Analysis and Visualisation for Twitter Network Data

Ayça Avcı (s4505972)
Robert Monden (s3851117)
Stella Tsoutsouri (s3904210)

Social Network Analysis course
December 15, 2022

1 Network statistics - Metrics

A variety of datasets were provided to us in order to choose which to utilise for analysis. We chose the PHEME dataset which contains tweet posts from multiple sources. The PHEME project focuses on rumour detection and veracity classification. The content of the dataset intrigued us, that's why we formed the decision to process them for our report. The tweets are categorised either as a rumour or no-rumour among different topics. The topics were categorised based on the hashtag used on the platform, in total eight. Those were: Charlie Hebdo, Ottawa shooting, Ferguson, Sydney Siege, Prince Toronto, German wings crash, Ebola Essien, Putin Missing.

Additionally, information about each user was included also in the dataset. We could examine the user's mentions and amount of tweets posted. The connections among users were also available where we could analyse who follows whom. Lastly, the reactions to tweets were also included. The data provided showed tweets in English and German language. We analyse and report the English tweets due to the team's knowledge in this language.

We examined 43732 nodes and 92607 edges. The resulting number of connected components was 30853. Moreover, the diameter of the strongest component was 29 while the density of the strongest component was approximately 0.0004. We have created multiple subgraphs from the network in order to analyse it and showcase our results. In Figure 1 you may examine a directed graph showing the vertices with different sizes. Each size is dependent on their degree. In total we observe one big vertex and four others in a bit smaller size. The numbers that are visible in each vertex indicate the user id, the edges indicate who is following whom.

Since, the account with number '537454328' demonstrates the biggest vertex, we could speculate that this account refers to a source that is considered credible by the users. This could be an institution, news medium or an individual with specific expertise. The other accounts with a high degree were: '250179380', '561587400', '81688768' and '427623553'. Furthermore, the vertices '250179380' and '561587400' are important for the network since their removal would exclude multiple nodes.

The importance of vertices for the network is calculated by the betweenness centrality measure. This measure indicates that if specific vertices are removed, access to the network would be restricted for some other vertices. We performed calculations in order to analyse the betweenness centrality in our network. As aforementioned, we created subgraphs in order to examine the network efficiently. We have included in this report a variety of different structured subgraphs. You may observe the results and their explanation in Figure 2, Figure 3, Figure 4, Figure 5, Figure 6 and Figure 7.

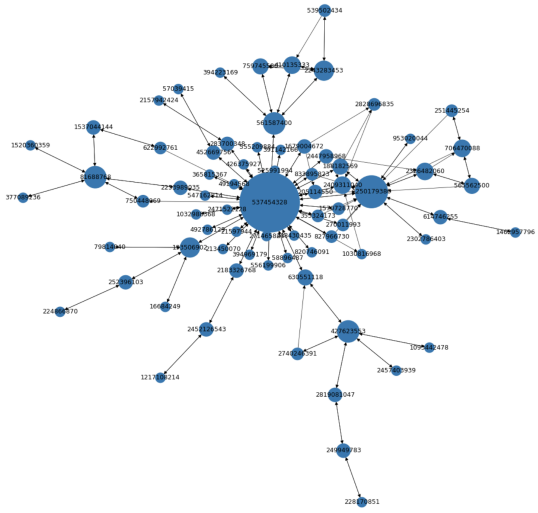

Figure 1: Graph showing nodes with their size dependent on their degree

Lastly, we performed calculations in order to analyse the closeness centrality in our network. This measure indicates which vertices are in a good position to spread information to the network. As aforementioned, we created subgraphs in order to examine the network efficiently. We have included in this report a variety of different structured subgraphs. You may observe the results and their explanation in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13.

2 Communities

As a follow up step we analysed the network in order to identify the groups that get formed. We identified the cliques and presented them by generating graphs. In total 50 graphs showcasing the different groups were created. Below we define in detail their specifications:

- Two graphs showed connections among 12 vertices
- 31 graphs showed connections among 11 vertices
- 17 graphs showed connection among 10 vertices

The connections of the nodes were based on users who tweet misinformation (rumours) and users who don't tweet misinformation (no-rumours). Furthermore, who from the users tweet and who just react to the tweet posts was used as a parameter. You may observe the graphs with 12, 11 and 10 vertices in the Figure 14, 15 and 16 respectively.

