Abeeha Zawar (SDP)—06225

Abeer Khan (CS)—05419

Maria Hunaid Samiwala (SDP)—05686

12/11/2022

Homework 2

Work distribution among group members:

Abeer Khan:

- The script for data extraction, data cleaning, and network construction
- The construction and visualization of the networks

Abeeha Zawar:

- Visualization and comparative analysis of the networks

Maria Samiwala:

- Visualization and comparative analysis of the networks

Network Construction:

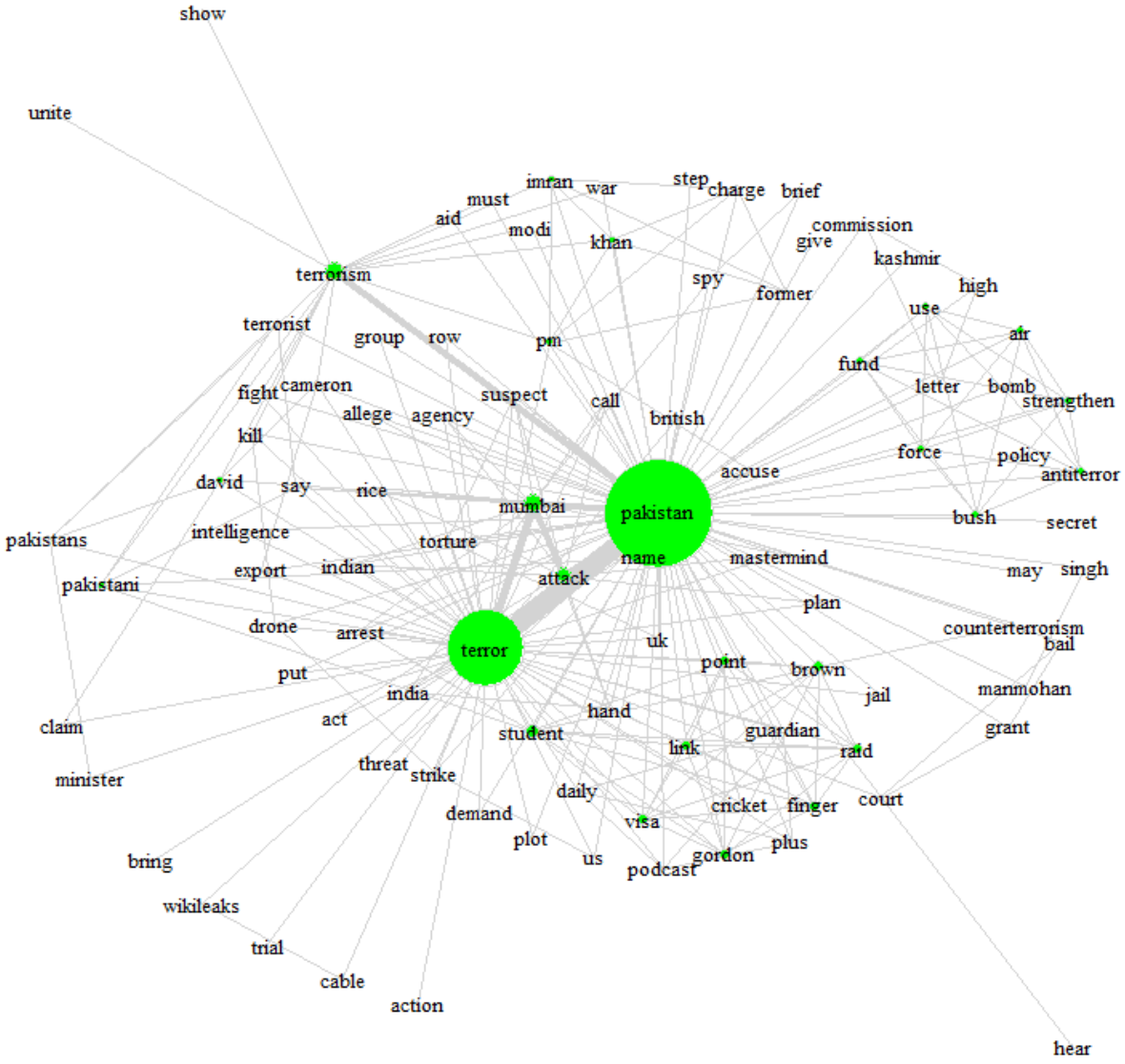
For the construction, we used hunspell stemmer and word stemmer and a stem completer to complete some of the stemmed words such as "sou" to "south" or "emerg" to "emerge". The network was stemmed twice to ensure proper data cleaning. The words were tokenized, stop words (where, when, who, what, how, etc.) and extra words (like the, can, get, got, etc.) were removed, as well as special characters were removed. We divided the code for the network in separate sections i.e. for flood and terror for our ease in visualization.

