Part B - defining and analyzing the network

Submitted by : Ofri Hefetz 209028067 & Shai Shani 206165318 & Gony Idan 315817601

Section A -Data editing and processing

Our data is based on transcripts of news broadcasts from the news network CNN. We focused on the news broadcast around major events in the US to examine the different shifts that the news takes during and after these events that shape and impact the nation. We will examine the main topics discussed in news broadcasts by analyzing the words spoken in each broadcast. Our data is based on Harvard Dataverse which scrapped the transcript on CNN broadcast available online between the years of 2000-2022.

We chose five major events between those years to examine and analyze:

1. **9/11 terror attack** - a devastating terrorist attack that occurred on September 11, 2001, resulting in the destruction of the World Trade Center towers in New York City and the loss of thousands of lives.
2. **Invading Iraq March 2003** - The United States initiated military action in Iraq in March 2003, leading to a protracted conflict and subsequent occupation that lasted for several years. Beginning march 19, 2003.
3. **Hurricane Katrina in August 23-31, 2005** - a catastrophic natural disaster, caused widespread devastation and loss of life, particularly in the city of New Orleans, Louisiana, and exposed significant flaws in emergency preparedness and response.
4. **Killing of Bin Laden May 2011** - Osama bin Laden, the founder and leader of the extremist group al-Qaeda, was killed in a covert operation by United States Special Forces in May 2011, bringing a significant milestone in the fight against terrorism.
5. **Bombing at Boston Marathon at April 15, 2013** - tragic terrorist attack that occurred in 2013, resulting in multiple casualties and injuries near the finish line of the renowned race.

For each of these events we extracted all the transcripts of news broadcasts in a period of a month, separated by each day. Since we are examining the effects over time, each graph will represent the context of all the news broadcasts per day. For each day and each transcript we removed all the stopwords, commonly used words, such as "a", "an", "the", and "in", due to their high frequency and low semantic significance. In addition we turned all the verbs and nouns into their basic form (root). This action helps reduce biases and makes the network less noisy. Finally we used word embedding to represent each word as a vector. We need to add which type we used.

Section B - All features will be taken into account during analysis

In order to examine the different phenomenons that may occur in our networks, for each event we will create graphs for 5 days before the event and 15 days after (the selection of the number of days is done arbitrarily). For example for 9/11 terror attacks we will examine the graphs created from news broadcasts between 6.9.2001 to 26.9.2001. In total we will have 20 graphs per event, and 100 graphs overall. In the following figures 1,2,3 we can see the distributions of the number of nodes and edges in all our graphs. The average number of nodes are 296,544 and the median number of nodes are 304,538. The average number of edges are 309,291 and the median number of edges are 315,020.

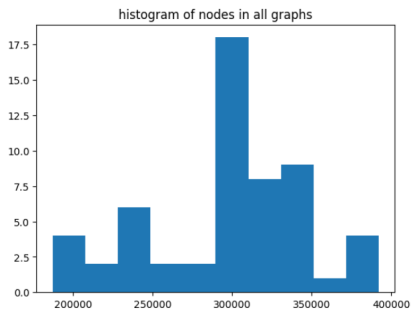

Figure 1 - Distribution of the number of vertices

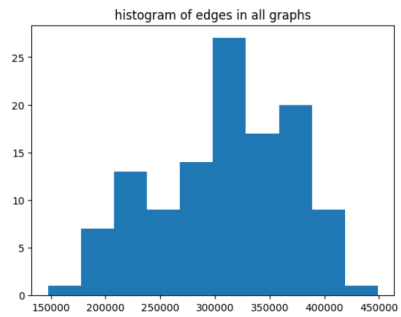

Figure 2 - Distribution of the number of edges

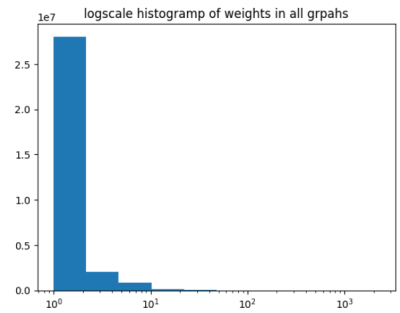

Figure 3 - Distribution of weights

As we explained before, our graphs are weighted based on the number of time two words appeared in the same sentence. In the following graph we can see the distributions of the weights in all the graphs. The average weight is 1.554 and the median weight is 1.

