Barack Obama Retweets Network

Mateusz Zaremba

December 16, 2019

1 Methods

1.1 Network size metrics

The data was downloaded from WEBSITE and prepared using R; on the ‘Network Repository’ website we can find already calculated network data statistics, like number of nodes, number of edges, average degree etc.; these are also already calculated at the repository’s website (Rossi and Ahmed, 2015). I created a network graph in python programming language using networkx library and compared the obtained statistics with the originals. Turns out they are exactly the same:

Table 1: The results are the same - we can be pretty confident we have created our network graph in python correctly.

	Calculated Statistics	Website Statistics
Number of nodes	9631	9.6K
Number of edges	9775	9.8K
Average degree	2.0299	2

1.2 Network Structure Metrics

Our graph in undirected so it does not contain in- or out-degrees. Instead, we found the highest degree node - it had an id of 2506 and degree of 7655; this was likely Barack Obama’s twitter account since it is the most connected.

1.3 Network Density

The network’s density is equal to 0.000210789557302036, which is close to zero, meaning our graph is close to being ‘fully disconnected’.

1.4 Shortest Path Between Two Nodes

We will find the shortest path between top two nodes with the highest degree and between the nodes with the highest and the lowest degree.

1.4.1 Top Two

The top two nodes with the highest degree have ids consequently: 2506 (Barack Obama) and 9302. Shortest path between these two nodes is: [‘2506’, ‘8474’, ‘9302’]; the length of this path is 2, which means they are not connected directly.

1.4.2 Highest and Lowest Degree

As already mentioned, the id of the node with the highest degree is: 2506 and the id of the node with the lowest degree is: 2709; the shortest path between these two nodes is: [‘2506’, ‘2709’] and its length is 1, meaning the nodes are connected directly.

1.5 Identifying Network Communities

We managed to identify 138 community groups:

```
Counter({1: 7414, 6: 611, 7: 158, 12: 114, 4: 106, 2: 102, 50: 84, 5: 69,
88: 66, 61: 61, 127: 53, 35: 49, 19: 39, 25: 38, 17: 27, 111: 27, 77: 26,
23: 25, 112: 25, 0: 22, 49: 22, 26: 20, 54: 15, 37: 14, 102: 14, 22: 13,
41: 12, 48: 12, 55: 12, 9: 11, 15: 10, 32: 10, 75: 10, 97: 10, 34: 9, 38:
9, 82: 9, 85: 9, 13: 8, 3: 7, 29: 7, 84: 7, 105: 7, 108: 7, 68: 6, 78: 6,
83: 6, 117: 6, 86: 5, 110: 5, 8: 4, 27: 4, 31: 4, 62: 4, 67: 4, 79: 4, 93:
4, 100: 4, 129: 4, 10: 3, 20: 3, 36: 3, 40: 3, 47: 3, 66: 3, 70: 3, 80: 3,
81: 3, 87: 3, 90: 3, 95: 3, 104: 3, 107: 3, 114: 3, 125: 3, 133: 3, 134:
3, 139: 3, 11: 2, 14: 2, 16: 2, 18: 2, 21: 2, 24: 2, 28: 2, 30: 2, 33: 2,
39: 2, 42: 2, 43: 2, 44: 2, 45: 2, 46: 2, 51: 2, 52: 2, 53: 2, 56: 2, 57:
2, 58: 2, 59: 2, 60: 2, 63: 2, 64: 2, 65: 2, 69: 2, 71: 2, 72: 2, 73: 2,
74: 2, 76: 2, 89: 2, 91: 2, 92: 2, 94: 2, 96: 2, 98: 2, 99: 2, 101: 2, 103:
2, 106: 2, 109: 2, 113: 2, 115: 2, 116: 2, 118: 2, 119: 2, 120: 2, 121:
2, 122: 2, 123: 2, 124: 2, 126: 2, 128: 2, 130: 2, 131: 2, 132: 2, 135: 2,
136: 2, 137: 2, 138: 2})
```

It was calculated that the sizes of the communities range from 2 to 7414. We can again assume that the biggest community is likely to be centred around Barack Obama.

1.6 Network Structure Connectivity

Investigation reveals that the Barack Obama network is fully connected, and it has no subcomponents; this is not surprising because the edges are retweets, nodes are twitter users and the network consists only of users who retweeted Barack Obamas posts.

1.7 Network Hubs/Brokers

Betweenness and closeness centrality was successfully calculated (Although, it took almost 40 minutes to compute) and sorted from the highest to the lowest score for top 20 results; but a `(PowerIterationFailedConvergence(...), 'power iteration failed to converge` `within 100 iterations')` error kept occurring when calculating node eigenvector centrality; replacing `nx.eigenvector_centrality` with `nx.eigenvector_centrality_numpy` solved the issue (Stack Overflow, 2019) and after the fix, the computation was almost instantaneous.

Barack Obama's node (id = 2506) had the highest score in closeness, betweenness and eigenvector centrality, which means:

- The highest closeness centrality score - it is the farthest away from all other nodes in the network or – it takes the most time to spread information sequentially from it to other nodes (Sciencedirect.com, 2019).
- The highest betweenness centrality score – it has the highest number of distinct paths that strictly contain it in between (Sci.unich.it, 2019).
- The highest eigenvector centrality score – it is the most influential node in the network.