For the floods network (Figure 2), articles with the words ‘Pakistan’, ‘Floods’ (capital F) & ‘floods’ (small f) in the title are retrieved. We were only able to recover 86 articles before we exceeded the API call limit. Similarly for the terror network (Figure 1), articles with the words ‘Pakistan’, ‘Terror’ (capital T) & ‘terror’ (small t) in the title are retrieved. Networks without these terms in the article’s title were also generated for further analysis (Figure 3 & 4).

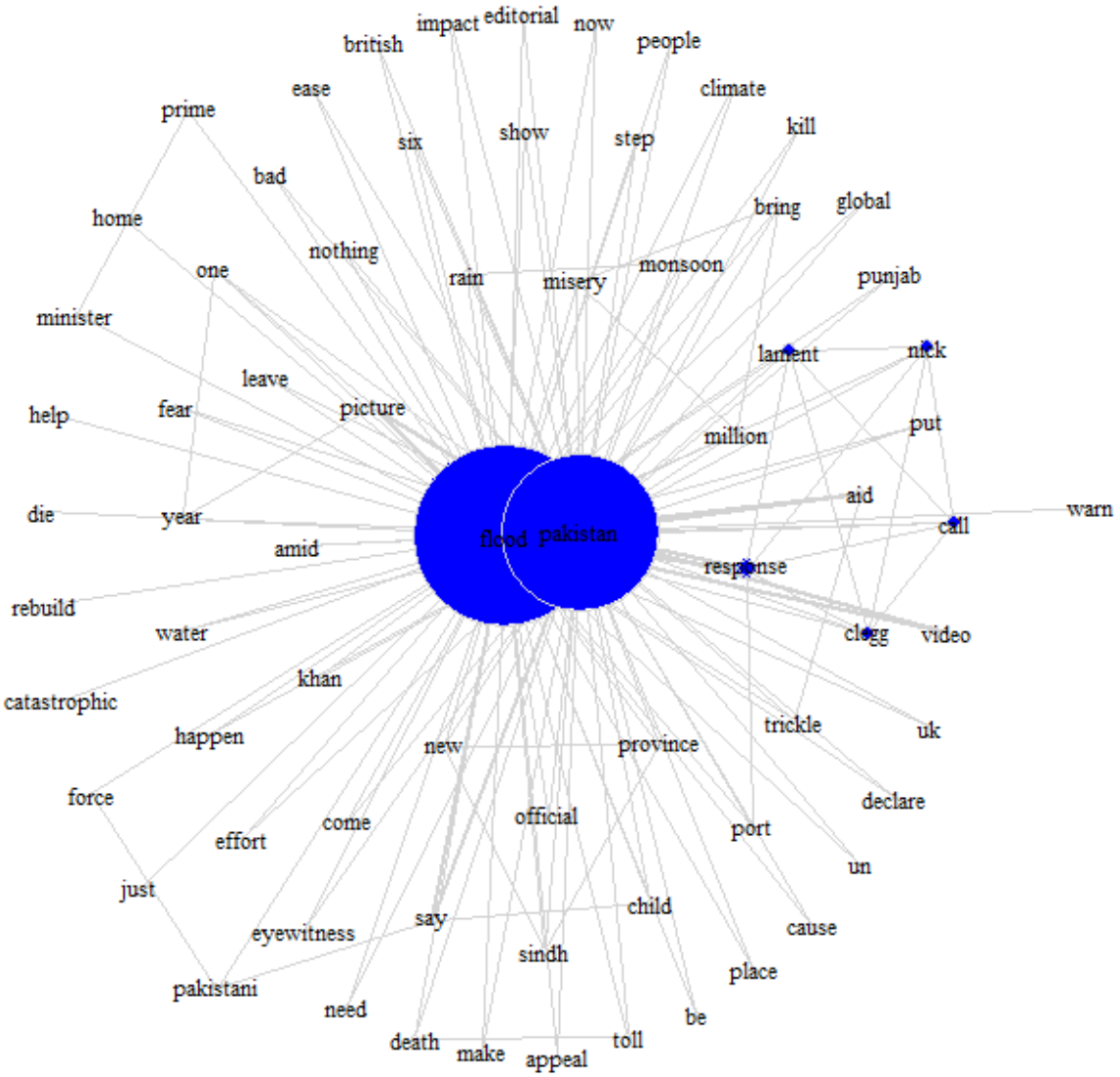
Additionally, the colour and size of vertex were used as visual elements for analysis. The higher the degree of connections the nodes have, the bigger in size they are, and the ones with above average degree have a different colour. Force directed layouts were applied to all the networks.