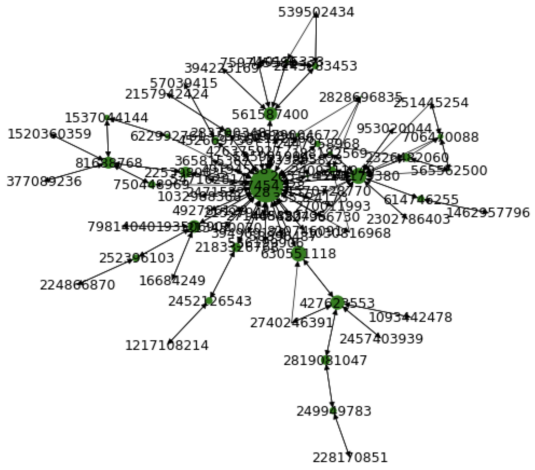

Figure 2: Subgraph showing betweenness centrality. In this graph we observe one vertex in the middle that shows the highest betweenness centrality and four other vertices in smaller size that indicate a significant importance for the network.

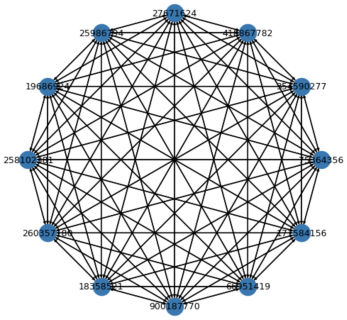

Figure 14: Graph with 12 grouped vertices

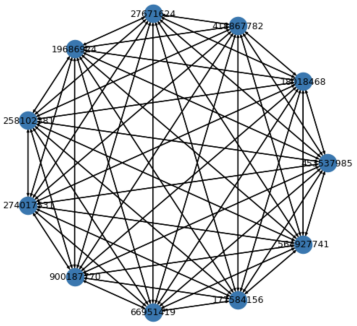

Figure 15: Graph with 11 grouped vertices

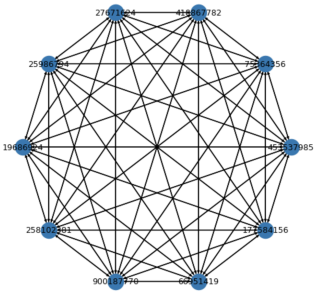

Figure 16: Graph with 10 grouped vertices

Moreover, we examined how communities get formed while implementing the Girvan-Newman algorithm. This algorithm provides the ability to detect communities while removing edges from the graph. The edges with the highest betweenness centrality are removed gradually. The graph produced shows two different communities as seen in Figure 17. The two communities are distinguished by the colour green and blue. Both communities have a big vertex in the middle and multiple others ones with different sizes spread through the graph.

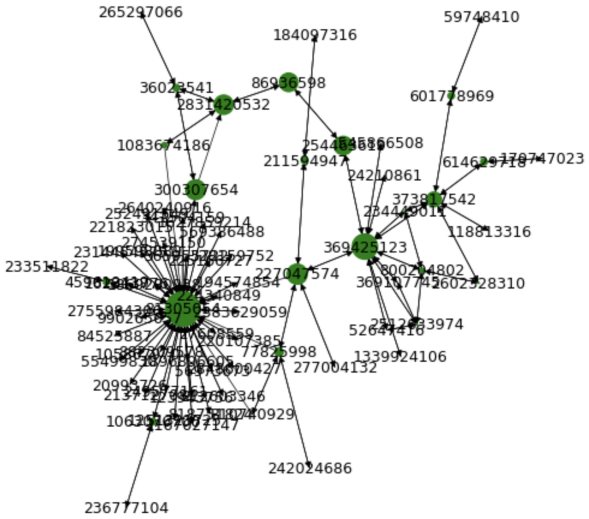

Figure 3: Subgraph showing betweenness centrality. In this graph we observe one vertex that shows the highest betweenness centrality and seven other vertices in smaller size that indicate a significant importance for the network.

