**1st Assignment: Compute different measures on a large graph**

**Description of the Network:**

The network under analysis represents the directed email interactions within the Democratic National Committee (DNC) based on the 2016 email leak. Each node in the network corresponds to an individual within the dataset, and each directed edge represents an email sent from one person to another. Given that an email can have multiple recipients, a single email generates multiple directed edges, resulting in a network where the number of edges is approximately twice the number of emails in the dataset.

**Summary of Network Properties:**

* **Number of nodes:** 1891
* **Number of edges:** 39264
* **Average degree:** 41.53
* **Node with highest degree:** 1874 (calculated using the sum of in-degree and out-degree)
* **Highest Degree Value:** 5813
* **Highest In-Degree Value:** 2951 (Might be the Backup server node)
* **Highest Out-Degree Value**: 2862 (Might be the broadcast node)
* **Giant component size:** 520
* **Average shortest path length (Giant Component):** 3.05
* **Graph density:** 0.011
* **Average clustering coefficient (Simple Graph):** 0.209
* **Transitivity (Simple Graph):** 0.089
* **Assortativity (Simple Graph):** -0.301
* **Diameter (Giant Component):** 8
* **Diameter path (Giant Component):** [480, 808, 761, 1838, 278, 1669, 1159, 1274, 1667]

**Visualizations:**

**Graph:**

**A network of black and blue dots

Description automatically generated**

**Giant Component:**

**A blue and black network

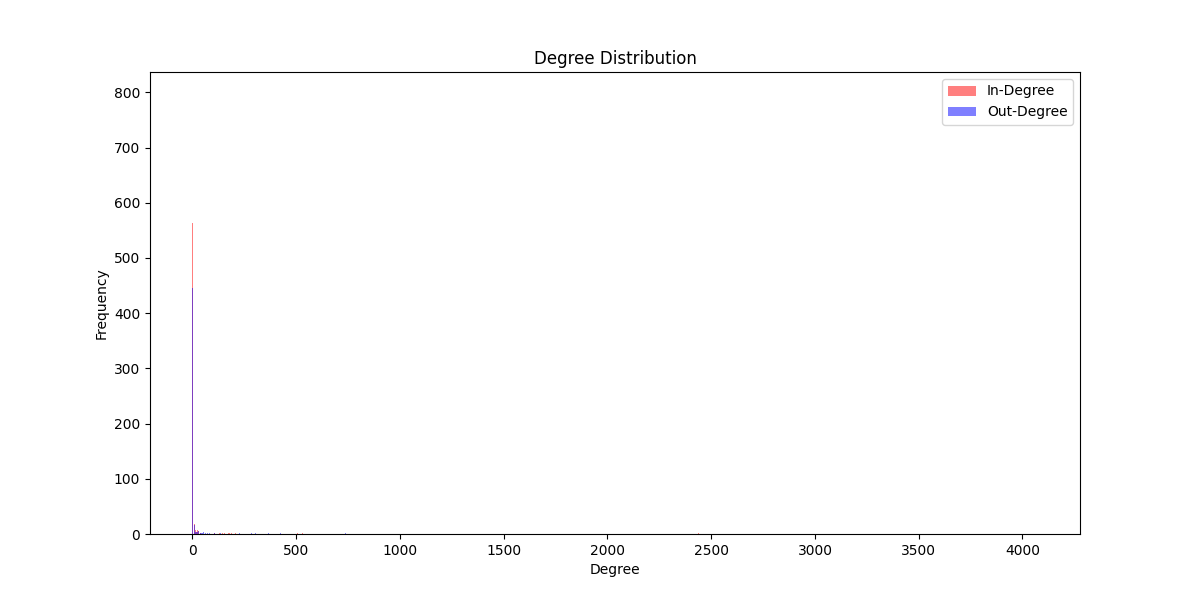
Description automatically generated**

**Detailed Analysis:**

**Degree Distribution:**

* + The degree distribution of the network exhibits a heavy-tailed behavior, typical of many real-world networks. The in-degree and out-degree distributions both show that a small number of nodes (individuals) have a very high degree, acting as hubs in the network, while the majority of nodes have a low degree.
  + The most common in-degrees and out-degrees are low, with the highest frequencies for nodes having degrees of 1 or 5.

