Twitter Sentinel – A Networks Backed Approach to Controlling Fake News

Susan Koruthu (A0231905L) *Felipe Chapa Chamorro (A0179033E)* *Widya Gani Salim (A0231857Y)* *Gino Martelli Tiu (A0231956Y)*

1. BUSINESS CONTEXT & MOTIVATION

For many people, Twitter has become an alternative source for breaking news. With 59% of its 436M users [1] leveraging it for this purpose, ensuring content veracity becomes paramount. This is especially so in the high stakes area of politics, where misinformation has caused far-reaching polarization as seen in both the 2020 US presidential elections and Brexit.

Owing to the ease of creating and joining groups, the major concern at hand is the proliferation of echo chambers on Twitter. Based on a study[2], fake news is: (a) 70% more likely to be retweeted, (b) be six times as fast in reaching the same quantity of people and (c) have "cascade depths" or unbroken retweet chains that are 10 about 20 times that of facts.

Hence, a control and mitigating mechanism is required to: (a) identify influencers in powerful communities, (b) determine the nature of said entities and (c) take required measures for those with malicious intent.

In particular, the team aims to show the proposed end-to-end (E2E) process needed to disrupt the spread of fake news by **limiting the activity of infected nodes**.

2. DATASET & HIGH-LEVEL METHODOLOGY

2.1 Data Overview

The FakeNewsNet dataset was utilized – more specifically, the Politifact subset.

The structure of the original data is provided below for ease of reference. Information from the 2 sides - (a) user and (b) news & interaction - were processed using Python to come up with the main dataframes.

USER METADATA

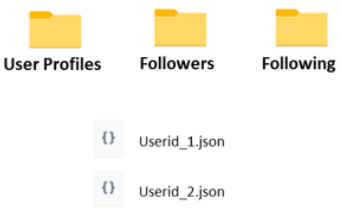

Figure 1. Social context data structure

NEWS & INTERACTION METADATA

Figure 2. News data structure

Dataframe outputs are: (a) **tweet_retweet.csv** and (b) **follower_followee.csv**.

2.2 High Level Approach & Toolkit

Outlined as follows is the team's blueprint for approaching the problem. A section is dedicated for each component to be discussed in more depth.

Table 1. Methodology and toolkit

Components	Key Output	Techniques/ Tools
Data insights	Master dataframe and key statistics	TigerGraph Data mining
Network form	Data in node-edge form as per package requirements	MapReduce NetworkX iGraph
Communities	Communities and points of influence	* Leading Eigenvector * Asynchronous Label Propagation

[1] Mitchell, Amy et.al., "News on Twitter: Consumed by Most Users and Trusted by Many" https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/

[2] Vosoughi, Soroush et al., "The Spread of True and False News Online " https://www.science.org/doi/10.1126/science.aap9559

Components	Key Output	Techniques/ Tools
	Identification of malignant communities and nodes	* Fluid Communities * Kernighan Lin Maximization
Disruption	Disabled malignant communities by limiting the activity of least number of nodes possible	*Spark GraphFrames *Targeted attack based on highest degree

3. EXPLORATORY DATA ANALYSIS

3.1 Key Statistics

Using a combination of TigerGraph and data mining techniques, the below key figures and insights were derived from the dataset:

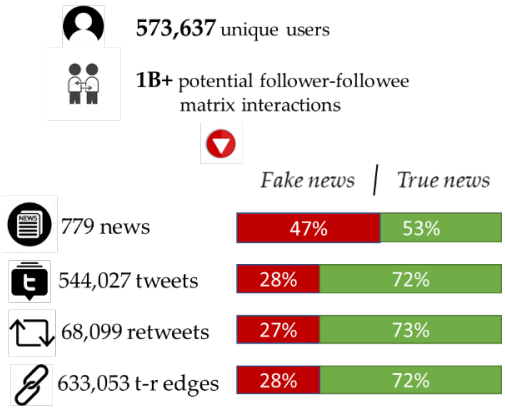

Figure 3. Network data statistics

3.2 EDA Insights

In order to determine how to best frame the business objective as a network problem, data exploration was conducted to respond to 2 key questions expounded as follows:

3.2.1. Which user activities or interactions are more effective in spreading fake news?

Twitter activities and interactions that were considered for edge links were as follows: (a) follower-followee relationships, (b) tweet-retweet activities and (c) likes.