Network Visualization:

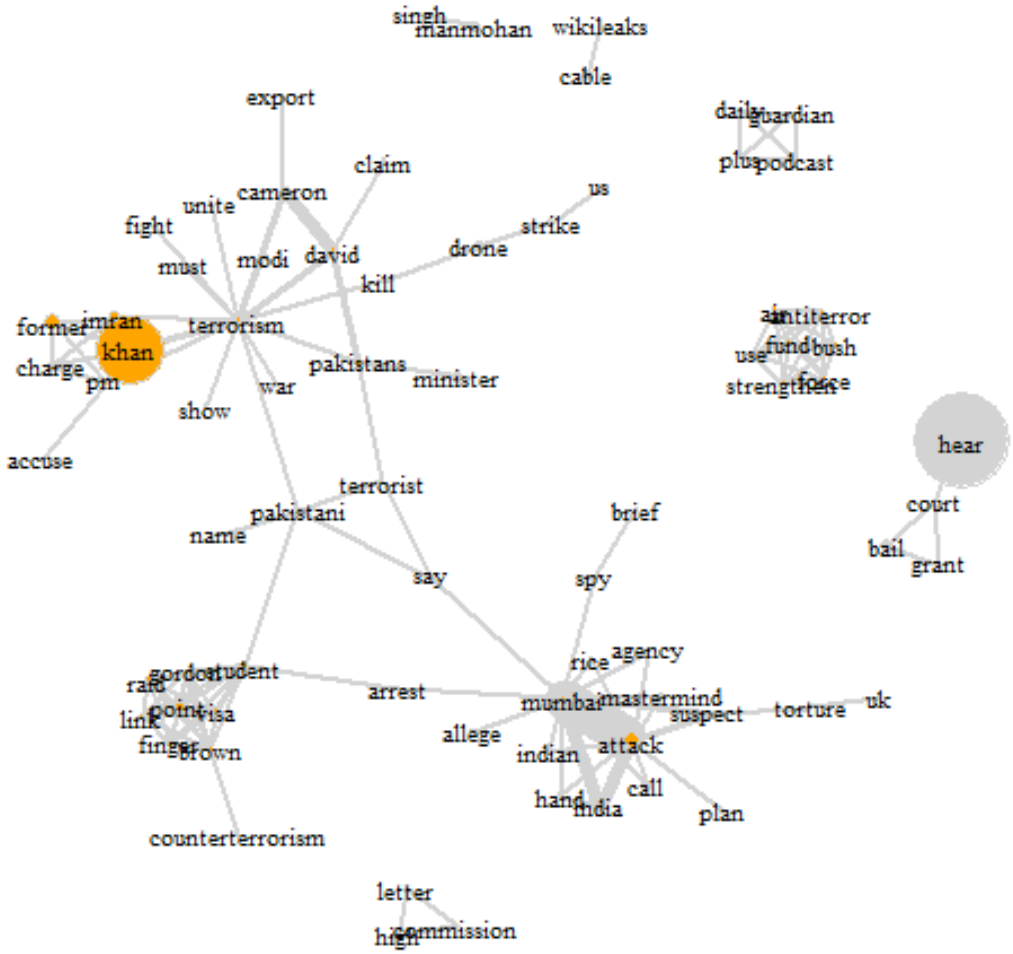
1. Network with ‘Pakistan’ and ‘Terror’ in Title (Figure 1)