3 HITS Implementation

The hub and authority scores were calculated using the HITS algorithm. Specially, we used the `networkx` implementation. Although the raw values included the hub and authority scores for all nodes, for the visualisation only those belonging to the largest connected component were included.

Figure visualises the different hub scores of the nodes in the connected component where user 'Goalies-NeverSay' is most influential. As we would expect, the nodes that are more centrally located have a higher hub score than the leaf nodes. In a similar fashion, Figure shows the corresponding authority scores.

At first glance the figures appear identical, however subtle differences can be seen near the node representing the user with id '1250179380', and near the leaf nodes right of it.

The five nodes with the higher hub scores are:

- Node 537454328 (hub score of 0.1076)
- Node 1250179380 (hub score of 0.0293)
- Node 188182569 (hub score of 0.0282)
- Node 827966730 (hub score of 0.0270)
- Node 355324173 (hub score of 0.0270)

Similarly, the five nodes with the highest authority scores are:

- Node 537454328 (authority score of 0.1306)

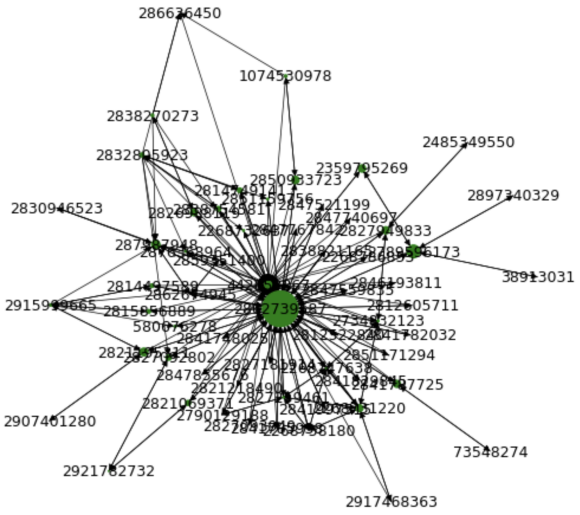

Figure 4: Subgraph showing betweenness centrality. In this graph we observe one vertex in the middle that shows the highest betweenness centrality and one other vertex next to it in smaller size that indicates a significant importance for the network.

- Node 1250179380 (authority score of 0.0619)
- Node 2409311040 (authority score of 0.0287)
- Node 188182569 (authority score of 0.0233)
- Node 205114550 (authority score of 0.0222)

4 Longitudinal Analysis

The evolution of the number of reactions over time was analysed for five of the source tweets. In Figure 20, we see an initial spike in reactions within the first thirty minutes for four of the five source tweets.

For the same four source tweets another spike can be seen around twelve hours after the source tweet was posted.

A similar phenomenon cannot be observed for the source tweet concerning Putin's disappearance. There is a slight increase in the number of replies around the thirty minute mark, however the number of replies seems to grow linearly.

One possible explanation for the discrepancy can be the nature of the topic: perhaps the topic is simply much less divisive than the other ones. However, this hypothesis would not explain the much larger spikes of replies to the German Wings source tweet.

5 Discussion and Conclusion

In this report we analysed a twitter network by implementing multiple social network analysis methods. We focused on the network metrics, its communities, the HITS Implementation as well as the longitu-

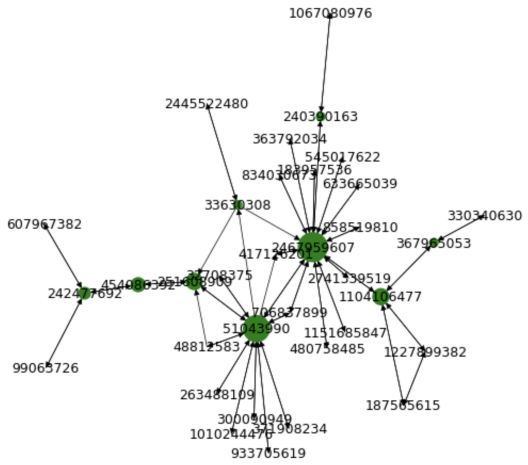

Figure 5: Subgraph showing betweenness centrality. In this graph we observe that vertices show relatively almost the same betweenness centrality. The scale of the network is also smaller.

dinal analysis. We identified the total of nodes and edges. We calculated the components produced for users of the network focusing on the parameter of who follows whom. We examined the closeness centrality in order to observe which vertices are the more popular ones on the network. Also, we examined the betweenness centrality in order to observe vertices are significantly important for the network. The positioning of these nodes provides the opportunity for other nodes being part of the network. In this specific case it would affect the users and with whom they interact.

