SMAII CA2

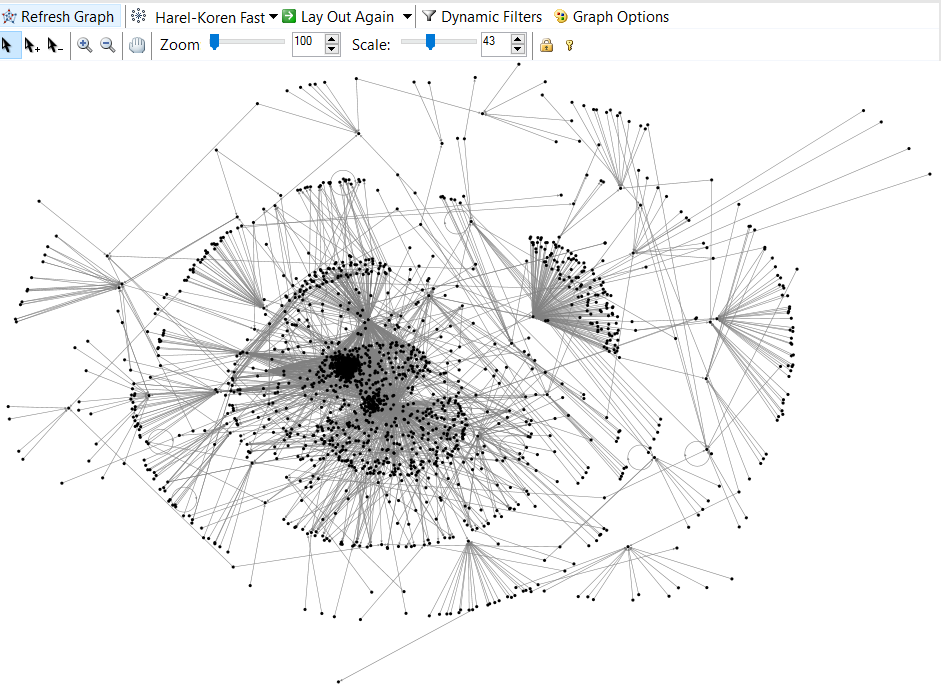
Adam Kielthy

X00111393

Network Visualisation

The network I have chosen is based on the phrase “person of the year”, and consists of twitter users and their tweets. My intention is to identify who is talking about this topic, whether there is conversation going on or just simple tweeting, and to identify groups of people talking about this, for example celebrities, politicians and news agencies. I expect Donald Trump to be the centre of the network, with a high volume of political chatter, and also news stories. I am also interested in the size of networks based on political opinion, and I will use this to guess whether general opinion is positive or negative.

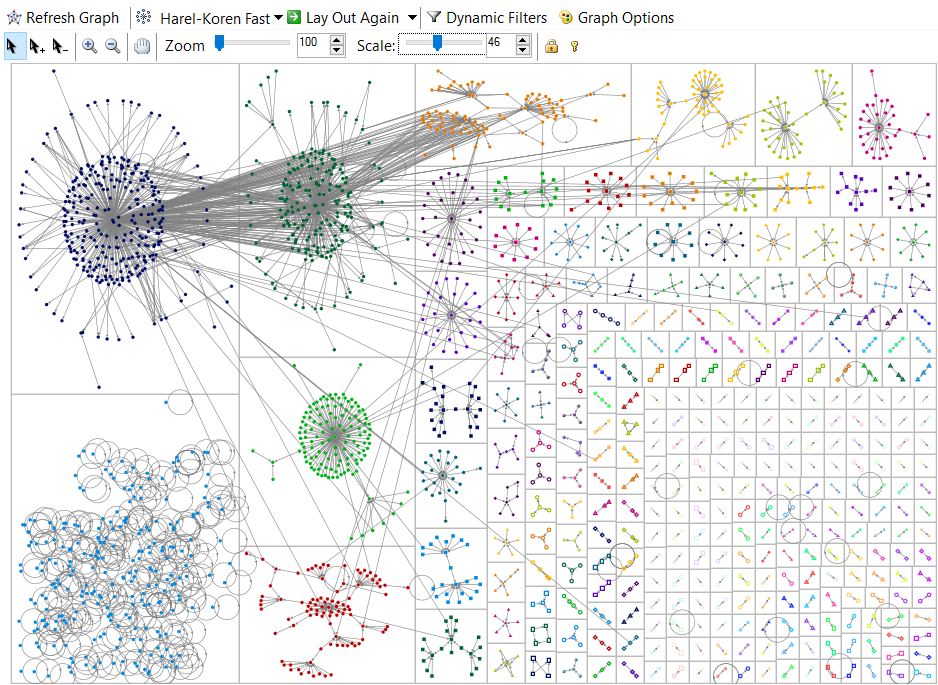
After trying out a variety of layout options, It was clear that “Harel-Koren Fast Multiscale” was the most appropriate for my network.



As shown above, there are some clear clusters and smaller networks centralised around a single person. The above graph also clearly shows that there is a high amount of unidirectional communication.

To make this graph clearer I created groups based on clusters, and sorted users into boxes based on their group. I decided to use the default colour settings, allowing NodeXL to colour users based solely on the group they belong to.

