CPT_S 591 - ELEMENTS OF NETWORK SCIENCE - PROJECT REPORT

SOCIAL NETWORK ANALYSIS - FACEBOOK & TWITTER

Pallavi Arivukkarasu and Pallavi Sharma

Department of EECS

Washington State University, Pullman, WA, USA

Abstract

Social Network Analysis (SNA) is the process of performing analysis on social media networks to determine smaller structured groups through the study of graph theory and network science. It also aims to capture the key individuals that have a significant impact on the network and map relationships between the nodes that ultimately forms the base of the network. Apart from these reasons, social network analysis also determines the information flow, which is critical for improving communication and gathering knowledge. This paper describes our motivation to perform Social Network Analysis, the patterns/trends we were able to read, and how it is beneficial to track rumor spreading.

1. Introduction

Social media forms an indispensable part of one’s life in today's world. Regardless of gender and age, people are active on at least one of the social networking websites such as Twitter, LinkedIn, Facebook, and Instagram, among others. The social media fact sheet released by the Pew research center in 2021 reveals that today 7 in 10 Americans use social media to connect and share information. It also states that Facebook is the most widely used social media platform, followed by Instagram, LinkedIn, and Twitter. Global social media statistics show the same trends. According to the statistics released by Global WebIndex in 2022, more than half of the world’s population, i.e., 58.4%, uses social media. The average time spent on social media daily has been recorded as 2 hours and 27 mins.

All these statistics demonstrate that the user base for online social networking platforms is enormous. When viewed as networks, these platforms are a graphical structure with nodes being represented by the individuals and relationships among these nodes defined by their connections. For instance, for networks like Facebook and LinkedIn, the nodes would be the users, and the edges between the nodes would signify if they are friends with each other. An in-depth analysis of the structure of these social networks can help us find great insights relating to human behavior and relationships. This social network analysis can serve to compare and contrast different social networks. We can apply the discovered patterns to real-world problems, such as devising an efficient way to prevent rumor spreading and implementing a marketing strategy to maximize profits. The utility of Social network analysis in real-world situations and the zeal to enhance our technical abilities motivated us to perform the analysis of two popular social networks- Facebook and Twitter. Our approach and applied methodologies were synonymous with the many research studies being done on social networks.

For this project, we procured the dataset and explored basic network properties associated with the network. We then proceeded towards visualization of the network and computation of centrality measures. We compared the analysis and results for both networks. Epidemic diffusion models were applied to the networks to understand which model resembles human behavior in the context of rumor spreading.

The rest of the report is organized as follows: Section 3 will formally define the problem statement, section 4 will enlist the algorithms and models we have applied for our analysis, and Section 5 will take you through our systematic implementational journey. This will be followed by presenting our results in section 6, a description of related work in section 7, and a conclusion in section 8.

2. Problem Definition

Through the social network analysis, we intend to answer the following questions. We have divided these into the below categories:

1. **Comparison**:

   We wanted to know how similar and different the two social networks are in their network structure and properties. We also wanted to observe if they generate a similar trend when put under similar conditions.

2. **Real-World and Random Networks**:

   We were interested in determining if the social networks possess characteristics similar to real-world networks or random networks such as Erdős–Rényi or Barabási–Albert graphs.

3. **Small-World Phenomena:**

   We wanted to verify if the social networks chosen by us exhibit small-world phenomena as proposed by Milgram.

4. **Rumor Spread:**

   We wanted to determine the epidemic model that can best describe human behavior in the use case of rumor spreading. This can be a significant step toward implementing a model that helps control the spread of misinformation.

5. **Community Detection:**

   We are interested in finding communities, i.e., sets of strongly connected nodes within a network for both these networks. This can give us an insight into how many communities can be present within a social network, which could be a game-changer in the field of customer segmentation and marketing.

3. Models/Algorithms/Measures

The questions in our problem statement inspired us to apply the below techniques.

3.1. **Network Analysis:** We performed network analysis to understand our networks better. This helped us gain a fundamental understanding of the structural properties of the networks. For instance, we compared the degree distribution of Facebook and Twitter networks with real-world networks and random graphs. We plotted histograms and curves of frequency vs. degree on absolute and log scale to understand the distribution of degrees across the network. We plotted both the in-degree and out-degree of the nodes for directed graphs. We also computed the central nodes based on different centrality measures such as degree, betweenness, closeness, eigenvector, and page rank. We. also determined graph properties such as diameter, average path length, average local clustering coefficient, and connected components.

3.2. **Visualization of Networks:** We plotted the networks using different libraries to understand and visualize the properties computed in the Network Analysis section. We also tried to visualize the central nodes by plotting these networks while considering centrality measures. For instance, in the case of degree-based visualization, we intended to make the central nodes bigger in size and different in color to visualize them better across the network.

3.3. **Diffusion Modeling:** Social networks provide an excellent platform for information diffusion where millions of people can share information. However, this information is not always accurate. Many times, misinformation or rumors get transmitted through a social network. This transmitted information is called a Network rumor. Diffusion models can help prevent the spread of a rumor by giving us an idea about the extent of the spread in a defined period. To see how a rumor spreads quickly across the networks, we implemented four types of epidemic diffusion modeling. We applied these epidemic models because the spread of misinformation in a social network is conceptually similar to the propagation of viral infection in an epidemic model.

3.4. **Community Detection:** We implemented the Community Detection concept to determine the number of communities with common interests in the network. We used the Louvain method for the partition process. This algorithm is a greedy optimization method that runs in n times log n time complexity and hence, takes comparatively lesser time to run for such large networks.

4. Implementation/Analysis

4.1. Language and Libraries used

We used python from start to end to implement this project. Through this project, we got an opportunity to explore a variety of libraries built for dealing with network structures. Below are a few libraries that we used extensively:

1. NetworkX: We used it to read data, visualize, obtain structural properties, and compute centrality measures.
2. Ndlib: This library was used to study and implement different diffusion models for our large and complex networks.
3. Community: We used this library to apply the Louvain method to our networks.

The table below shows the entire list of libraries we utilized and the model we used it for.

