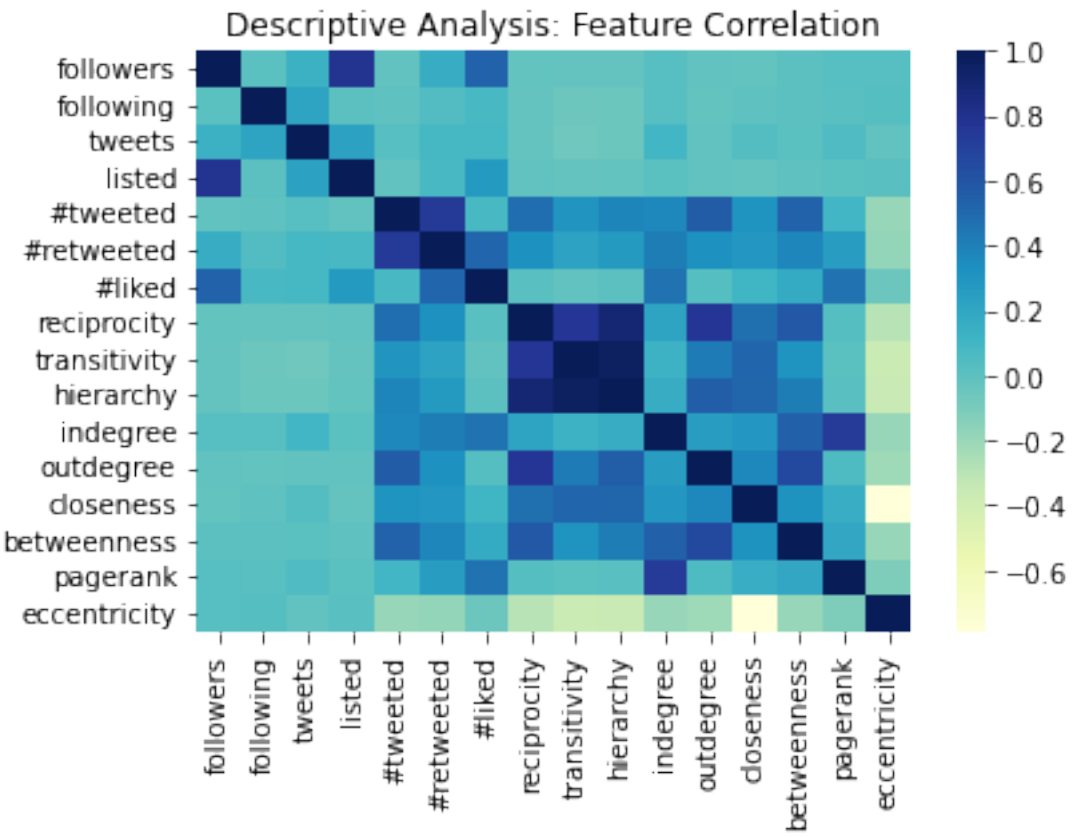
---

1 SELECTION

The 16 features selected for this analysis are detailed in the Part I: Descriptive Analysis. They cover node attributes of expansiveness (whether someone is a social butterfly or wallflower), attractiveness (whether someone is popular or unpopular), as well as relevance (whether someone is locally important or important gloably). For example, a person can be generally active on social media with a high number of tweets, while another person can be active with a high number of tweets at one particular point in time discussing one particular issue. Both of these traits are captured by the features: *tweets* (global) and *#tweeted* (local).

Notable observations from the correlation matrix:

- *#liked* number of tweets and being *listed* on someone's public lists is positively correlated with the number of *followers*.
- *#liked* number of tweets are positively correlated with the *indegree* of your tweets being replied to as well as the *pagerank* likelihood of being connected to other high rank positions.
- *#tweeted*, *reciprocity*, *hierarchy* are all positively correlated with the *outdegree* of replying to tweets and the *indegree* of your tweets being replied to.