Section C - Defining the analysis and research:

- **Degree Centrality:** measures the number of connections (edges) a node (word) has in the network. Nodes with higher degree centrality are more connected to other nodes, indicating their importance.
- **Betweenness Centrality:** calculates the extent to which a node lies on the shortest paths between other nodes. Nodes with higher betweenness centrality act as bridges or intermediaries between different parts of the network.
- **Closeness Centrality:** metric measures how close a node is to all other nodes in the network. Nodes with higher closeness centrality can quickly access information from other nodes and have a higher influence.
- **Eigenvector Centrality:** This metric considers not only the direct connections of a node but also the connections of its neighbors. Nodes with higher eigenvector centrality are connected to other important nodes in the network, indicating their significance.
- **PageRank:** assigns importance scores to nodes based on their incoming links and the importance of the linking nodes. Nodes with higher PageRank scores are considered more influential in the network.

We will use these metrics in order to determine which nodes/words have the most effect indicating the main topics spoken in the relevant day news broadcast.

- **Clustering Coefficient:** This metric measures the extent to which nodes in a network tend to cluster together. It provides insights into the level of community structure within the network.
- **Community Detection:** This involves identifying groups or communities of nodes that have stronger connections among themselves compared to connections outside the community. Various algorithms, such as modularity optimization or hierarchical clustering, can be used for community detection.

These metrics will help us learn about which are the topics discussed in the news broadcast for the relevant day. By examining the communities created in the graphs we can generate the main topic for each community, and by these we can examine which communities were more relevant each day.

- **Spl -** can help identify the strength and nature of connections between different words or topics, providing a quantitative measure of their proximity or distance within the network. can be valuable for understanding how certain events/themes relate to each other and evolve over time in the news broadcast
- **Modularity** is a widely used measure in community detection algorithms to evaluate the quality of the identified communities within a network. It quantifies the degree to which nodes in the same community have more connections with each other than expected by chance. It can help identify distinct themes, subtopics, or narrative clusters within the news broadcast, providing valuable insights into how news broadcasts evolve and shift during the major events.

Defining the comparison between the networks

As stated before, our graphs represent the connections between words in news transcripts, and we have a graph for each day around major events. Our goal is to compare between the graphs by timeline and using the metrics stated above to examine the different phenomenons that occur around major events.

1. **Temporal Comparison**: we will examine changes in network structure, the emergence or disappearance of communities, shifts in central nodes/influential topics, and differences in the strength of connections between words. This can provide insights into how news broadcasts have evolved and how the focus of discussions has changed over the days.
2. **Event-based Comparison:** compare the networks corresponding to different events, Looking for variations in community structures, key topics discussed, and the centrality of specific words or themes. This comparison can help understand how news coverage differs across events and how certain topics gain prominence or decline in different contexts.

Phenomena to Examine:

1. **Communities:** Analyze community structures within networks to identify cohesive groups of words or topics that are densely connected. This can help uncover distinct themes, narratives, or subtopics within the news broadcasts.
   This will contribute to the research questions by revealing how different events give rise to specific themes or narratives, shedding light on the framing and discourse around these events.
2. **Proliferation**: Study the spread or proliferation of certain words or topics across the network over time. This can provide insights into the diffusion of information or the rise and fall of specific issues or discussions.
   This will contribute to the research questions by identify how certain topics gain prominence, evolve, or fade away, providing insights into the changing focus and public attention around major events.

Testing and Simulation:

1. **Community Detection Algorithms**: Apply community detection algorithms such as modularity optimization, hierarchical clustering. Using these algorithms to identify and analyze communities within the network.
2. **Temporal Analysis:** Compare network snapshots at different days to examine the evolution of community structures and topic proliferation over time.

Section D -Visualization of the network(s)

In order to visualize our graphs we plotted an interactive graph for the top 50 edges with highest weight. We also created and interactive dashboard that let's the users change through the graphs for each day of the relevant event and examine the change over time.