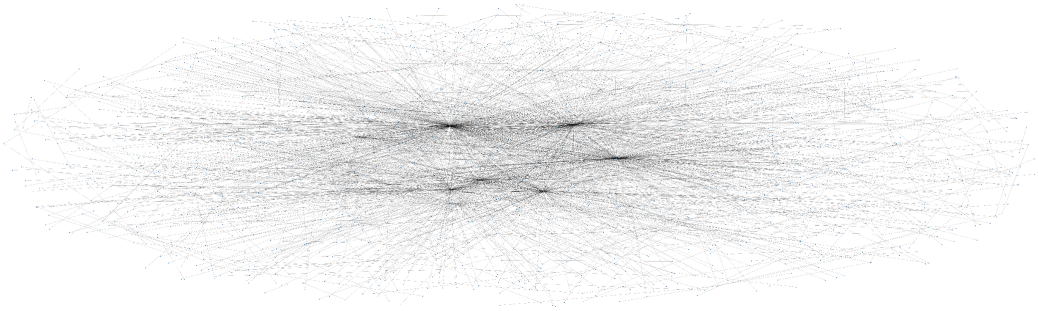
Hessel Eekhof - s3398641

Kylian de Rooij - s4549503

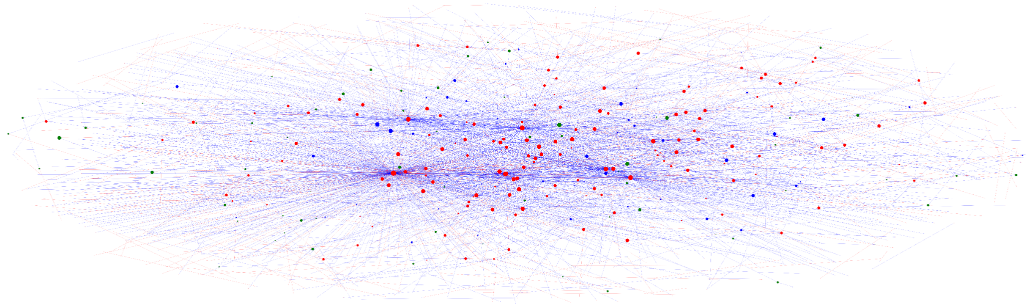
We chose to focus our attention on the network of tweets containing the word ‘Halsema’. Femke Halsema is a well-known Dutch politician from the GroenLinks party, and is currently the mayor of Amsterdam. Because of this she is often in the center of controversy. Twitter is no exception to this.

The expectation is that the network would be a ‘polarized crowd’. The reason for this is that political issues or figures are usually looked at from different points of view. Hence, the network of people consists of multiple groups not necessarily tightly connected to one another.



This is the network without any extra features. It is hard to make out the nodes and the edges because the network is quite big. To avoid having to visualize an extremely large network, we decided to only include tweets which originate from one user and which mention or reply to another user. Tweets which exist on their own were not taken into account.

From this visualization certain patterns can already be identified. It can be seen that a lot of edges convert to a few central points. In this network, there are six of them. Three of them seem to be the most important, whereas three other nodes seem less so. The rest of the participants in the network seem to be clustered around these central entities, as if they radiate some kind of ‘gravity’. However, this visualization is lacking in additional information. The following is more valuable in this regard.



The colors of the edges in this visualization provide information as to the kind of relationship that exists between two nodes. If the edge is blue, that means that there exist more tweets which mentioned the other node than tweets which replied to the other. Red edges indicate that the opposite is the case. Green edges, although rare, signal that the amount of mentions and the amount of replies are equal. The more tweets exist between two nodes, the thicker the edge is.

As can be seen, most of the edges are blue. This means that mentions are more common than direct replies. The Twitter users in this particular network seem to mention one another frequently, but direct conversations in the form of replies seem to be more rare.

The colors of the nodes indicate their proportion of indegree and outdegree. Red nodes have higher outdegree (they send more tweets), whereas blue nodes have higher indegree (they get replied to or mentioned more often). With green nodes, similar to edges, the proportion is equal. The sizes of the nodes depend on both their betweenness (how many paths cross through them) and their closeness (how far away they are from the other nodes).