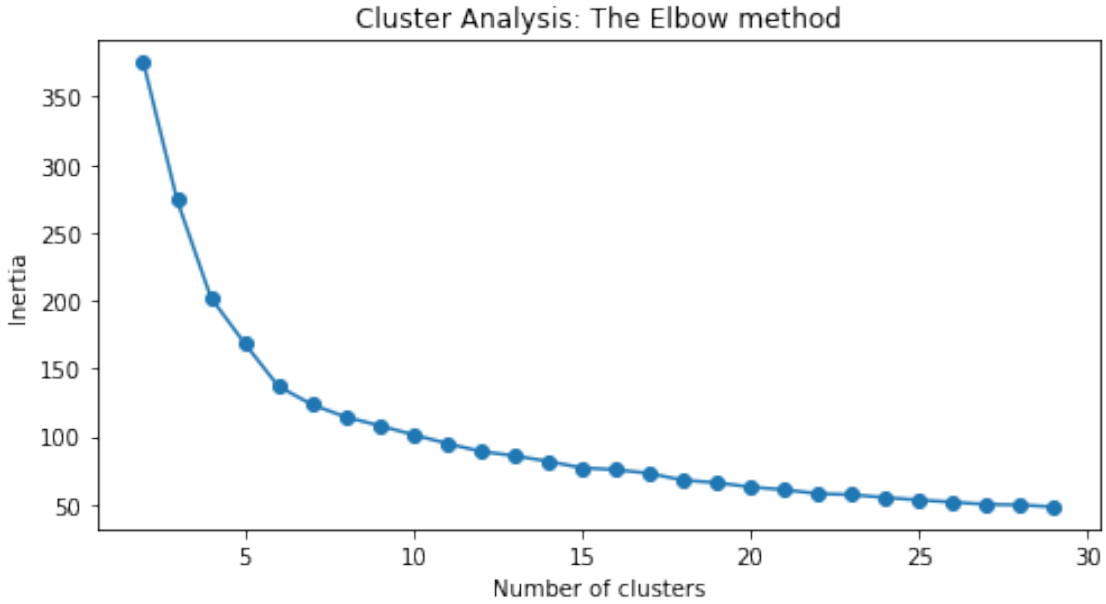


---

2 DETECTION

Using k-means clustering techniques, I detect the optimal number of clusters to fit the network data.