A graph of a degree

Description automatically generated

[\*Csv file of in-out-degree of nodes](file:///Users/mohammadhosein/Library/CloudStorage/OneDrive-unige.it/SEM2-2024/NA/Network-Analysis/Final%20Project/Frames/Assignment1Frames/degree_distribution.csv)

**Centrality Measures:**

* **Betweenness Centrality:** Most nodes have low betweenness centrality, suggesting that only a few nodes serve as critical bridges in the network. These nodes are crucial for maintaining connectivity.

A graph with a number of points

Description automatically generated with medium confidence

* **Closeness Centrality:** A few nodes have high closeness centrality, indicating their central position within the network, allowing them to communicate efficiently with many other nodes.

A graph with blue and black bars

Description automatically generated

* **Eigenvector Centrality:** High eigenvector centrality for a few nodes signifies their influence, as they are connected to other well-connected nodes, reinforcing their importance in the network.

A graph with a number of lines

Description automatically generated with medium confidence

**Shortest Path Length:**

* The average shortest path length within the giant component is relatively low (3.05), indicating a small-world property where most nodes can be reached from any other node through a small number of steps.
* The shortest path length distribution further supports this, showing a peak at 2-3 steps, with very few paths longer than 5 steps.

**A graph of a path length distribution

Description automatically generated**

**Graph Diameter:**

* The graph's diameter is 8. This reflects a compact structure in the DNC email network. Also, the path is not included the our highest (In-Out) Degree node.
* The Path : [480, 808, 761, 1838, 278, 1669, 1159, 1274, 1667]

**A red line and blue dots

Description automatically generated**

**Assortativity:**

* The negative assortativity coefficient (-0.301) suggests disassortative mixing, meaning that nodes with high degrees tend to connect with nodes of lower degrees. This is indicative of a hierarchical structure where highly connected nodes (hubs) interact with many peripheral nodes.

**Community Structure:**

* The network displays a clear community structure, with 100 distinct communities identified using the Info map algorithm. These communities are indicative of tightly knit groups within the DNC email network, with dense interconnections within each community and sparser connections between different communities.

A diagram of a network

Description automatically generated

**Communities guesses:**

* Leadership and Executives
* Campaign Staff
* Communications and Media Relations
* Policy and Research Teams
* Fundraising and Finance
* Regional or State Committees
* Volunteers and Grassroots Organizers
* IT and Infrastructure Support:

The Infomap algorithm was applied to the DNC email network, which consists of 1891 nodes and 5598 edges. The analysis identified 100 communities, with 67 being non-trivial.

* **Initial Codelength:** 7.1624
* **Final Codelength:** 5.9550
  + **Modules Codelength:** 0.5270
  + **Leaf Nodes Codelength:** 5.4279

**Conclusion:**

The DNC email network exhibits several key characteristics typical of complex real-world networks:

* **Heavy-tailed Degree Distribution:** Highlighting the presence of hubs that are crucial for maintaining network connectivity and communication.
* **Centrality Measures:** Identifying critical nodes that serve as bridges, central communicators, and influential members within the network.
* **Small-World Property:** Facilitating efficient communication and rapid dissemination of information.
* **Community Structure:** Reflecting the organizational subgroups within the DNC, likely corresponding to different teams or departments.
* **Disassortative Mixing:** Indicating a hierarchical structure where central hubs interact with many peripheral nodes.

**2nd Assignment: Sending information through a network**

**Objective:**

The second assignment aims to simulate the spread of information in a network, particularly focusing on how information propagates when there are both gossipers (who spread true information) and malicious nodes (who tamper and spread false information). The goal is to understand the impact of different parameters on the diffusion of true and false information within the network.

**Network Description:**

This analysis is conducted on the directed network of emails in the 2016 Democratic National Committee (DNC) email leak. The network represents email communications, where nodes correspond to individuals, and directed edges represent the act of sending an email.

**Key Steps in the Code:**

1. **Parameter Setting:**
   * **Gossipers and Malicious Nodes:** Gossipers are nodes selected to start spreading the true information. Malicious nodes, chosen based on their degree, are those that tamper with the information.
   * **Information Initialization:** Gossipers start with the "super secret sentence (True info)", while malicious nodes start with no information but will tamper with true information when they receive it.
2. **Information Spread Simulation:**
   * **Neighbors Consideration:** Each node considers its neighbors (both successors and predecessors) to decide whether to accept and forward the information based on a threshold.
   * **Threshold Check:** If a sufficient fraction of a node's neighbors have forwarded the information, the node will adopt and potentially forward it.
   * **Malicious Behavior:** Malicious nodes change true information to false information before forwarding it.
3. **Visualization and Analysis:**
   * **Frames Creation:** The spread of information is visualized and saved as a series of images (frames).
   * **Cosine Similarity Analysis:** The final state of the network's information spread is analyzed using cosine similarity to understand the similarity of the information held by different nodes.
   * **Graphs and Bar Charts:** The impact of various parameters (threshold, number of gossipers, number of malicious nodes) on the spread of true and false information is visualized using graphs and bar charts.
4. **Parameter Variations:**
   * The simulation runs for different values of thresholds (0.1 to 0.5), gossipers rates, and malicious rates to study their impact on information diffusion.