This produced the image below:



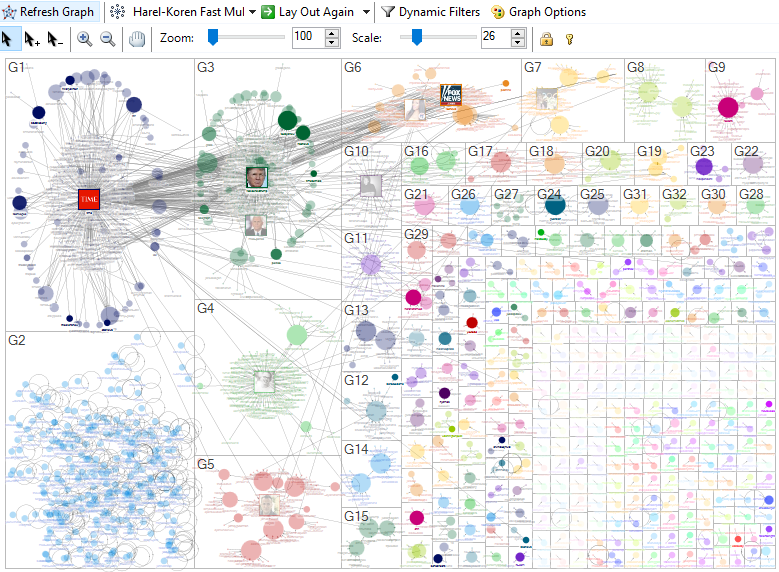
Next I added some visuals to make the network easier to analyse.

Vertex Size was generated using the users in-degree, with a range of 10 – 250. This makes the range larger and shows clearly who is the most tweeted user and therefore the centre of the network.  
Opacity was generated using the users followers count, but this allowed outliers such as YouTube, CNN and the BBC which have millions of followers to take control of the graph and make it difficult to retrieve any information.  
To make the graph more legible I then set the outliers to be ignored when generating vertex opacity. Although this did remove the outliers completely, it allowed for a more complete image overall.  
Also, minimum opacity was set to 30, so that the most important people in the network were obvious and highlighted, but regular users were also still visible.

Groups 1 to 32 are labelled to make referencing them easier, all other groups are too small to be labelled without the label covering the entire group.

Users with an in-degree higher than 30 have their shape set to the users twitter profile picture, this allows us to easily see popular users such as Time, CNN, Fox etc.  
A Dynamic filter was also used to remove users with less than 300 followers, this was done to reduce clutter and make the graph more informative through clarity. I avoided going above 300 users as this would remove a high number of twitter users who are just regular users and have no inflated following through fame.

With the above additions, the graph below was produced.



After the above modifications, we can now use a nodes size (In-degree) and opacity(Followers) to determine their importance in the network. More often than not, the largest node in a group is the focus of the cluster, G8 is a notable example to this where two users are at the centre of the network.

In-degree was chosen over out-degree and degree because there is a very low amount of reciprocal tweets, and one person rarely tweeted enough times to distinguish them from the rest of the sample.  
This low reciprocal tweeting statistic also means that no cliques can be seen in this network.

In the graph, Donald trump and Time are the two most important users according to the measures applied to it. This is not at all surprising considering the search term revolves around these users. Some other highlighter users who received a lot of tweets on the subject are anti-trump journalist Shaun King, Vice-President Mike Pence, and various news outlets and journalists.  
These users were deemed to be of high importance because they were mentioned in over 30 tweets, and are in the highest bracket of followers in the network.   
It is also worth noting that Donald trump had the highest betweenness centrality in the network, with no other user coming close, this solidifies the idea that he is central to the entire network.

G1 & G3 are the biggest clusters in the network, they are both of similar size with 290 and 260 users respectively. They are also quite similar in the contents.   
G1 is mostly news outlets and journalists reacting to Time, with some scattered regular users. While G3 contains Donald Trump & Mike Pence, who are surrounded by some news outlets, but mostly twitter accounts whose sole purpose is to support Donald. The top hashtag in this group was “#maga”, Trumps campaign hashtag. This suggests the tweets in this network are generally supportive of his campaign.  
Although these networks are the most populated and contain the central figures, they are the least dense groups in the entire network, with a density of 0.004 and 0 respectively.

G6 is another group which consists of news outlets, and then high volumes of regular twitter users tweeting them. This is likely news outlets referencing Time and Donald trump in the tweet, and then the smaller nodes are users reacting to the news story.

G10-G30 seem to be clusters of regular twitter users, all focusing on one single user who has a particularly high follower counts. The lack of replies implies that the users are referencing the user, but not talking to them, or that they are mentioning the user to inform them, again, rather than actually conversing.

A lot of two and three node groups exist, with one node of high importance and others tweeting to that user. The higher importance users are mostly tv shows and artists, including jimmyfalon and dj\_khaled.

This network has a large amount of singletons, G2 has 200 singletons, making it the third largest group. This could be due to the search term being big news, and people are including it in their tweets to express their opinion on the current events.

The non-reciprocal nature of this network is strange, it’s unusual for a network of 2000 users to have less than 10 bi-directional connections. This may be caused by how topical the search term was, that there was such a high volume of tweets the NodeXL import limit was reached and the network contained tweets that were too new to have replies.  
If the same search query was executed now I would expect the graph to more resemble a typical network, with less users but more communication between each user.  
Due to this, the occurrence of hubs, bridges and cliques