2.1 The Elbow Method

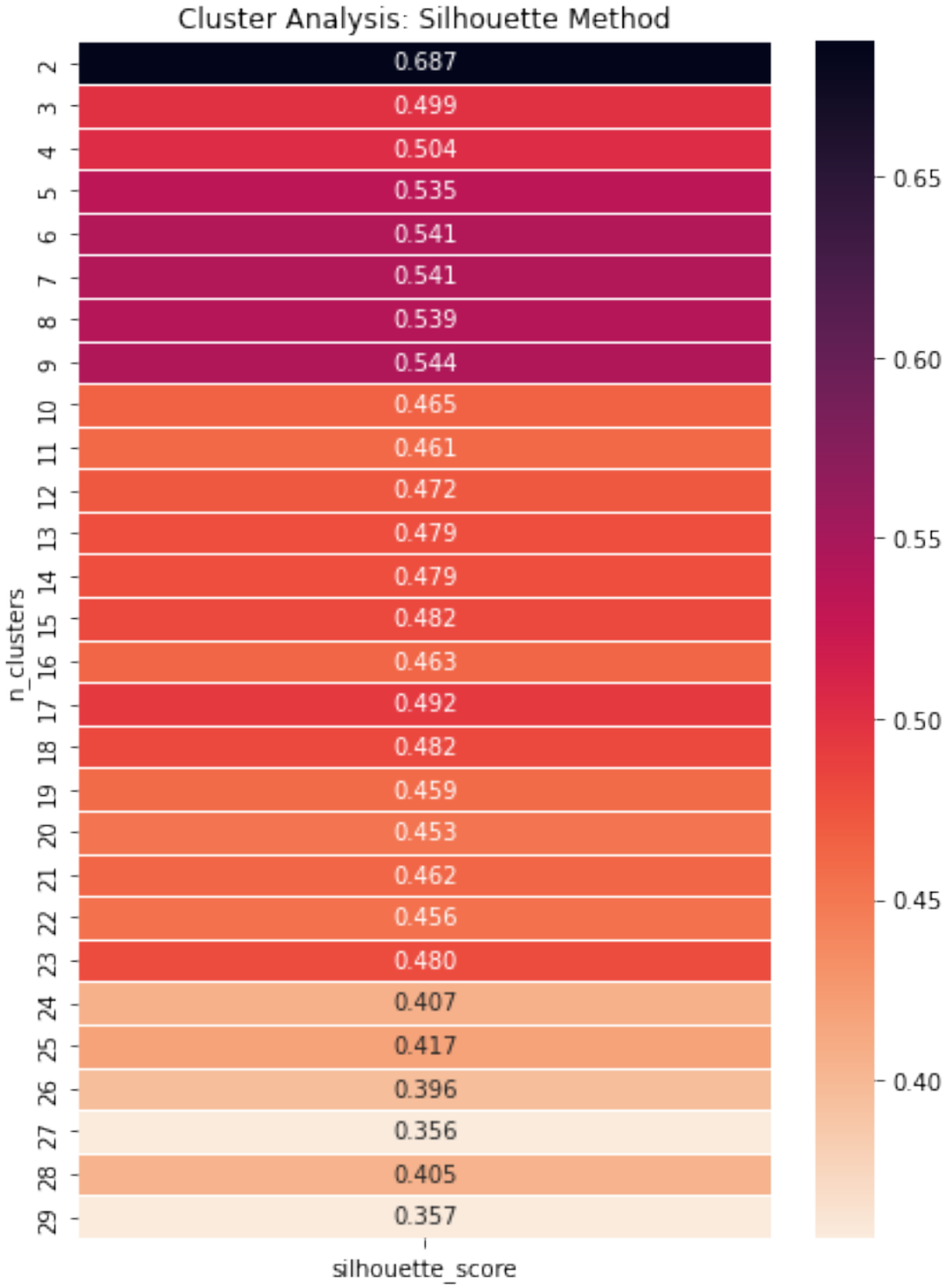


KneeLocator

Since it is not visibly clear, just by looking, which point of the curve is the knee or point of the maximum curvature. I run the *kneelocator* function and find the optimum number of clusters at at k-cluster=7.

```
7
```

2.2 The Silhouette Method



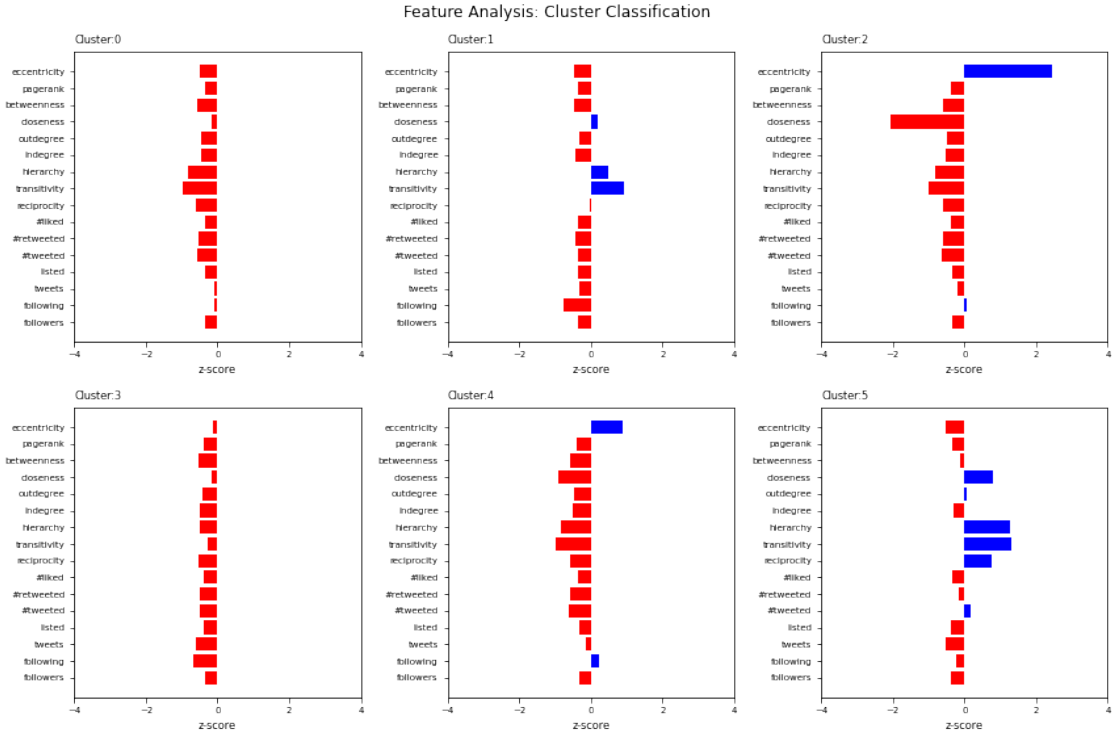
The highest silhouette score obtained for k-clusters between 2 and 30 is 0.687 at k-cluster=2.