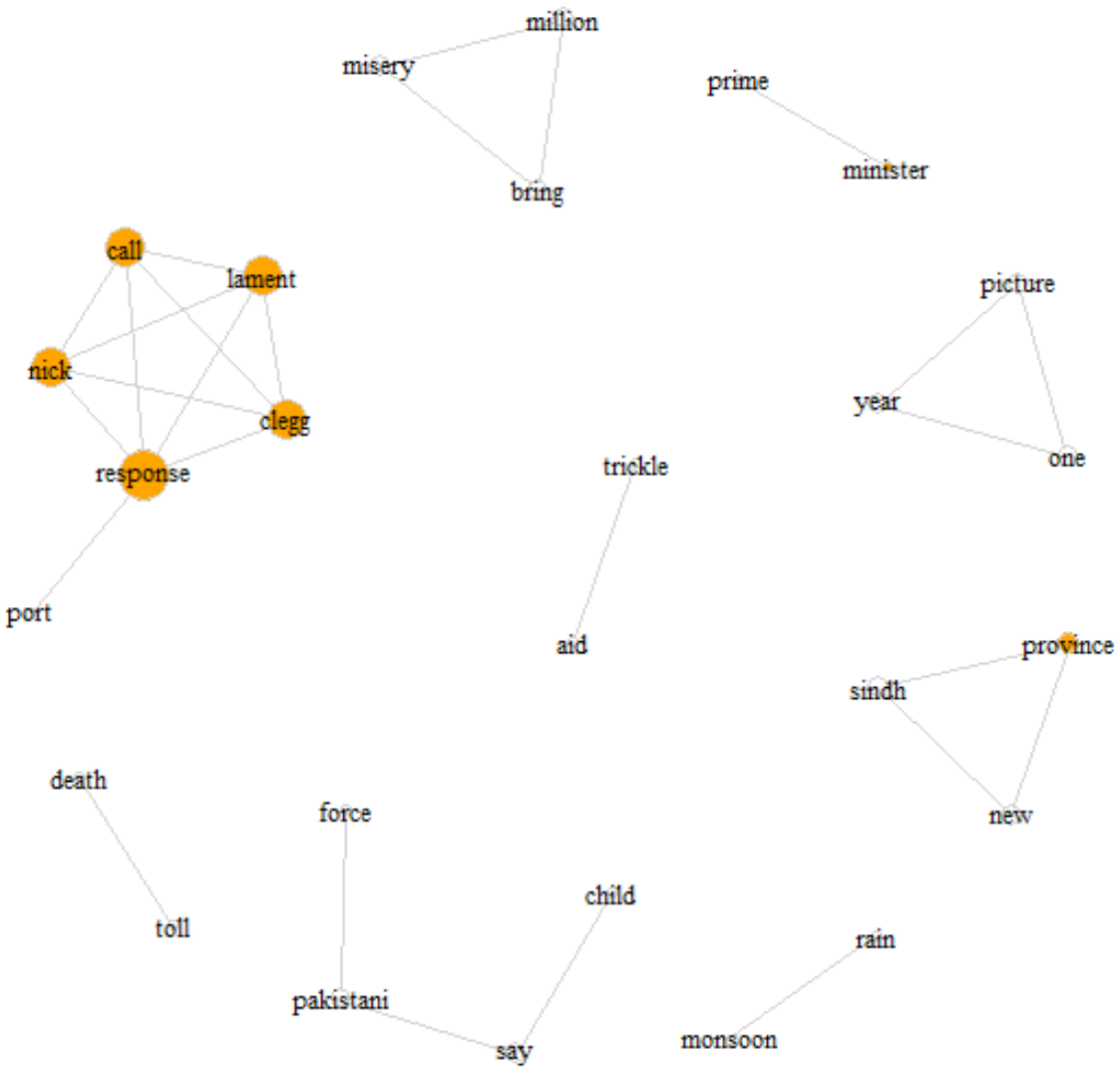
2. Network with ‘Pakistan’ and ‘Floods’ in Title (Figure 2)



3. Network without ‘Pakistan’ and ‘Terror’ in Title (Figure 3)



4. Network without ‘Pakistan’ and ‘Floods’ in Title (Figure 4)



Network Analysis:

The central nodes (with the highest degrees) around Figure 1 for the Pakistan with Terror network seem to be ‘attack’, ‘mumbai’, ‘student’, ‘terrorism’ (despite the double stemming), ‘raid’, ‘brown’, and ‘visa’. They all seem to indicate political situations—‘brown’ is interesting because of how terrorism in the West, since 9/11, terrorism is seen as a Middle East or Muslim world problem, where terrorists are characterized by brown people; ‘visa’ points to how difficult immigration for travel or work or personal reasons has become a major impediment for brown people since the War on Terror (carried out by U.S. President, George Bush–interesting to note ‘bush’ and ‘us’ are also present in the network). Mumbai seems to be another central node, when coupled with ‘india’/’indian’, ‘modi’, ‘manhoman’, & ‘singh’ (most probably, alluding to both the current and formerIndian Prime Ministers) points to India as a whole being a major theme in this network. This shows how central India is as an international player when it comes to questions surrounding Pakistan’s security or politics because of the region’s geo-political history. There are a lot of international players involved like UK, India, US, Mumbai, Kashmir in this network. There are also individuals involved like Imran Khan, Modi, Manmohan Singh, Bush, David Cameron, as well as the word British.