We managed to create cliques based on users that post misinformation and those who don't. Also, by using who will create posts and who will just react to them contributed to creating additional cliques. Using the Girvan-Newman algorithm revealed two specific communities as part of the network.

By implementing the HITS algorithm we were able to calculate the hub and authority scores for the network. Also, the longitudinal analysis provided information in regards to reactions over time. The timespan used was from 1 minute until 5 days.

Due to the complexity of social media networks subgraphs were created to analyse specific sections of it. For the implementation of this analysis NetworkX framework, Python coding language and matplotlib library were used.

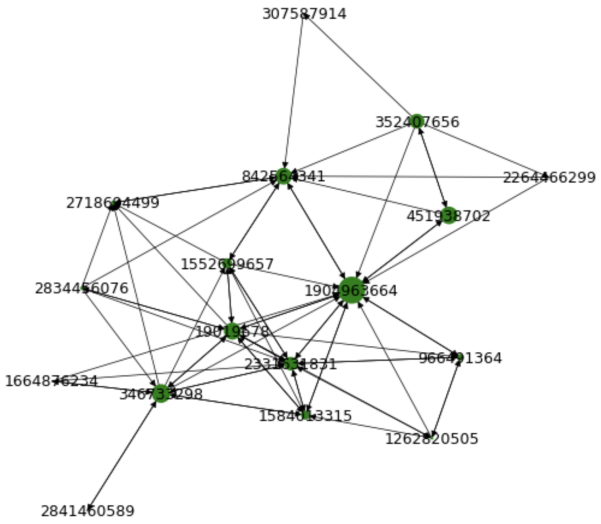

Figure 6: Subgraph showing betweenness centrality. In this graph we observe that vertices show relatively almost the same betweenness centrality. The scale of the network is also smaller.

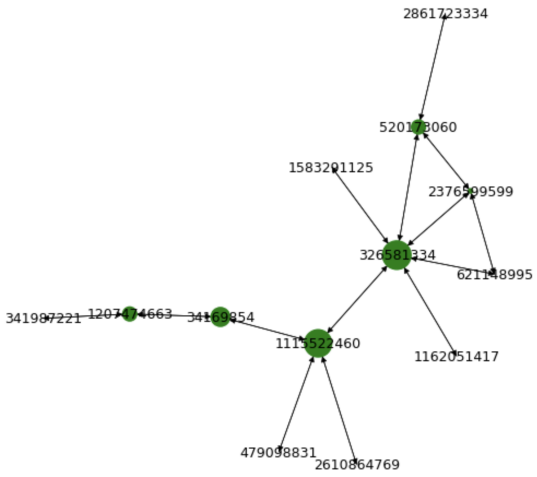

Figure 7: Subgraph showing betweenness centrality. In this graph we observe that vertices show relatively almost the same betweenness centrality. The scale of the network is very small thus the vertices have a significant importance.

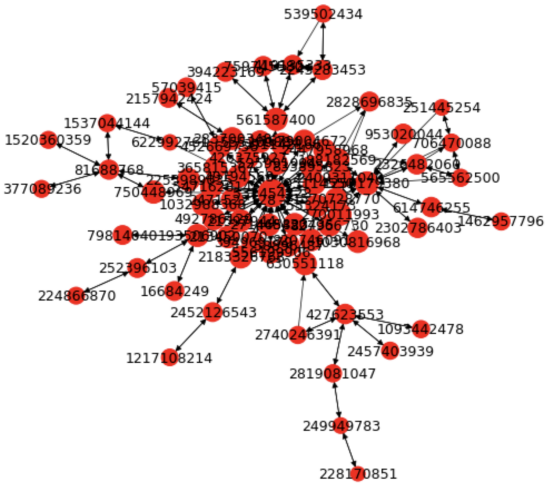

Figure 8: Subgraph showing closeness centrality. In this graph we observe one vertex in the middle showing the highest closeness centrality. The vertices around it show a smaller closeness centrality but still are considered significant for the network in order to spread information in the furthest vertices.