The correlation between the above interactions and the number of fake news related instances relative to the activity in question are shown as follows:

Table 2. Correlation: Fake news count vs interaction

Interaction	Correlation
Tweet-Retweet	0.344
Follower-followee	0.375
Like-Tweet	-0.013

Each relationship is plotted and explained as follows:

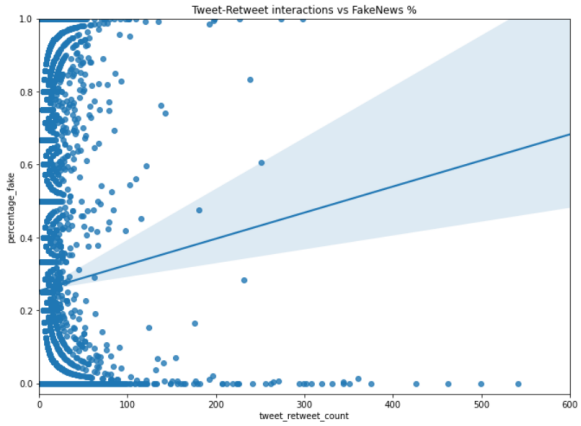

Figure 4. Tweet-Retweet vs Fake news % scatterplot

Tweet-retweet vs Fake news % shows the distribution of fake news % vs number of tweet-retweet interactions a given user has. Modeling the relationship this way shows a positive correlation between interactions and fake news.

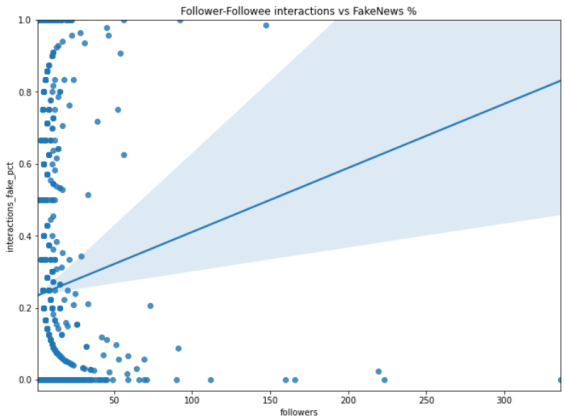

Figure 5. Follower-followee vs Fake news %

Follower-followee vs Fake news % *as* the network structure results in a slightly stronger positive correlation, however, the number of interactions is sparse in comparison.

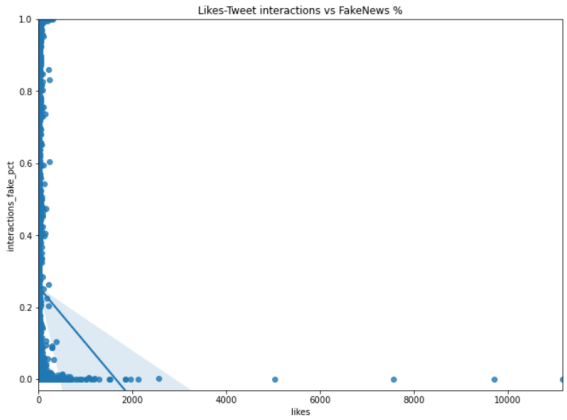

Figure 6. Likes-Tweet vs Fake news % scatterplot

Likes-tweet vs Fake news % shows a high number of interactions with a negative correlation between variables. At the same time, most tweets seem to have similar and small amounts of likes. As such, modeling interactions this way seems like a weaker option.

From the aforementioned results, the team chose to model **tweet-retweet and follower-followee relationships** as 2 network graphs for assessment. This decision was made based on 2 arguments:

- Said edge relationships showed the highest correlation to fake news activity.
- Network graphs based on this structure make intuitive sense from a user perspective, as not all types of Twitter user actions are created equal. To be more specific, retweeting and following accounts based on content shared are more deliberate decisions and can be argued to have more weight than say, liking a post.