Table 1: Libraries

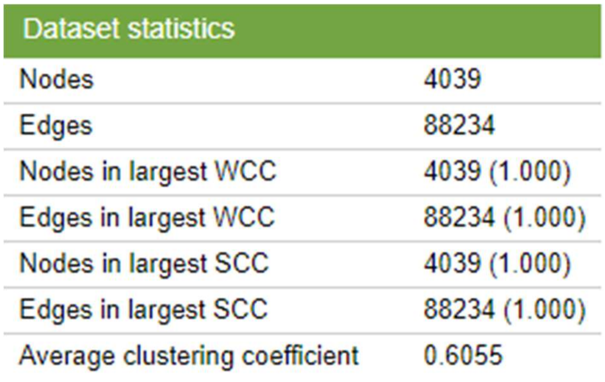
Models / Techniques	**Libraries**
Network Analysis	NetworkX, SNAP for Python
Visualization of Networks	NetworkX, Graphistry, GraphViz, Matplotlib, Seaborn, PyVis
Diffusion Models	Ndlib, Pyplot Viz, Bokeh Viz
Community Detection	NetworkX, Community, Matplotlib

4.2. About the data

Since we wished to conduct our analysis and test our hypothesis on at least one large network, we obtained our network datasets from SNAP (Stanford Large Network Dataset Collection). SNAP contains large network datasets, and it contains graphs describing citation networks, social networks, communication networks, web graphs, and much more. We selected Facebook and Twitter datasets for our analysis. SNAP compiled both datasets in the year 2012. The Facebook data is smaller in size than Twitter. While the Facebook data has been hand-labeled and compiled by Prof Julian McAuley (currently in UCSD) and Prof Jure Leskovec (currently at Stanford University), the Twitter dataset was taken from publicly accessible data. Both these datasets were cleaned and analyzed to invent a novel machine-learning algorithm to detect social circles by these professors.

- Facebook: This dataset contains a list of 88,234 edges, with each edge signifying a friendship between two users. There are a total of 4,039 users and hence, 4039 nodes. Since friendship represents a bidirectional connection, this network is undirected. The table below shows complete information about the network:

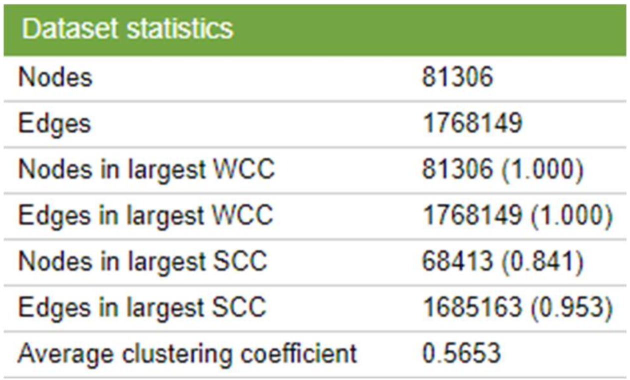
Table 2: Facebook Dataset Statistics



Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055

- Twitter: This dataset contains 81306 nodes and 1768149 edges. A connection or edge from Node A to Node B represents that user A follows user B in this network. Hence, unlike Facebook, this is a directed graph. The table below shows complete information about the network:

Table 3: Twitter Dataset Statistics



Dataset statistics	
Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653

4.3. Understanding the Networks

Since social networks are small-world networks, we believe our networks should exhibit small-world phenomena. We tested our hypothesis using the following activities:

- Both the datasets were edge lists, and hence, we used the read_edgelist() function provided by NetworkX to convert them into a network.
- Verification of the network created from the edge list by running through the info() function of the graph. It gives out the nodes and edges of the graph created. We matched this with the information published on the SNAP website to ensure data sanity.
- We also calculated the number of weakly and strongly connected components from the networks. The size of the largest component was also computed.
- The average path length, i.e., the average smallest distance between two nodes of a network, and the diameter of the networks were computed.
- The average clustering coefficient, i.e., the average of each node's local clustering coefficient, is also calculated.

4.4. Network Visualization

Since social networks are usually closely-knit dense networks; we believe that Facebook and Twitter networks should present a similar structure overall. To test this hypothesis further, we performed the following activities:

- We tried to visualize our network to get an intuitive sense of the state of the network. The idea was to realize better the network properties computed before. We used the following methods to achieve this:
- We used the draw() function of networkX to draw both of our graphs; however, the graphs obtained couldn’t depict the connections and nodes.
- We then utilized libraries that are used to visualize large networks. We started with the GraphViz library; however, it failed to produce a clear visualization
- We then used the Py Vis library to draw the networks. Using this library, we could capture the true essence of the Facebook network. Due to the very high computation time required (more than 24 hours), we could not achieve the visualization of the Twitter network.
- We utilized the Graphistry python API to visualize the Twitter network. It not only produced a clear and distinguishable visualization but also required minimal computational time.

4.5. Distribution-based Analysis

We believe that these social networks are real-world networks. We applied the degree distribution analysis to test our hypothesis:

- We plotted histograms representing the frequency of degree on the Y-axis and the degree of the network on the X-axis.
- Since histograms did not give a clear picture of the distribution of high-degree nodes, we decided to plot the line curves that presented a true representation of high-degree nodes
- Since Twitter is a directed network, we plotted curves for both in-degree and out-degree.
- To verify if the distributions follow the power law, we plotted these distributions on a log-log scale, i.e., taking log on both the X and Y-axis.
- We utilized matplotlib and seaborn libraries to achieve this.

4.6. Centrality Based Analysis

This analysis was performed to determine how many central nodes are common when centrality is determined using five centrality measures. This also serves as a way to compare the two networks. We believe that the similarity of nodes among different centrality measures for both the social networks should be almost equal or very close to each other.

- The centrality-based analysis estimates how vital a node or edge is to detect the flow of information and connectivity in any given network.
- We have implemented the five measures of centrality analysis: Degree, Betweenness, Closeness, Eigenvector, and PageRank.
- We used the networkX functions- degree_centrality(), eigenvector_centrality(), pagerank_centrality(), closeness_centrality() and betweenness_centrality()
- We also plotted the networks using the degree centrality where we showed the nodes with higher degree with different color and larger size for better visualization.

4.7. Diffusion Models

We hypothesized that the SEIR epidemic model best describes human behavior regarding rumor spreading. Our approach to connecting epidemic models with the real-world rumor spreading problem is as follows:

For the SI model:

- The infected population(I) can be represented as the number of users who received, read, and were convinced by the rumor. These nodes in social networks are likely to be the transmitters of rumors.
- The susceptible population (S) can be represented as the users who don’t have any clue about the rumor.

For the SIS model:

- This stands similar to the SI model, with a slight change that here, we consider that human memory is limited; hence, after a point of time, the infected population can become susceptible again.