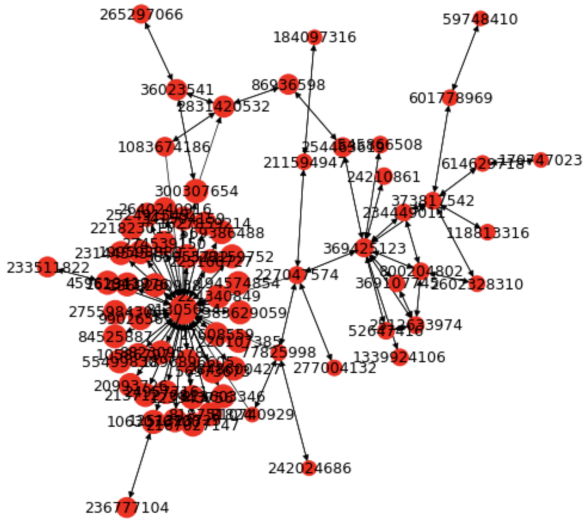

Figure 9: Subgraph showing closeness centrality. In this graph we observe one vertex in the middle showing the highest closeness centrality. All the rest of the vertices around it show a smaller closeness centrality. Therefore, there is no significant difference in their importance on the network.

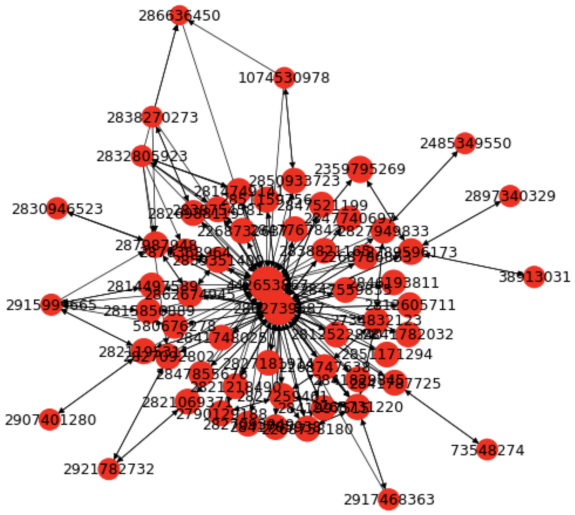

Figure 10: Subgraph showing closeness centrality. In this graph we observe two vertices in the middle showing the highest closeness centrality. All the rest of the vertices around it show a smaller closeness centrality. Therefore, there is no significant difference in their importance on the network.

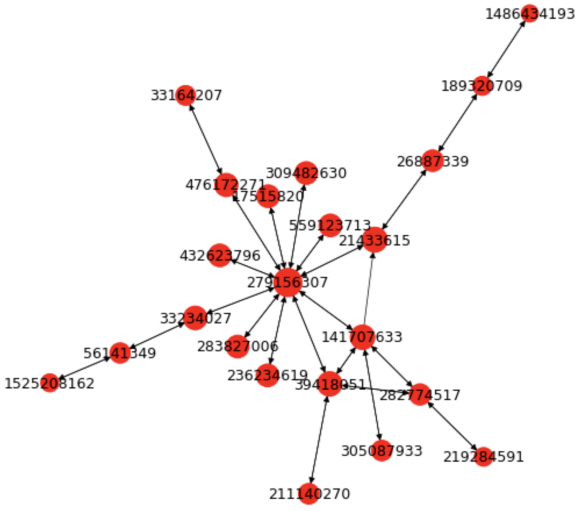

Figure 11: Subgraph showing closeness centrality. In this graph we observe one vertex in the middle showing the highest closeness centrality. All the rest of the vertices around it show a smaller closeness centrality. Therefore, there is no significant difference in their importance on the network. The scale of the network is also smaller.