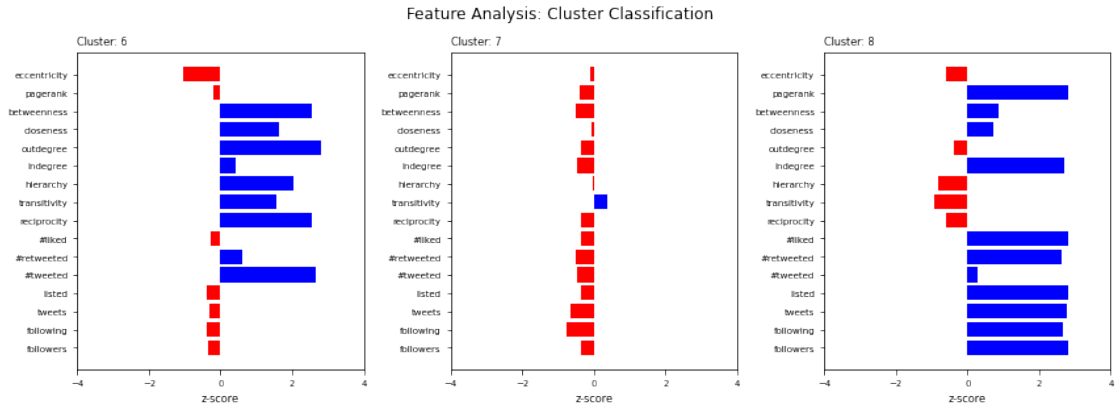
3 CLASSIFICATION

Using the elbow and silhouette method as a guide, I narrow the range of k-clusters and carefully examine the groupings that display the most consisent and meaningful result. I find k-cluster=9 to have the best grouping for this network data.

I classify the different clusters by plotting the mean scores of each feature in a cluster and comparing them to the mean scores of features in other clusters. Distinct clusters are identified by the features that display the largest deviation from the mean score.

3.1. Mean Score Plots





At k-cluster=9, 3 distinct clusters are detected. The remaining 6 exhibit non-distinct or low value features. Non-disinct clusters are therefore labelled as **Minor** types, while the 3 distinct clusters are labelled as such:

1. **Observer** -> periphery users that are far away from the centre of the discourse and have little participation or involvement.
2. **Activator** -> active and bridge users that reply to tweets and have a high interaction and engagement with other users, whilst having a wide reach in spreading information across network.
3. **Leader** -> popular and public users who have many followers, are publicly listed and have a high global tweet count, that receive many replies to tweets and interact mainly with other users of similar status.

Below, I list the cluster sizes and usernames of members assigned to each distinct cluster group. I also re-plot the mean scores to include only the disctinct clusters.