For the SIR model:

- In this model, we relate the status Removed with the number of users who read the news and either didn't believe in it or didn't have the time to forward it. Hence, during a defined course of time, a user can go from the status of susceptible to infected to removed. This is a great model for describing human behavior when encountering false information through a post or tweet via social media.

For the SEIR model:

- We have an extra status E in addition to the SIR model in this model. This status Exposed can be represented by the users whose news feed has received the news but hasn't noticed it yet. This model truly represents human behavior as it covers all possible cases that may happen when a user gets to know about misinformation[4].

We applied all the above models to test our hypothesis. We used the ndlib library to implement these. Throughout the analysis across both the networks and models, we tried to keep all the testing conditions and parameters the same.

- The initially infected nodes were the top 5% nodes with the highest degree of centrality measure.
- The infection probability was kept at 5%, while the removal probability was 1%
- The number of iterations considered was 200.
- We plotted the Diffusion Trend and Prevalence plots for all these models. We also plotted the infection spread for all the models for each network for better comparison.

4.8. Community Detection

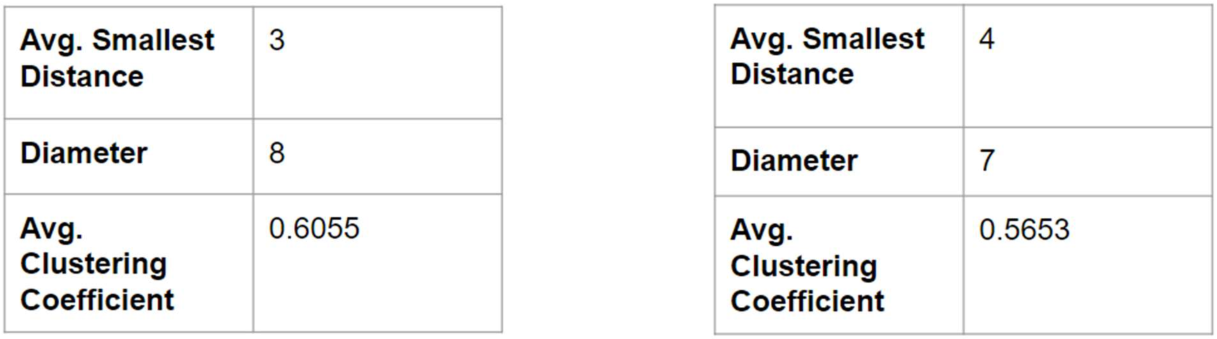
- In our quest to compare networks and find densely connected nodes in a network, we applied Louvain community detection to our networks.
- The rationale behind choosing this algorithm was its faster implementation and suitability for large networks. We used the community library for this purpose.
- The idea was to see if the number of communities detected in both networks was comparable.
- We also tried to visualize these communities by using the draw function of NetworkX and keeping the opacity of edges as 0.1 and 0.5.

5. Results and Discussion

5.1. Small World Phenomena

As described in the implementation section, we calculated the network properties of both the networks. Our hypothesis that social networks exhibit small-world phenomena is valid. As we can see from the table below, both Facebook and Twitter have a smaller average distance and a higher average clustering coefficient, concluding that these networks exhibit a small-world phenomenon.

Table 4 : Small-World Phenomenon Metrics



Avg. Smallest Distance	3
Diameter	8
Avg. Clustering Coefficient	0.6055

Avg. Smallest Distance	4
Diameter	7
Avg. Clustering Coefficient	0.5653

While we have utilized the primary characteristics of small-world networks to prove the validity of our hypothesis. We feel that plotting the average path length distributions of the networks could have strengthened our claim further.

5.2. Network Visualization

Our theory claims that since both networks belong to the category of social networks, they should look visually similar, irrespective of the difference in the size of both networks. We were correct in our claims, as we can see that the images below follow similar patterns. We have no isolated nodes. They both have a dense center which is the largest connected component. Our empirical data also strengthens this claim. The Facebook network is one strongly connected graph that can be clearly inferred from the image. Twitter, on the other hand, is one weakly connected graph; however it has 12,448 strongly connected components. But the largest strongly connected component accommodates 84% of the nodes, which can be seen in the dense center of the network in the image.

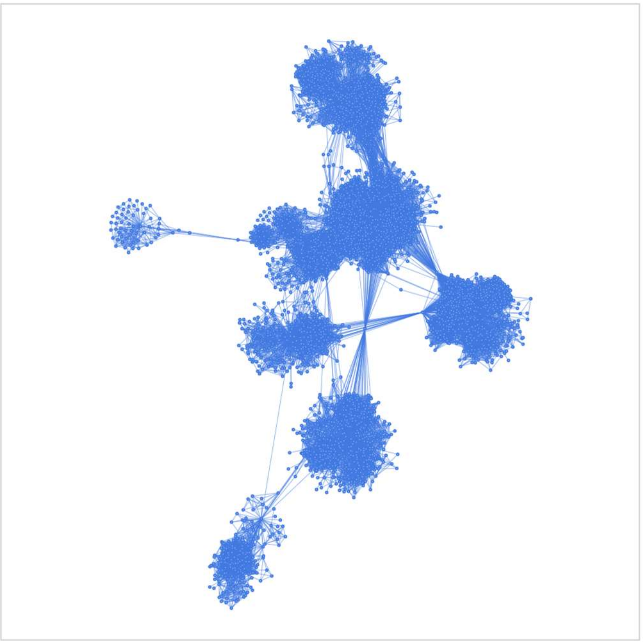


Figure 1: Facebook Network Visualization

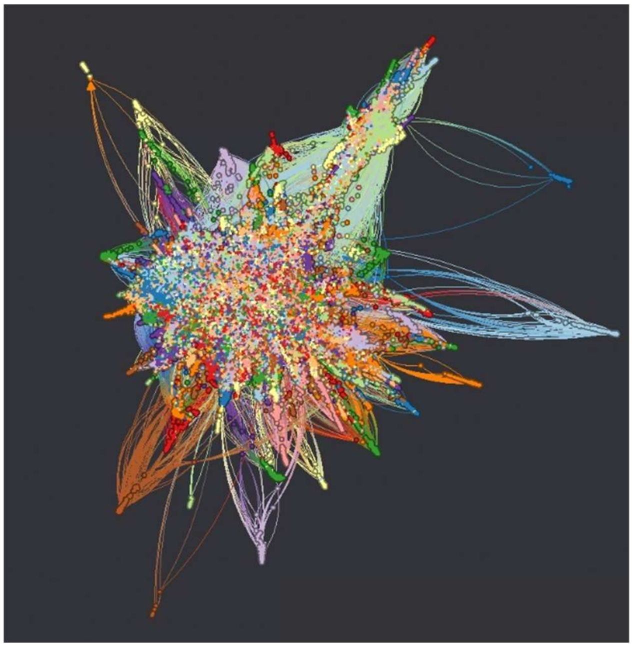


Figure 2: Twitter Network Visualization

5.3. Distribution-based Analysis

We started with the belief that Facebook and Twitter's social networks should be more similar to real-world networks over random networks. We were correct in our hypothesis as the degree distribution plots show that both the networks depict a long-tailed distribution, indicating that the networks indeed are real-world networks.