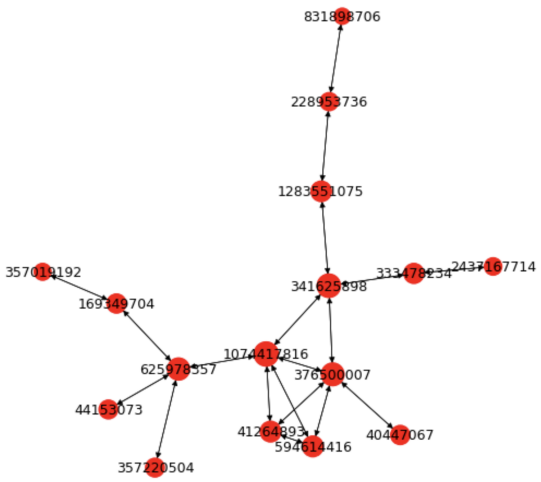

Figure 12: Subgraph showing closeness centrality. In this graph we observe four vertices in the middle showing the highest closeness centrality. All the rest of the vertices around it show a smaller closeness centrality. Therefore, there is no significant difference in their importance on the network. The scale of the network is also smaller.

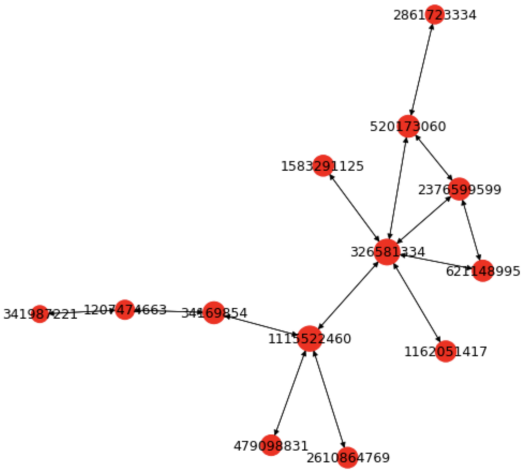

Figure 13: Subgraph showing closeness centrality. In this graph we observe two vertices in the middle showing the highest closeness centrality. All the rest of the vertices around it show a smaller closeness centrality. Therefore, there is no significant difference in their importance on the network. The scale of the network is also significantly smaller.

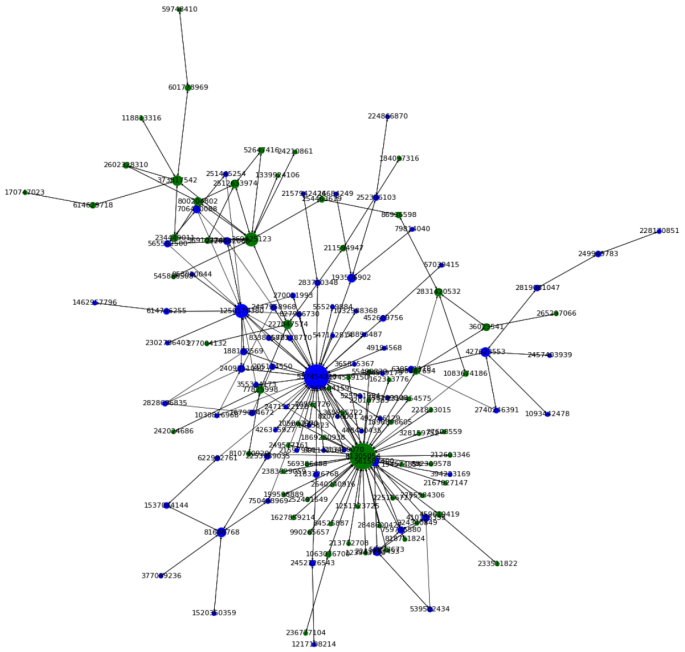

Figure 17: Graph showing the result of using the partitioning algorithm Girvan-Newman