K-Cluster Sizes:

Cluster	id
0	2027
1	408
2	534
3	424
4	2625
5	200
6	66
7	370
8	21

K-Cluster Types:

Cluster 2: Observer

List too large to display

Features: +ve eccentricity and -ve closeness

Cluster 6: Activator

---

	label
0	JayMzansi10
1	WolfDen12427788
2	Motheo2009
3	peezyjr
4	Custoza1
5	Light50995046

6	MEB40122141
7	Khadijahhosana
8	Africanist14
9	nellyngwenya7
10	aredi1234
11	PNMaster_
12	Ashante14527204
13	BFSF1212
14	Nkweeng1
15	BrandonKaule
16	Mamotlokoa02
17	Great_lioness
18	Mashabela_Paul
19	happyerics
20	Khumbuz98119386
21	Island_Tribe17
22	Fokazingzing
23	TopThestreets
24	TjRaisibe
25	CynthiaTshaka
26	tchuene
27	6b0594175411491
28	Indigenous_SA
29	PatrioticOneSA
30	better_SA_fan
31	Kelebile14
32	ThaboSchoeman
33	bushyza1
34	OfNkululeko
35	lyrlady
36	ladycovid
37	Noma_7006
38	sepedirock
39	siza_mhayise
40	Pheaga1
41	Shokwakhe16
42	nanoza23
43	Sipho_Nkosi
44	Galela15505225
45	Ingonyama1993
46	Sesie52422225
47	Majola74537213
48	Wabolethu02
49	Faye33938341
50	mashabelane
51	TP16176433
52	YenaAyaKwini
53	PutSAns_1st

54	Cyza
55	MthembuM2
56	21DaysLaw1
57	LimpopianMartin
58	Moemise_Kgabo
59	Aphanempolai
60	masotobe39
61	Dennis92070975
62	Tshima_Maminkie
63	mothol96901389
64	Maps_pj
65	MarioKhumalo

Features: +ve betweenness, closeness, outdegree, hierarchy, transitivity, reciprocity, #tweeted*

---

*Table list is sorted by feature values.

Cluster 8: Leader

---

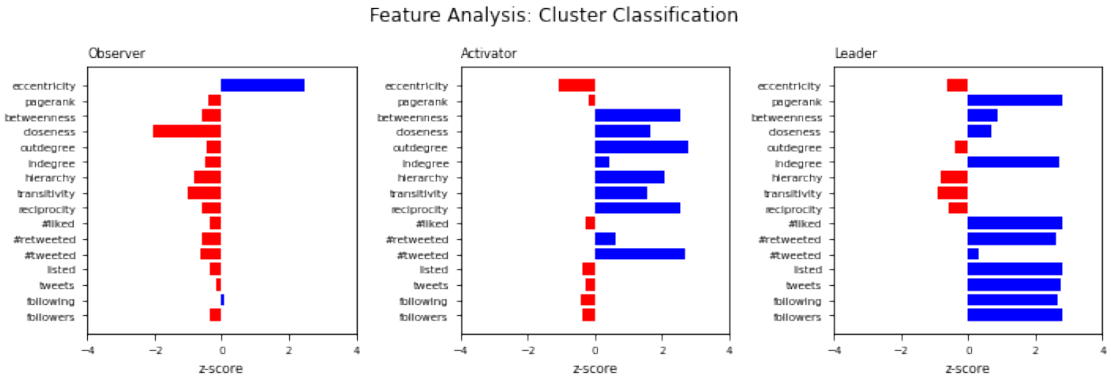
	label
0	MbuyiseniNdlozi
1	Lerato_Pillay
2	EFFSouthAfrica
3	CyrilRamaphosa
4	AdvoBarryRoux
5	HermanMashaba
6	MbalulaFikile
7	Julius_S_Malema
8	ZungulaVuyo
9	casspernyovest
10	MmusiMaimane
11	News24
12	Abramjee
13	KimKardashian
14	akreana_
15	davido
16	elonmusk
17	TiAmoNtombonina
18	RudyGiuliani
19	Cristiano
20	wizkidayo

Features: +ve pagerank, indegree, #liked, #retweeted, listed, tweets, following, followers

---

*Table list is sorted by feature values.