3.2.2. Based on the network structure, which nodes actually wield significant influence in a community?

Page Rank analysis was done to determine the important nodes in the network. One thing to note was that Page Rank analysis failed to converge for the follower-followee network due to dead ends and issues on how connected the network structure is. Hence, said analysis is only applicable to the tweet-retweet network – a finding that will later inform our choice of network formulation.

Note also that Page Rank was done for 2 versions of the dataset: (a) all news and (2) only fake news. Interestingly, pages ranked as influential in the "fake news" dataset also figured highly in the "all news" one. The below table illustrates this finding.

Table 3. Fake news spreaders PageRank ranking

Username	All news rank	Fake news rank
US_Dem_Voices	8th	1st
trumpcardiac	21st	2nd
KagForce1	17th	3rd
JamesOKeefeIII	31st	4th
Rtwefm	56th	5th

Analysis of the ranked nodes for the fake news subset tags the following users as the most influential fake news spreaders.

3.2 Measuring Information Flow

Leveraging on network definitions, the team chooses the concept of degree to numericize how well information flows across the network. Measuring the change in this value will later indicate how well the team's approach works in disrupting the flow of fake news.

4. NETWORK PROBLEM FORMULATION

With insights derived from exploratory data analysis, the direction of how to formulate the business objective as a network problem is clear.

(1) The network will model tweet-retweet interactions with node-edge definitions shown as follows:

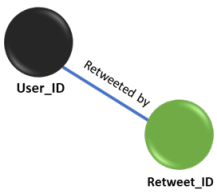

Figure 7. Tweet-Retweet network schema

(2) Flow of information is measured using degree. The same measure will indicate effectiveness of the disruption strategy chosen.

5. COMMUNITY DETECTION

5.1 Rationale for Community Detection

Moderating activity and banning users is an expensive control measure both financially and reputation-wise.

Hence to streamline the disruption process undertaken downstream, the team decided to implement the following flow:

(1) Cluster nodes into communities
(2) Only target top communities
(3) Limit the activity of top users that have high degree

Since the goal is not to disrupt the functionality of the entire network but to control the spread of fake news, community detection is a logical first step. This is primarily because deploying a targeted monitor and control effort to top communities will make the most business sense in terms of feasibility and expected return.

5.2 Community Detection Algorithms Explored

To this end, four community detection algorithms were explored.

1) **Asynchronous Label Propagation:** This algorithm repeatedly sets a label to each node depending on the label that appears most amongst its neighbors. It is done without waiting for updates from all nodes, hence asynchronous.

2) **Kernighan Lin Maximization:** Nodes are assigned communities. The algorithm then checks if moving a node will increase modularity. If it does, then it is moved. Otherwise, it remains as is. This continues until every node is moved at least once. The steps above are repeated until no increase in modularity is observed.

3) **Fluid Communities Algorithm:** Initialize communities. For each node, sum the densities of the surrounding communities. The node is assigned to the community with max density. When a node assignment is changed, the community density is adjusted. This is iterated until convergence occurs.

4) **Leading Eigenvector Algorithm (LEA):** Starts with all nodes being part of the same community. A modularity matrix is used to split the graph to maximize modularity. This continues until maximum modularity is achieved.

Modularity-wise, LEA showed the best cohesion and was therefore utilized to create the communities.

5.3 Insights from Detected Communities

To gather some insights from the detected communities, the maximum and the average of 4 network metrics – Degree, Hub Score, Authority Score and Betweenness – from the entire network, as well as from the top 5 communities generated by the chosen algorithm, were observed.

Table 4. Maximum Network Metrics

	Degree	Hub Score	Authority Score	Betweenness
All	1177	1.00	1.00	976.00
1	146	1.00	1.00	66.00
2	28	1.00	1.00	10.00
3	31	1.00	1.00	8.00
4	19	1.00	1.00	8.00
5	16	1.00	1.00	5.00

It is noticeable from Table 4 that there exist nodes which have extremely high degree of 1,177 and betweenness of 976. Since communities are sub-graphs of the entire network, it is expected that the highest degree and betweenness of each community are much lower than the entire network.