The first plot is the number of nodes vs. degree and the second plot depicts the number of nodes on a logarithmic scale vs. degree for the Facebook network.

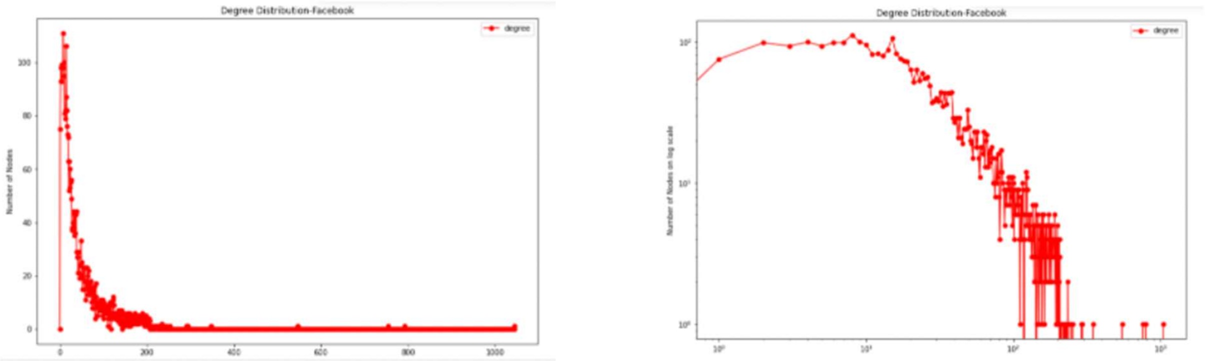


Figure 3: Degree Distribution Analysis for Facebook

We first plotted a degree histogram for the Twitter network. As you can see, there is not much information as the number of degrees increases, so we then plotted the in-degree vs. out-degree on a logarithmic scale which is mainly considered for directed networks.

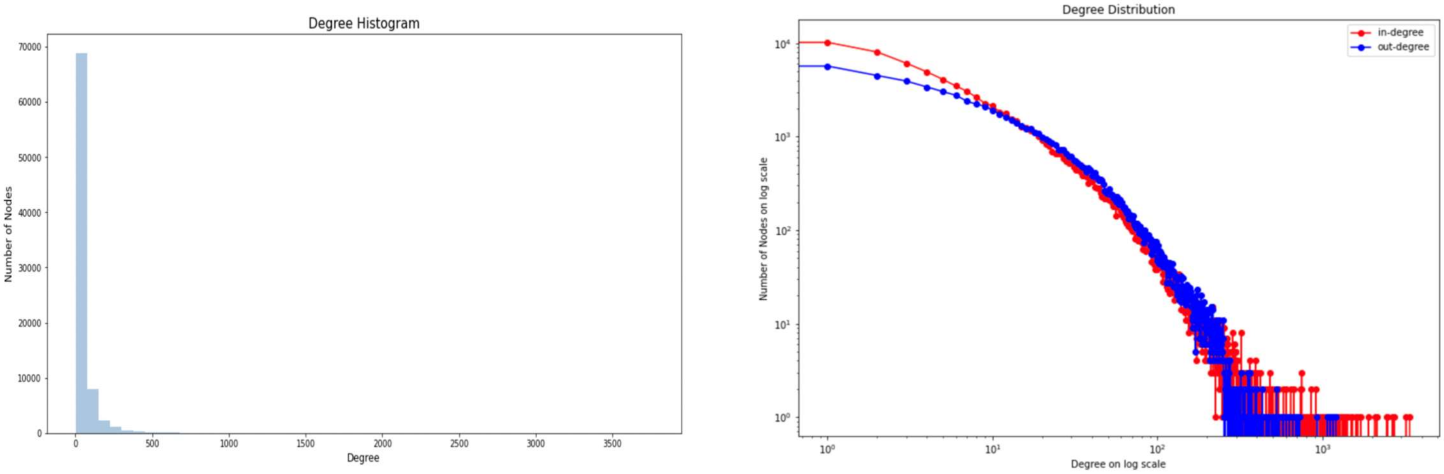


Figure 4: Degree Histogram for Twitter

The long-tailed distribution plots for the number of nodes vs. in-degree and the number of nodes vs. out-degree are given below:

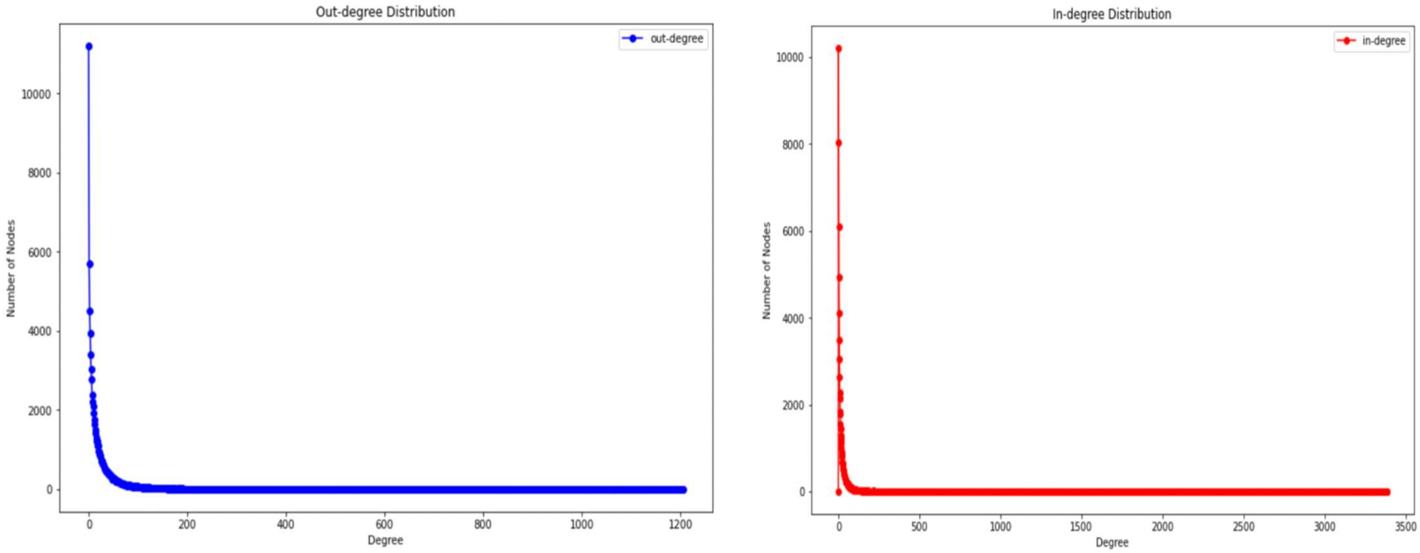


Figure 5: In-degree and Out-degree Distribution for Twitter

We believe that plotting average path length distributions of the networks could have bolstered our claims.

5.4. Centrality-Based Analysis

We calculated the top 10 central nodes for both networks based on the five centrality measures. The results are presented below:

For Facebook:

- Degree Centrality: [107, 1684, 1912, 3437, 0, 2543, 2347, 1888, 1800, 1663]
- Closeness Centrality: [107, 58, 428, 563, 1684, 171, 348, 1663, 414, 2543]
- Betweenness Centrality: [107, 1684, 2543, 1912, 1085, 0, 698, 567, 1663, 428]
- Eigenvector Centrality: [1912, 107, 1663, 2543, 2464, 2142, 2218, 2078, 2123, 1684]
- Pagerank Centrality: [2543, 107, 1684, 0, 1912, 348, 686, 1663, 414, 698]

For Twitter:

- Degree Centrality: [813286, 115485051,40981798,3359851,43003845,22462180,34428380,59804598,7861312,15913]
- Closeness Centrality: [115485051,813286,7861312,15485441,16303106,1183041,783214,14075928,90420314,972651]
- Betweenness Centrality: [15485051, 40981798,3359851, 7861312,11348282,17093617, 18396070,14230524, 18996905, 117674417]
- Eigenvector Centrality: [40981798,22462180,43003845,34428380,31331740,27633075,18996905,83943787,17868918,117674417]
- Pagerank Centrality: [115485051,116485573,813286,40981798,7861312,11348282,17093617,15439395,18396070,14230524]

We observe that there is a similarity of 50% among the top 10 nodes detected by each centrality measure in Facebook. We also realized a similarity of 30% among the top 10 nodes detected by each centrality measure for the Twitter network. Since we claimed that the similarity in nodes should be close to each other, our hypothesis stands incorrect. In our opinion, this difference is due to the vast difference in the size of the two networks. We conducted a similar analysis by taking a larger set of central nodes; however, the percentage of similarity among different measures decreased with an increase in size, suggesting that in large networks, central nodes can be quite different based on the centrality measure selected for calculation.

Plots for degree-based network visualization can be seen below:

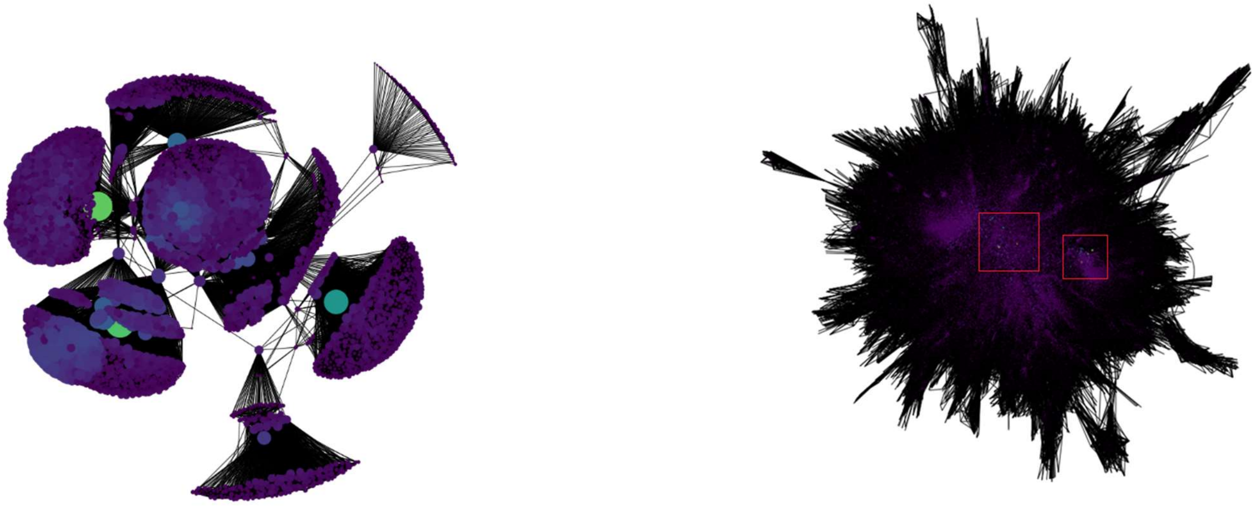


Figure 6: Degree Centrality for Facebook & Twitter

5.5. Diffusion Modeling

We obtained differing diffusion trends by applying the SI, SIS, SIR, and SEIR models to both networks.

For Facebook, both the models' SI and SIS show that the users become infected in minimal iterations. SIR and SEIR models paint a realistic picture showing that the rumor does take some time to spread; it affects 90% of users in the SIR model and around 65% of users in the SEIR model. The inclusion of different statuses representing users who don’t believe in the rumor (R) and users who have received the rumor but haven't had the time to view it (E) make the SIR and SEIR better suitable for assessing the extent of a rumor spread. The plots are shown below:

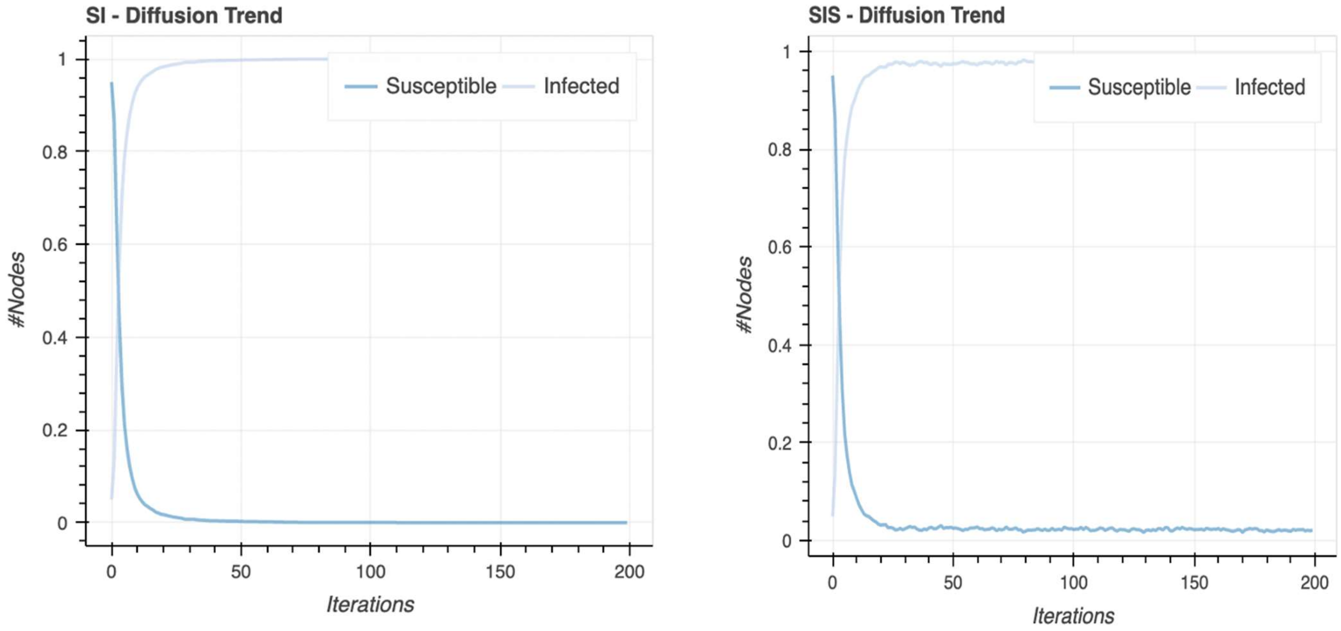


Figure 7: SI and SIS Diffusion Trends for Facebook

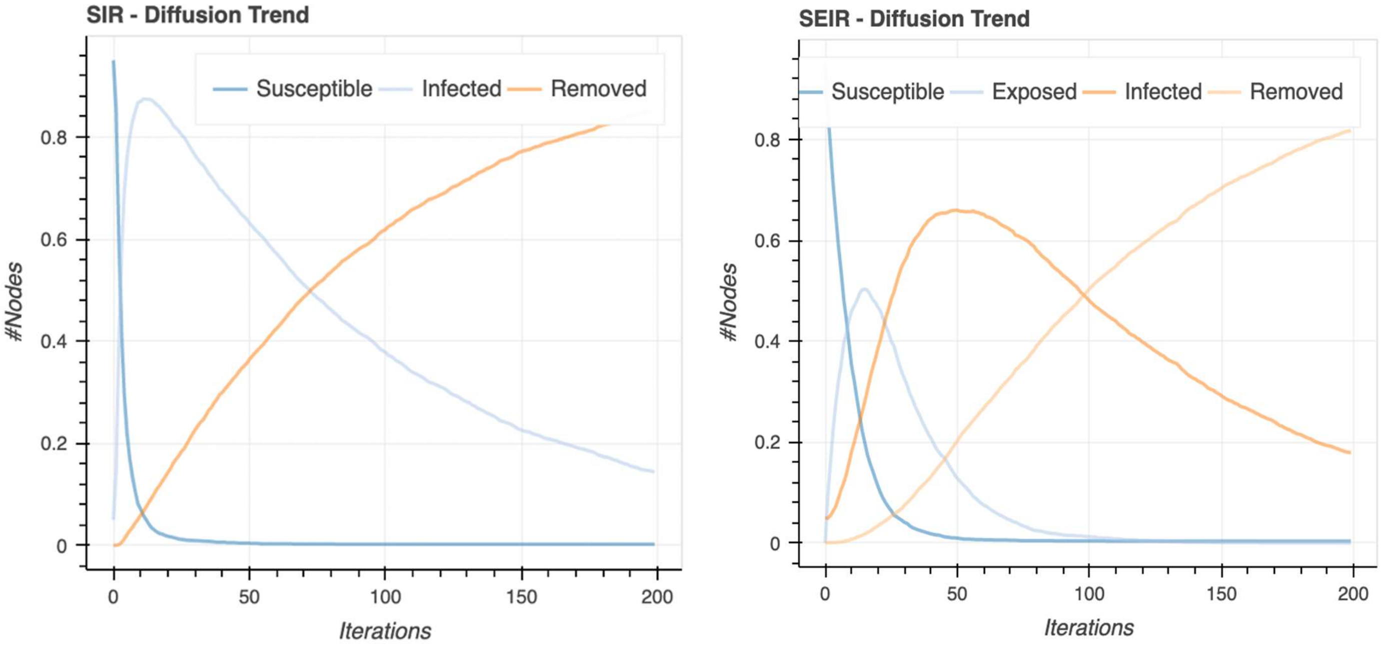


Figure 8: SIR and SEIR Diffusion Trends for Facebook

We experience a similar trend for Twitter; the SI and SIS models show a similar trend, with all 100% of the users getting infected quickly. The SIR and SEIR diffusion plots show trends similar to the Facebook network and hence, resemble human behavior more intuitively. The plots are shown below:

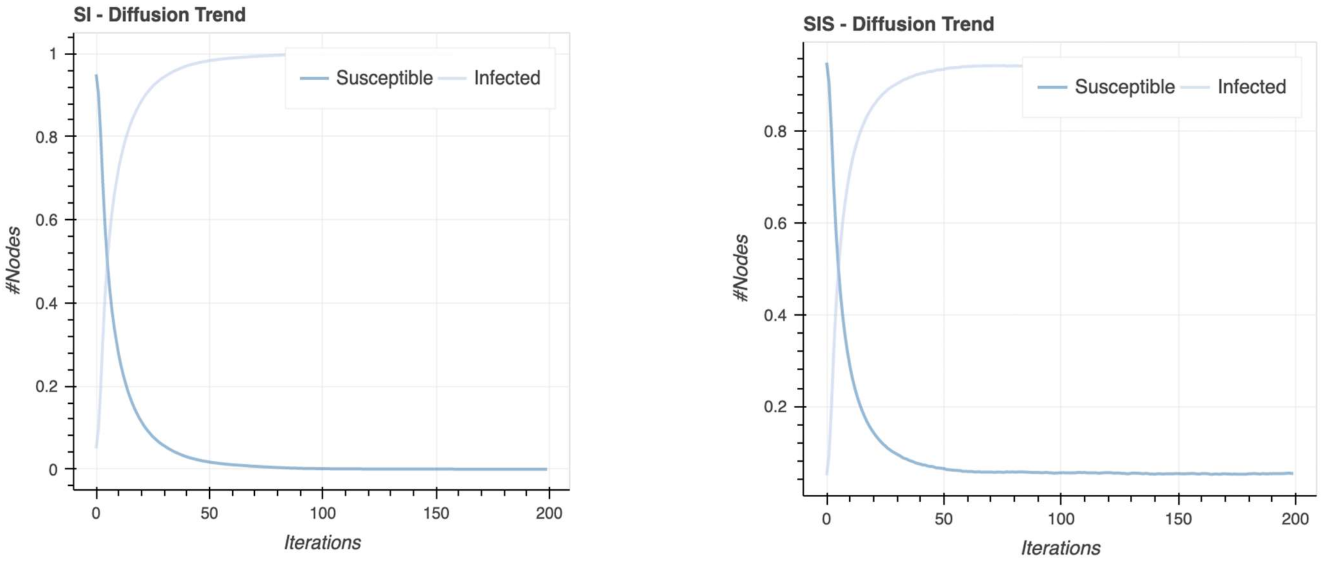


Figure 9: SI and SIS Diffusion Trends for Twitter

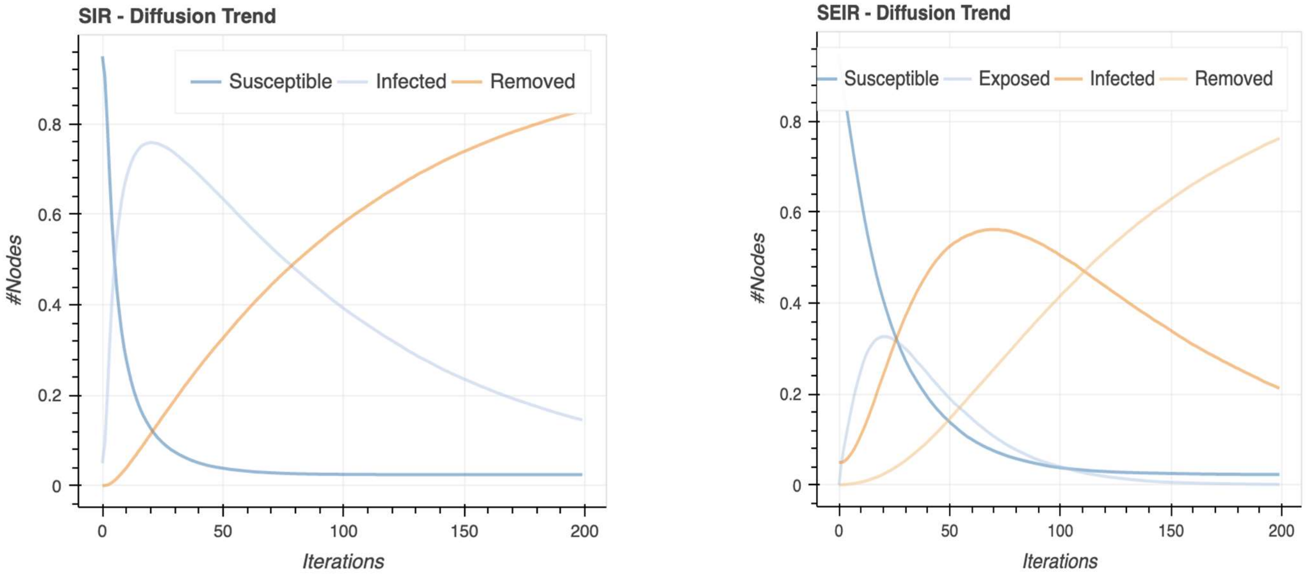


Figure 10: SIR and SEIR Diffusion Trends for Twitter

We decided to plot the infected trend from all the models for each network in a single graph for better comprehension. Plot 1 below shows the infected trend for the Facebook network, and plot 2 represents these trends for the Twitter network.

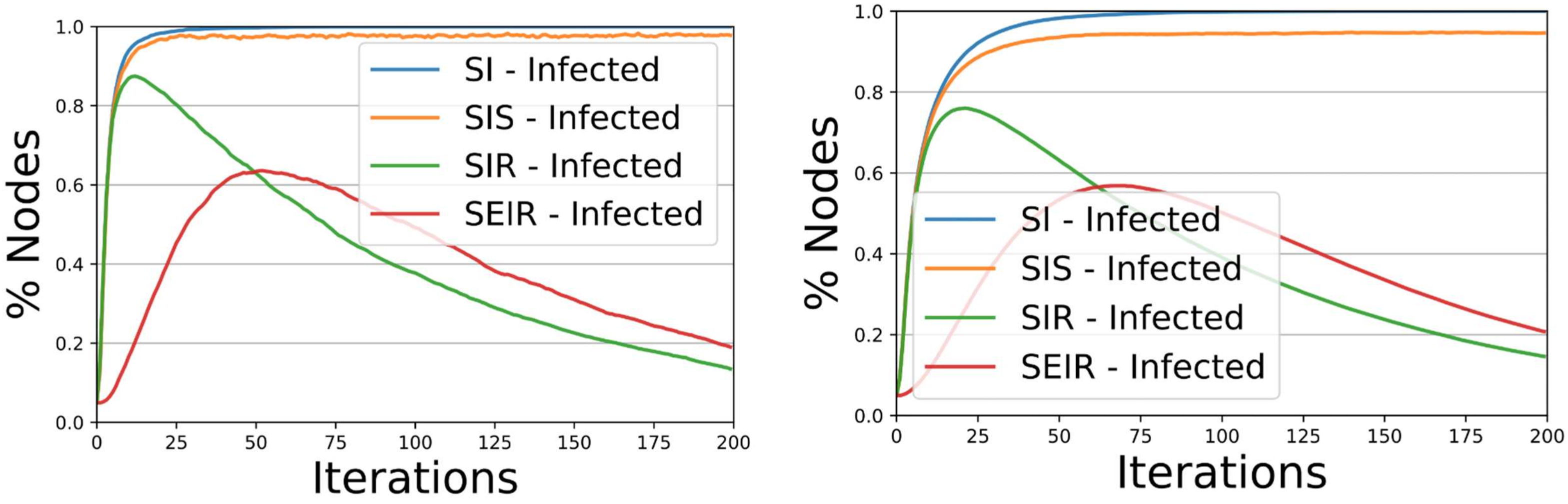


Figure 11: Infected trends for Facebook and Twitter

In the SEIR model, we observe that a maximum of 62% on Facebook and 58% of users on Twitter can get affected by a rumor, which seems logical considering that not all users will be aware of the rumor at all times. Some would not believe it, some would forget it, and some would not notice it. Since the SEIR model accounts for all these cases and shows that it has good control over the spread through empirical analysis, our hypothesis stands valid. We believe that tweaking the parameters and conditions could have given us more insights about these models.