3.1 Mean Score Plots (excl. Minor)



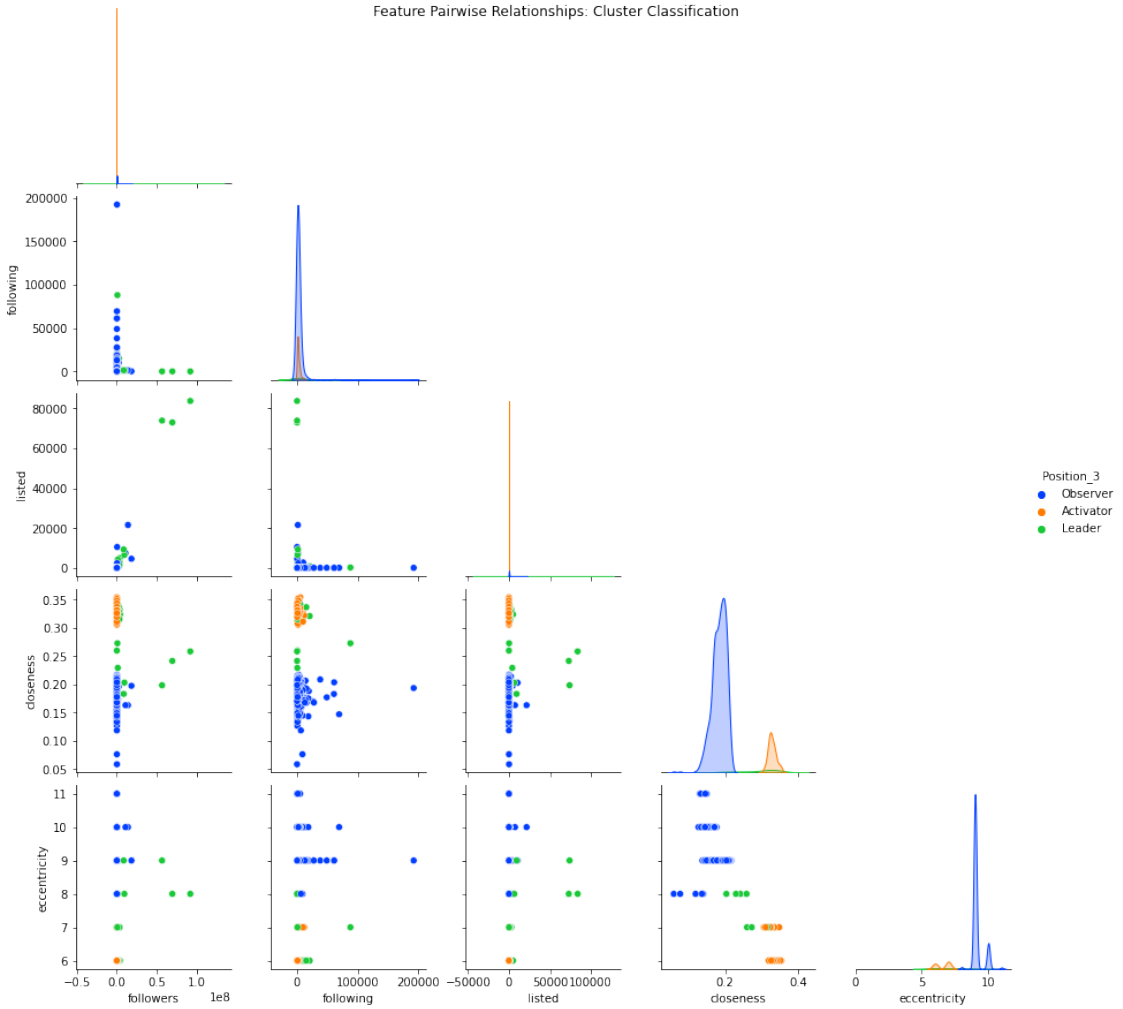
---

4 EVALUATION

Using pythons seaborn library, I plot the pairwise relationships of detected clusters to assess the fit of nodes within a cluster and the relationships between nodes and their features.