2 Results

It was also assumed that the biggest community of 7414 would be Barack Obama's first-degree neighbours. This cannot be really well seen on the graph without any colouring or sizing:

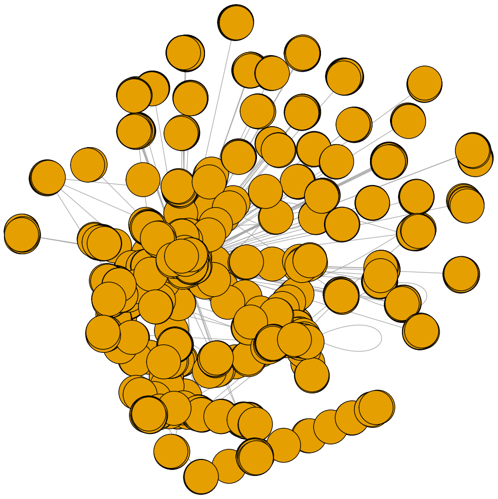

Figure 1: Graph without colouring or sizing

But we can immediately see Barack Obama's node when we distinguish the node's degree using sizing and colouring:

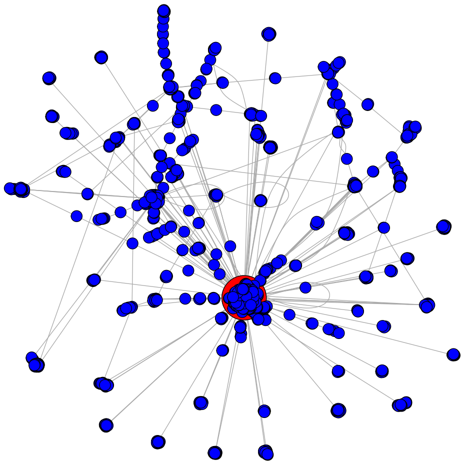

Figure 2: Graph with node sizing and colouring

The graph becomes even more readable when we use colouring to represent the node's score of closeness and the node's sizing to represent its measure of betweenness.

The same graph was plotted using `Large Graph Layout` function:

Interesting results were achieved using Gephi (Gephi.org, 2019), were it is clearly visible that there is one central node in the network (Barack Obama).

3 Conclusion

We managed to successfully identify the node which is very likely to be Barack Obama's twitter account – the nodes id was 2506; find the shortest path between top two nodes with the highest degrees as well as the node with the highest and the lowest degree; identify 138 communities within the network and their range.

The size of the network proved it quite difficult to work with; calculation of the node betweenness and closeness centrality took approx. 40 minutes. Plotting did not take as much time - it took only a few moments; a big improvement was noticed when the nodes labels were not being displayed. Nonetheless, an attempt was made to speed up the process

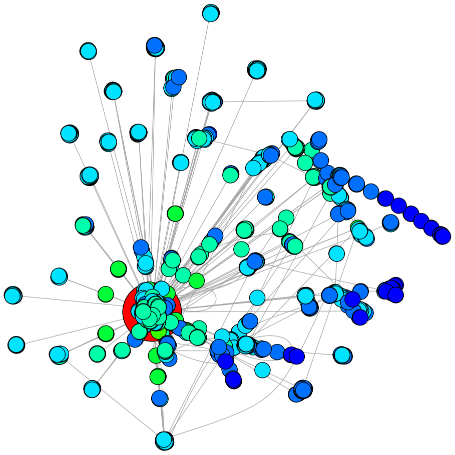

Figure 3: Colour - closeness. Size - betweenness

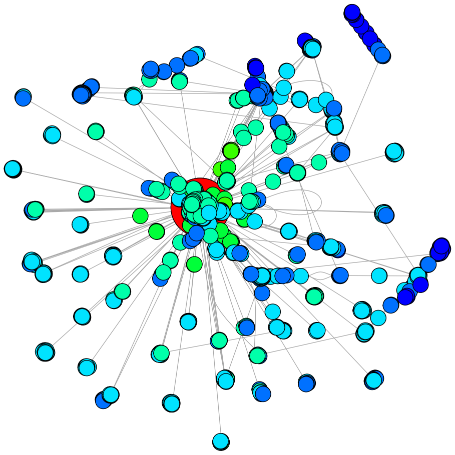

Figure 4: Plotted using Large Graph Layout

using - `with_lgl(...)` (Rdrr.io, 2019) function for `Large Graph Layout` but no improvement in rendering speed was noticed apart from a slightly better visual representation of communities.

One of the things that could be done in the future, when analysing even bigger networks, to improve calculation and render times, could be a usage of a GPU for network analysis. (Mathworks.com, 2019); e.g. Geforce GTS 250 has 450 cores working in parallel versus 4 on the CPU (Kajan and Slačka, 2019).

Overall, the data was well described and of a good quality; the original analysis provided a lot of good insight (Look table 1) and the size of the data was appropriate for a comprehensive analysis.