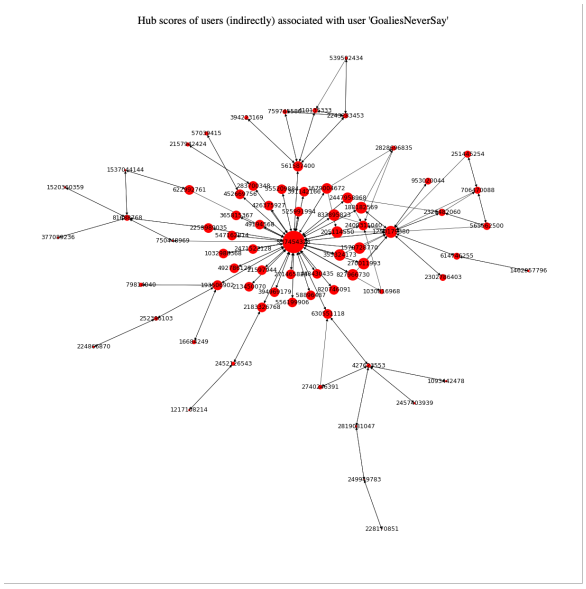

Figure 18: Hub scores of users associated with user 'GoaliesNeverSay'. The nodes are scaled by a factor 10^4.

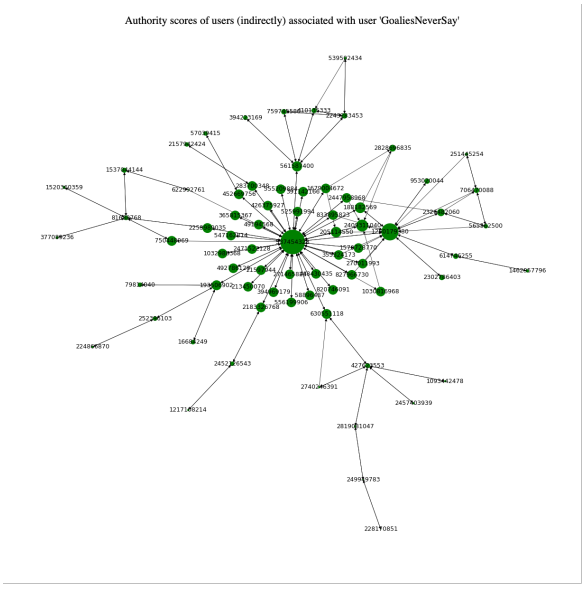

Figure 19: Authority scores of users associated with user 'GoaliesNeverSay'. The nodes are scaled by a factor 10^4.

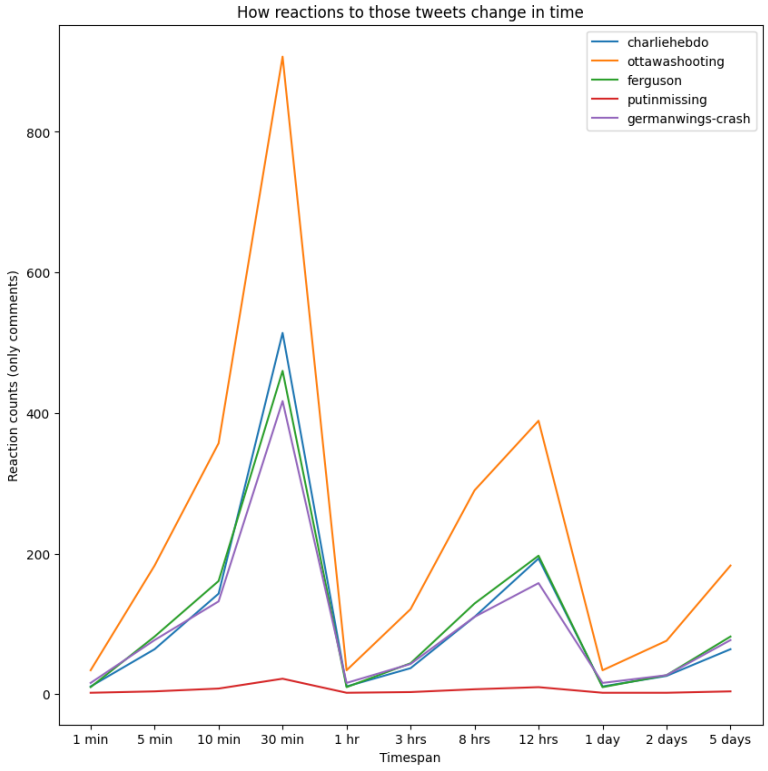

Figure 20: The number of reactions to each tweet over time.