To interpret the results, lets say we want to consider the relationship between a node's distance from other nodes in the network and the probability of being added to a public list. From the plot, this can be found where the *closeness* feature on the y-axis meets the *listed* feature on the x-axis. Here, we fail to find a significant distinction in being publicly listed for nodes in the different clusters, except maybe for a few stray nodes in the *Leader* cluster. It is also evident that *Activator* nodes are positioned close to conversations in the discourse. Similary, when viewing the relationship between *indegree* and *pagerank*, not much seperation between node attributes exists except for a few *Leader* type nodes with exceptionally high *indegree* values. The *#retweeted* and *#liked* relationship appears to record high values for *Leader* types, showing a positve relationship between their discourse related tweets being retweeted and being liked. *Activator* clusters are significantly distinct across *outdegree*, *transitivity*, *reciprocity*, and *hierarchy* features.

4.1 Pairwise Relationship Plots




Feature Pairwise Relationships: Cluster Classification
Position_3
Observer
Activator
Leader
pagerank
indegree
#tweeted
#retweeted
#liked

5 SUMMARY

In this section, I use feature based techniques to conduct analysis on the positions and roles that define users in the #PutSAFirst twitter reply network. Below is a summary of notable findings.

I identify 3 distinct clusters that fairly categorise the roles one could expect from a discourse network on a social media platform like Twitter and classify them as such:

1. **Observer** -> periphery users that are far away from the centre of the discourse and have little participation or involvement.
2. **Activator** -> active and bridge users that reply to tweets and have a high interaction and engagement with other users, whilst having a wide reach in spreading information across network.
3. **Leader** -> popular and public users who have many followers, are publicly listed and have a high global tweet count, that receive many replies to tweets and interact mainly with other users of similar

status.

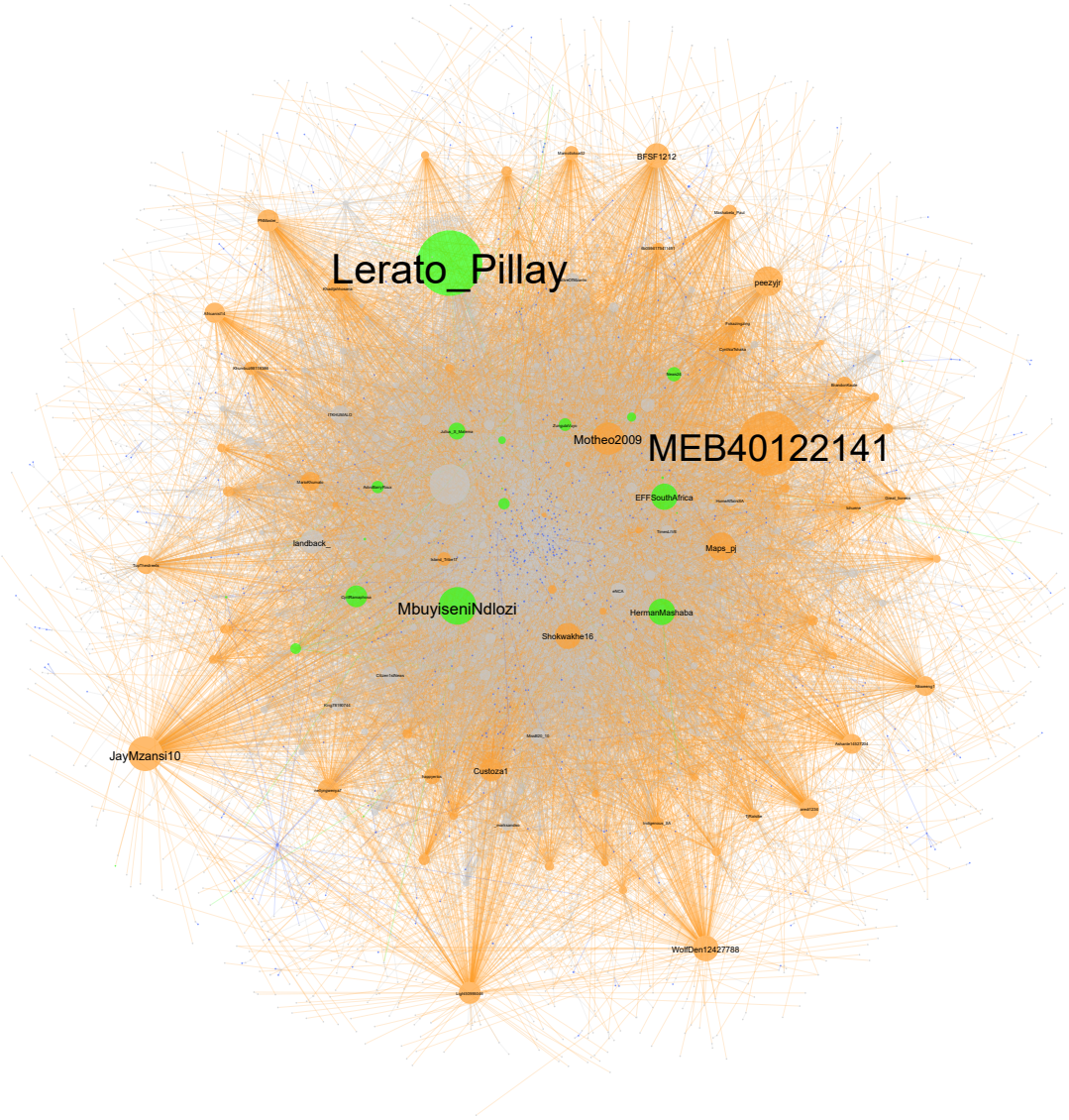
Unlike the 5 distinct roles detected in the other online discourses, this network only produces 3 distinct roles. This is possibly explained by its higher clustering property, and therefore a higher incident of redundant ties. It might be that there is a large duplication in the roles prominent users play in this network and therefore very little separation is detected between activities. Another possibility is that the feature based technique used in this analysis are somewhat limited. Thus it might benefit to compare it to outcomes from other network clustering methods, like **blockmodelling**.