Table 5. Average Network Metrics

	Degree	Hub Score	Authority Score	Betweenness
All	1.46	0.000030	0.0000085	0.20
1	2.06	0.0032	0.13	0.44
2	2.08	0.0097	0.086	0.22
3	2.12	0.011	0.055	0.11
4	2.05	0.0064	0.051	0.29
5	2.01	0.015	0.090	0.17

From Table 5, it can be observed that in general, the entire network has lower network metrics than all communities. This could suggest that the network is less connected as compared to the detected communities. Moreover, community 1 seems to have the highest betweenness and authority score as compared to both the entire network and other

communities. Contrasting to the point earlier, this could indicate that community 1 is more connected than others.

Meanwhile, community 5 has the highest hub score and its authority score is also only slightly lower as compared to community 1. This could suggest that this community contain more highly influential users than other communities. Lastly, it is notable that degrees are comparable between the communities, depicting that, on average, the number of connections that the users within the communities have is comparable.

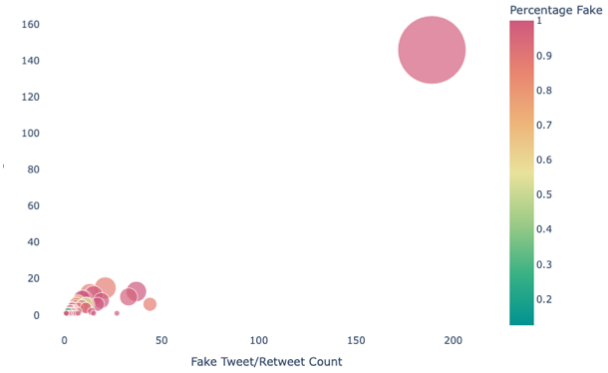

Figure 8. Fake Tweet Count vs Percentage Fake vs Degree for Community 1. Points sized based on degree and colored based on % of Fake Post.

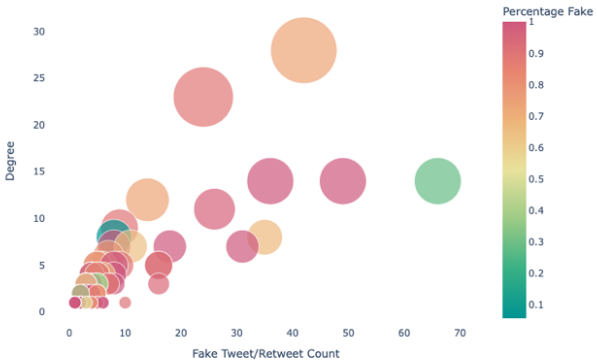

Figure 9. Fake Tweet Count vs Percentage Fake vs Degree for Community 2. Points sized based on degree and colored based on % of Fake Post.

Further key observations that can be derived from the communities are as follows:

- **High percentage of fake news posts per user:** The percentage of fake news of most of the entries are more than 60%, very few users had lower percentage than that. Within each community, there are always a cluster of users who have only posted fake news but have low tweet/retweet counts. These can represent fake accounts which sole purpose is to propagate fake news and supporting influential spreaders.

- **Clusters of fake accounts intentionally created to support the spread of fake news:**

- **The degrees of influential users:** The most influential fake news spreader in the network has a degree of more than 150 (as shown in **Figure 8**), while on average, the degree of other influential accounts ranges from 40 to 60. Interestingly, the highest degree of users with a low percentage of fake news posts is only 15 (as shown in **Figure 9**) – showing that they are less popular.

- **Top 5 community structure:** From the top 5 communities, 1 community has only one hub which coincidentally is also the most influential fake news speared. Another community contain the least influential spreaders with low counts of fake tweet/retweet counts. Meanwhile, the other top 3 communities have a mix between less and more influential spreaders.