**Results Analysis:**

* **Informed Nodes:** The number of informed nodes (both correct and incorrect) is tracked at each step of the simulation.
* **Correct vs. False Information:** The number of nodes holding correct versus false information is recorded.
* **Impact of Parameters:** The graphs and bar charts illustrate how different combinations of gossipers and malicious rates, along with varying thresholds, affect the diffusion of information.

**Results :**

**Scenario 1: Gossipers are High Degree Nodes**

**Description**: In this scenario, gossipers are selected as nodes with high degrees. These nodes are central in the network, meaning they have many connections and can potentially spread information rapidly to a large number of nodes.

**Observations:**

1. **Information Spread:** The true information propagated quickly through the network due to the high connectivity of gossipers.
2. **Malicious Influence:** Depending on the placement of malicious nodes, they could intercept and alter the information. However, the high degree gossipers ensured that true information reached a significant portion of the network before malicious nodes could tamper with it extensively.
3. **Resulting Graphs:** The graphs would likely show a rapid increase in nodes holding true information initially, with a potential gradual increase in false information if malicious nodes are effectively positioned.

**Scenario 2: Malicious Nodes are High Degree Nodes**

**Description: In this scenario, malicious nodes are selected as high degree nodes. These nodes have many connections, allowing them to quickly spread false information once they intercept true information.**

**Observations:**

1. **Information Spread: The true information may start spreading, but the high degree malicious nodes quickly intercept and alter it, resulting in a rapid spread of false information.**
2. **Malicious Influence: The influence of malicious nodes is significant in this scenario, leading to a higher prevalence of false information in the network.**
3. **Resulting Graphs: The graphs would likely show an initial spread of true information, followed by a sharp increase in false information as the high degree malicious nodes begin tampering and spreading the altered information.**

**Scenario 3: Both Malicious and Gossipers are Low Degree Nodes**

**Description: In this scenario, both gossipers and malicious nodes are selected as low degree nodes, chosen randomly. These nodes have fewer connections, making the spread of information slower.**

**Observations:**

1. **Information Spread: The spread of both true and false information is slower due to the low connectivity of the nodes involved.**
2. **Malicious Influence: Malicious nodes can still alter information, but their impact is less pronounced compared to the high degree scenarios due to their lower connectivity.**
3. **Resulting Graphs: The graphs would show a gradual and slower increase in both true and false information, with the final distribution depending on the random placement and effectiveness of gossipers and malicious nodes.**

**Conclusion**

**Key Insights:**

1. **Node Degree Impact: The degree of nodes chosen as gossipers or malicious nodes significantly impacts the spread and final distribution of information within the network. High degree nodes facilitate rapid information spread.**
2. **Malicious Node Influence: High degree malicious nodes can quickly alter and spread false information, demonstrating their potential to disrupt information integrity in the network.**
3. **Random Low Degree Nodes: When both gossipers and malicious nodes are low degree and chosen randomly, the spread of information is slower and more localized, with a mixed distribution of true and false information.**

**Visual Representations:**

**The visual representations (graphs and charts) generated from the simulations provide a clear picture of how information spreads under different scenarios. They highlight the dynamics of true and false information propagation and the critical role of node degree in influencing these dynamics.**

**Future Considerations:**

1. **Threshold Variations: Further experiments with varying thresholds can provide insights into how sensitive the network is to changes in the fraction of neighbors needed to forward information.**
2. **Real-World Applications: Understanding these dynamics can help design more robust communication networks and strategies to counteract the spread of false information, particularly in scenarios involving high-degree malicious actors.**

**By analyzing these scenarios, we gain a deeper understanding of the network's behavior and the critical factors that influence the spread of information, both true and false. This knowledge is invaluable for developing strategies to manage and mitigate the impact of malicious actors in real-world networks.**