5.6. Community Detection

Our hypothesis regarding communities within a network stands incorrect. The number of communities detected are not close to each other. We found 15 communities with common nodes for the Facebook network. For Twitter, 71 communities were found. We believe that the larger number of communities in Twitter is due to the large size of the network and a greater number of strongly connected components. The plots showing the communities are presented below:

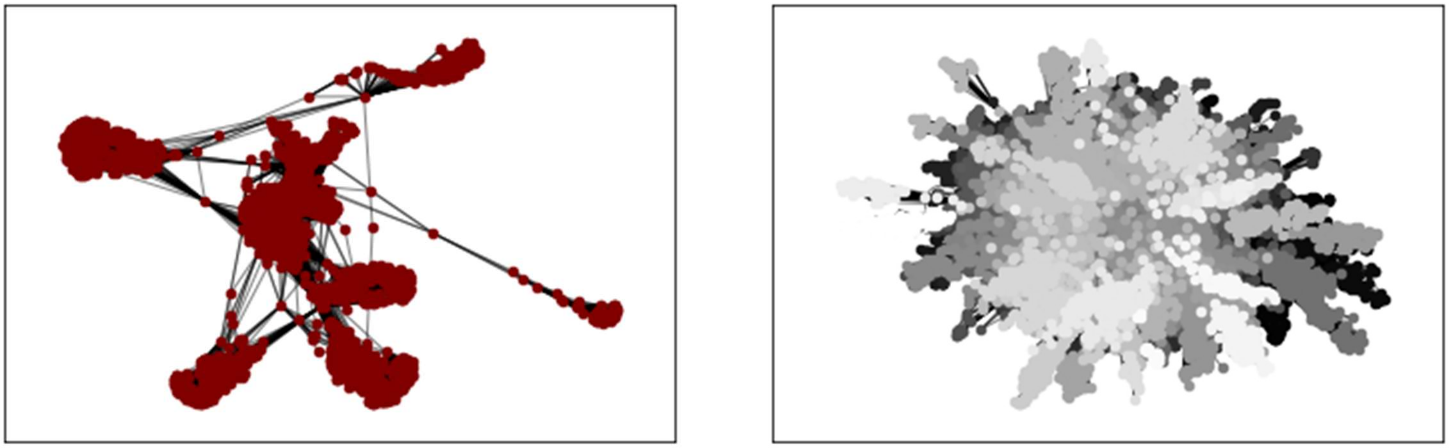


Figure 7: Community Detection for Facebook and Twitter

The communities are more apparent in the Facebook plot over Twitter. Applying more community detection algorithms could have helped us verify the number of communities present within the network.

5.7. Comparison of Networks

Throughout the analysis, we tried to find similarities between these networks. We have come up with the following conclusions:

- Both the networks are structurally similar.
- They are both small-world networks.
- They are both real-world networks.
- They show similar diffusion trends for all the four epidemic models.
- They possess communities within their network.

6. Related Work

There has been a plethora of research on information diffusion in social networks. The most widely used epidemic model was the SIR model; earlier, most of the research revolved around improving this model. However, the SEIR and SEIZ models are gaining popularity in recent times. These models are being extensively used to control the rate of information diffusion. One paper worked towards using SEIR as the model for rumor tracking; they also considered the effect of change in population i.e., the total number of users in their analysis[4]. Another paper implemented the SEIZ model for analyzing the misinformation spread during the Black lives matter movement[5]. We have applied four epidemic models to our networks in our project, including SEIR and SIR models. We decided to see which model best fits human behavior regarding rumor spreading. No specific papers compare these two social networks based on their structural properties and behavior; however, many papers go a notch further and compare these networks based on spam messages and profiles as well as usage. One paper compared Facebook and Twitter in terms of effective market strategies to implement. It was proved that both Facebook and Twitter required similar marketing strategies to attain success[6].

7. Conclusion and Future Work

With the increasing number of social media networks, every application has allowed people to stay connected every day. In this study, we have showcased that both Facebook & Twitter are small-world and real-world networks. We have also analyzed how quickly a rumor can spread in these networks using diffusion modeling. Since people with common interests have a lot of things to share, using community detection, we determined there are 15 communities on the Facebook network and 71 communities on the Twitter network.

For future research, more types of diffusion models can be implemented to study the flow of information. Parameter tuning for the diffusion models can be studied in-depth and applied to the models so that we can control the spread of rumors in a network.

8. Bibliography

1. https://snap.stanford.edu/data/ego-Facebook.html
2. https://snap.stanford.edu/data/ego-Twitter.html
3. Leskovec, Jure, and Julian Mcauley. "Learning to discover social circles in ego networks." *Advances in neural information processing systems* 25 (2012).
4. Dong, S., Deng, Y., & Huang, Y. (2017). SEIR Model of Rumor Spreading in Online Social Network with Varying Total Population Size. *Communications in Theoretical Physics, 68*, 545.
5. Leung, Xi & Bai, Billy & Stahura, K.. (2015). THE MARKETING EFFECTIVENESS OF SOCIAL MEDIA IN THE HOTEL INDUSTRY: A

COMPARISON OF FACEBOOK AND TWITTER. Journal of Hospitality and Tourism Research. 39.

6. Maleki, Maryam & Mead, Esther & Arani, Mohammad & Agarwal, Nitin. (2021). Using an Epidemiological Model to Study the Spread of Misinformation during the Black Lives Matter Movement.
7. https://snap.stanford.edu/snappy/index.html
8. https://ndlib.readthedocs.io/en/latest/reference/reference.html#
9. https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae
10. https://ndlib.readthedocs.io/en/latest/reference/models/epidemics/SIm.html

9. Appendix

https://github.com/Pals0405/NS_Project