In the succeeding section, communities with significant fake news activity are targeted for disruption.

6. NETWORK DISRUPTION

6.1 The Idea

Network disruption can be performed by removing edges between users. In the case of a tweet retweet network overlaid over a fake news dataset, this would entail preventing users that spread large amounts of fake news from tweeting. Users who spread a lot of fake news can be identified by looking at the degree of nodes in the network. Removing outgoing edges from a node is the equivalent of blocking a user's tweets.

6.2 A Comparison of Two Approaches

Both random and targeted attacks were attempted.

(1) **Random attacks** involve choosing nodes at random and removing all outgoing edges.

(2) **Targeted attacks** mean that outgoing edges from nodes are removed in order of descending degree. If degree is evenly spread between nodes in a

community, there would not be much difference between targeted and random attacks. When there is an uneven distribution, targeted attacks are more effective.

6.2 Evaluation Metric & Performance

Average degree is used as a measure of disruption in a network. A comparison of random vs targeted attacks on the top community is visualized below.

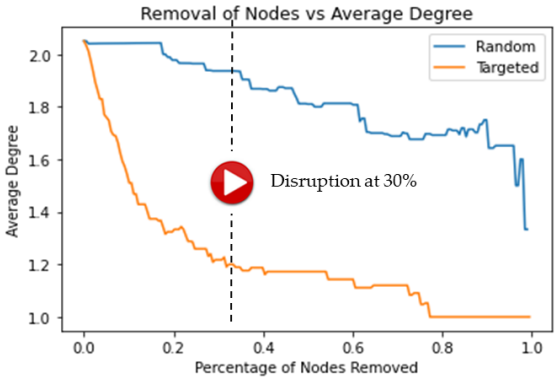

Figure 10. Network disruption sensitivity analysis

From the graph, it is clear that:

- Targeted attacks are more effective in curbing the spread of fake news in tweet retweet relationships.
- When ~30% of users are prevented from tweeting, a drop in average degree is seen for both types of attacks.
- The random attacks have a drop from 2 to 1.8 while targeted attacks have a more significant drop from 2 to 1.2 (approximately a 40% decrease).

Said insights are apparent in the pre and post disruption state of the network shown below.

PRE-DISRUPTION COMMUNITY 1

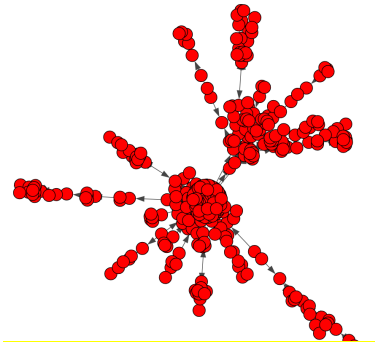

Figure 11. Community 1 before disruption

POST-DISRUPTION COMMUNITY 1

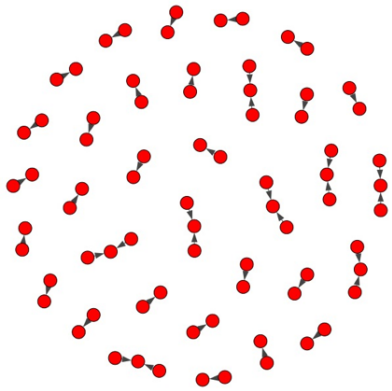

Figure 12. Community 1 after disruption

In particular, one can surmise that targeting users based on degree can show a significant change in network structure. In fact, we can see a very connected community becoming very disconnected and forming smaller groups of 2-3 users each - indicating a functional breakdown in the network's ability to spread information within a community

7. CONCLUSION AND BUSINESS APPLICATIONS

In summary, the team makes the below key points:

- **Fake news spread** is better represented through Tweet – Retweet relationships.
- **Targeted attacks** based on vertex degree is a very efficient disruption technique, reducing the **spread of fake news by 40%** (Avg. degree 2→1.2)
- More importantly, this approach makes the most business sense since a fair amount of disruption is caused to the flow of fake news while minimizing impact to the majority of business users.