Most nodes cannot be seen in this visualization. This is because their betweenness and closeness are so small that they are not visible. In a way, this provides the viewer with more information. ‘Irrelevant’ tweets are left out to highlight the ‘relevant’ ones more.

There are many relevant nodes. Among them are the nodes at the central points, discussed earlier. They are all red, indicating that they are very active in the sphere of Twitter. They send a lot of messages, and receive less. However, there are much more big nodes than just the ‘obvious’ ones. They seem to concentrate around the central points, which makes sense considering many paths cross through both them and the major nodes. Most of these nodes are red as well, but some of them are blue.

From these visualizations, it can be seen that the network is more a ‘tight crowd’ than it is a ‘polarized crowd’. There are no clearly distinguishable subgroups with little to no connection to one another. On the contrary: the network seems to be very much connected, clustered around a number of important nodes.

That is not to say, however, that there is no controversy involved. The people in the network might be replying to – or mentioning one another, but that does not necessarily mean that they are in agreement. They might voice their disagreement or criticize one another. Thus, this network indicates that people with differing political opinions might not always be trapped in a so-called ‘echo chamber’. It is also possible that they interact with one another even though they disagree. One example of this is the following tekst: “Wat een onzin. Volgens Halsema is demonstreren een grondrecht. Zie BLM verboden demo.” It is clear that this user is in disagreement with the user he or she is replying to.

Another conclusion which might be drawn is that the opposite is the case. People in this network might retweet one another extensively, or show support for one another by replying. Most people in the network might be more or less in agreement with one another. To support this conclusion, one might have to investigate the large number of retweets that occur in the network.

Other measures that we can investigate are the various centrality measures. An interesting measure to look at is the linguistic status measure. Pennebaker (2013) talks about this in his book *The Secret Life of Pronouns*. He explains how our use of pronouns and other function words tell us something about our personality.

We implemented this in our model by designing a linguistic status measure. This measure was calculated by dividing the amount of time a person used ‘we words’ by the total number of pronouns (‘I words’, ‘you words’ and ‘we words’). This leaves us with a value between 0 and 1, 1 being most powerful and 0 being least powerful. If a person does not use any pronouns, the score is set to 0.5, a neutral score.

We also decided to investigate the eigenvector centrality measure. This is a measure of centrality of a node, based on the centrality of its neighbors. It may prove as a useful measure, as nodes that would otherwise not have a high centrality, may now have a higher centrality due to them being connected to important neighbors.

In our network, ‘arjanbeets1967’ had the highest linguistic status with a measure of 1, which is the most powerful. The eigenvector centrality was 0.00115, which is not particularly high, but is substantially greater than the eigenvector of the person with the lowest linguistic status, which will be discussed later. The betweenness was 0.00458, which is also quite low. Finally, the closeness was 0.213, which is close to the average. Overall, nothing out of the ordinary, and thus does not fit Pennebaker’s predictions.

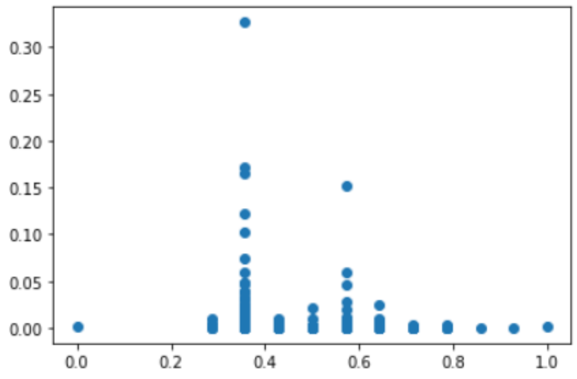
The lowest linguistic status we found in the network was ‘snijderpeter’, who had a score of 0.0, the lowest possible score. The eigenvector centrality was 0.000198, which is almost 10 times smaller than the highest score. Betweenness was 0.0 and closeness 0.192. Taking these scores into account, this fits Pennebaker’s predictions slightly better, but not quite.

The network centrality measures are very good when analyzing a network like this, as they provide relevant information about the connections that people have. However, they don’t tell us anything about the way the person talks, which is what Pennebaker’s linguistic status measure does. It tells us how powerful an individual is, based on the way they use pronouns. A weakness that Pennebaker’s measure has is that it doesn’t provide very accurate information with little text. In this data for example, there are only tweets, which can only be a maximum of 280 characters. As a result of this, some tweets do not contain any pronouns, thus making the measure practically useless.