In Figure 2 for the Pakistan and Floods network, the words ‘nick’ and ‘clegg’ are part of the central nodes. Upon researching, might allude to the former deputy Prime Minister of the UK (and current President for Global Affairs at Meta), alluding to the political aspect of this network as well. In this, there is notably only two international players involved in this network, i.e. UK, and UN. Along with ‘british’, the presence of UK is in the terror network as well. This points to the UK being an overall theme in the Pakistani context as determined by its co-occurrence around both the terms. Other geographical players are more local, like Sindh and Punjab. The words ‘khan’ and ‘pm’ or ‘prime minister’ playing a common occurrence might point to the importance of Imran Khan in both news coverages.

The coverage, as seen in figure 2, does not seem to be as political as the terror network. Other central nodes in the flood network seem to be ‘response’, ‘call’, ‘lament’. They indicate the urgency and magnitude of the disaster, and the occasional soft messaging used to appeal to emotion and appeal for action (‘rebuild’, ‘aid’, ‘child’, ‘misery’, and ‘need’ also seem to support this argument).

Whereas, the language used in the terror network is powerful, impactful (alarming, concerning), and strong—words such as ‘threat’, ‘strike’, ‘plot’, ‘kill’, ‘fight’, ‘torture’, ‘force’, ‘spy’, ‘bomb’, ‘counterterrorism’, ‘torture’, and ‘accuse’. Similarly, most of the language being used in Figure 2 for the Pakistan with Floods network is comparable to the terror network, i.e. strong, impactful language—such as ‘die’, ‘death’, ‘toll’, ‘warn’, ‘force’, ‘fear’, ‘kill’, ‘misery’, ‘catastrophic’, ‘appeal’, ‘impact’, ‘lament’, and ‘declare’. The words in common are ‘force’ and ‘kill’, alluding to the difference in coverage of these networks, but at the same time, the words for both have active and impactful connotations.

When comparing both Figure 1 and 2, we can see that that network included nodes and connections that have nothing to with Terrorism or Floods in Pakistan. In figure 1, irrelevant words such as ‘cricket’, ‘rice’, collocations such as ‘india’ & ‘indian’, ‘terrorism’ & ‘terrorist’, ‘pakistans’ & ‘pakistani’; in figure 2, extra words such as ‘be’ and ‘just’, and collocations such as ‘pakistani’. The words are present despite the double stemming.

The edge density for the terror network is 0.05668934, whereas for the flood network it is 0.0614241. Owing to the small difference, this means both the networks are very sparse because the frequency of the cooccurrence is very low. Furthermore, the global clustering coefficient for the terror network is 0.1685866, whereas for the flood network it is almost half at 0.0828841. This means that the words in the terror network are twice as connected to each other as the words in the flood network. Additionally, for the terror network (figure 1), average path length is 2.082457, whereas for the flood network (figure 2), it is 1.942488. This means Figure 2 is a smaller network, where the words are closer to reach each other. However, since the difference is not too much, we can conclude that in both the networks, the nodes are exactly (Figure 1) or roughly (Figure 2) 2 hops away from reaching each other.

In Figure 1,we can see that the nodes are closely connected with no notable clusters present in the graph. Here we can see in our network that all the nodes are directly connected to the central nodes (Pakistan and Terror), but don’t have any triangles or undirected networks. Similarly, Figure 2 has nodes that are closely connected, and all nodes connected to the two central nodes (Pakistan and Flood). However, when we compare Figure 1 & 3, and Figure 2 & 4, a different story emerges. Figure 3 and 4 are networks without the central nodes, ‘Pakistan and Terror’ and ‘Pakistan and Floods’ respectively. We see ‘former’, ‘imran’, ‘khan’, and ‘attack’ are the most notable nodes in Figure 3 when we remove terms terror and pakistan from the analysis, confirming the large role Imran Khan has with the terror network news coverage. There are also some isolated clusters, but most of the network is connected by a few hubs. Whereas in Figure 4, there are a lot of isolated clusters, meaning that the network was very much reliant on the term and context of Figure 2. This more so proves the aforementioned difference in global clustering coefficients for Figures 1 and 2, seeing how disconnected clusters in Figure 4 are. The clusters are in their own cliques, and there is no bridge between different clusters.