The presence of outliers can also significantly affect outcomes. In the observed network, the effect of high status individuals and celebrities on clustering is very evident in the *Leader* role. For example, just by looking, users in this role can be split into two different types. One type represents the (locally) popular users who are important within the discourse. They display high measures of *pagerank*, *indegree*, *#liked*, and *#retweeted*. Prominent users who have managed to achieve a large influence within the discourse, for example, are **Lerato_Pillay**, **MbuyiseniNdlozi**, **HermanMashaba**, **EFFSouthAfrica**, and **CyrilRamaphosa**. Conversely, the other type are (globally) popular users who are not actively contributing to the discussion but happen to be public figures. The latter account for large values in *listed*, *tweets*, *following*, and *followers* features. A commonly used ploy or device to bring awareness to a certain movement or increase visibility in online social media is to tag or reply to (unrelated) tweets sent by public figures with a message related to the cause. Hence, the addition of users **KimKardashian**, **elonmusk**, **RudyGiuliani**, and **Cristiano** in this network. One option to deal with these types of outlier effects and achieve more consistent and accurate results is to remove them from the data completely and assess the difference in clustering outcomes. This is beyond the scope of the current analysis.

In the *Activator* role, I find the user **JayMzansi10** leading in *outdegree*, *hierarchy*, *transitivity*, *reciprocity* features. Other highly active accounts are **WolfDen12427788**, **Motheo2009**, **peezyjr**, and **Custoza1**. These are all accounts that are strongly in support of the movement and whose strongly worded tweets also include the shaming of dissenters and exposition of illegal immigrant "sympathisers".

In the next section, I simulate multiple diffusion processes using the discourse network to determine whether certain positions in highly clustered communities promote or hinder the spread of information and ideas.

PutSAFirst Positions:


BFSF1212
Lerato_Pillay
peezyit
Motheo2009
MEB40122141
EFFSouthAfrica
landback_
Maps_pj
MbuyiseniNdlozi
HermanMashaba
Shokwakhe16
JayMzansi10
Custoza1
WolfDen12427788