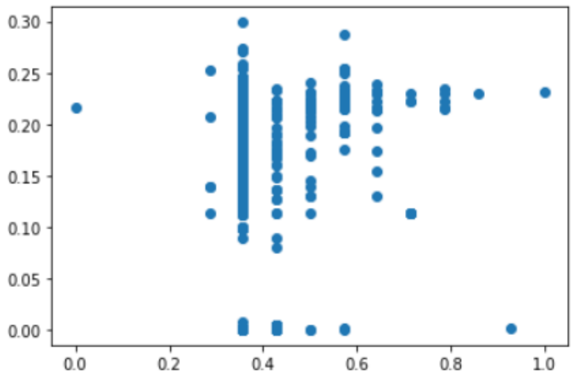
A better way to utilize Pennebaker's linguistic status measure is on bigger texts. An example that Pennebaker gives is a speech from John F. Kennedy, where there are more pronouns, providing a more accurate measure. On top of that, this speech is not part of a network, meaning that the centrality measures cannot be used, making Pennebaker's measure better in this case.

Overall, there is not necessarily a better measure, it depends on when it has to be used. If one is analyzing a network and its connections, the centrality measures will prove more useful, but if one is evaluating a larger text, not part of a network, Pennebaker's measure will be more useful.

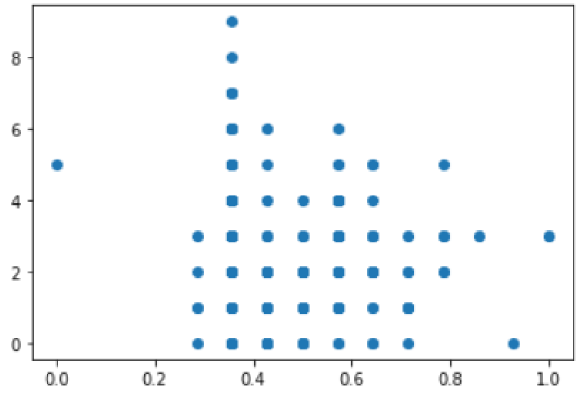
The graphs below show the relationship between the linguistic status measure and various centrality measures. As seen in the graphs, there are no real correlations between linguistic status and the centrality measures, which could indicate that Pennebaker's measure is indeed best used in other cases, namely where one is analyzing someone's power in a single text, rather than in a network.



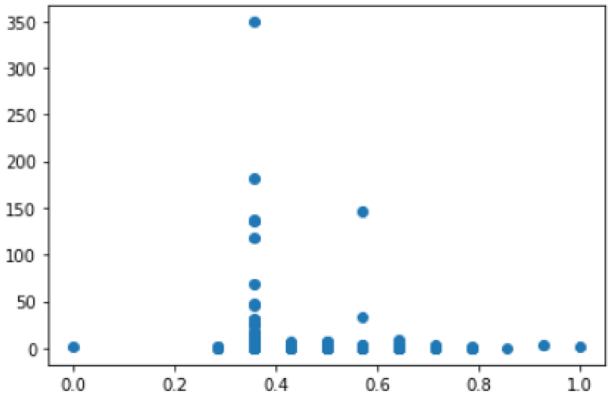
Ling status vs betweenness



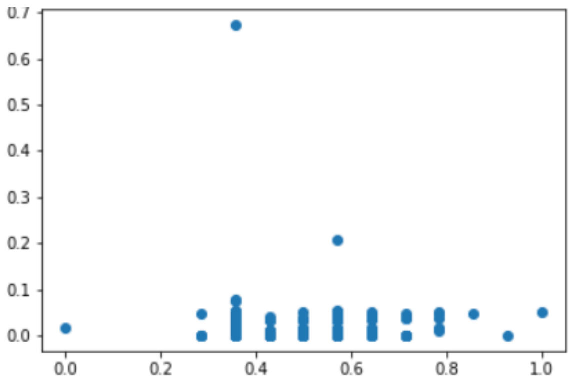
Ling status vs closeness



Ling status vs indegree



Ling status vs outdegree



Ling status vs eigenvector