**Networks and the Spread of Misinformation**

Victoria Chu

Shayna Kothari

Alex Krentsel



**Executive Summary**

Social media is a fast, easy access, low cost source of news for billions of users around the world. However, social media revenue streams incentivize clicks and views rather than quality, resulting in the creation and spread of fake news. The spread of fake news online, in particular, has become particularly relevant in the wake of revelations about the way that fake news affected the 2016 United States presidential election and other events around the world. In this project, we attempt to analyze the way that social media users spread fake and real news online, seeing what patterns emerge, how they influence the news networks, and how we can use them to prevent and predict fake news.

In particular, we are analyzing a dataset that includes a Twitter network and users who posted “real” and “fake” news to see how fake news spread online, looking at measures of connectivity within our networks to see what the differences between fake and real news datasets were. Additionally, we also specifically looked at the differences between fake and real news articles that mention current US President Donald Trump, seeing if particular issues in current events and/or differences in “shock value” lessened or greatened the disparities between the spread of real and fake news.

Our analysis showed that fake news consistently has a higher clustering coefficient than real news, indicating that fake news is more likely to spread via networks than real news. However, consistent with results from other studies on similar topics, articles that mentioned Donald Trump lessened the disparities between fake and real news, something we attribute to the “shock” often associated with articles about a current president. However, other measures of connectivity didn’t necessarily show a distinctive difference between real and fake news.

**Dataset and Objective**

In our research, we attempted to affirm findings and intuition from other research studies on fake news--in particular, “The Spread of True and False News Online,” by Soroush Vosoughi, Deb Roy, Sinan Aral, was an inspiration for our project. Vosoughi et al. had two main findings that we decided to investigate in our project:

1. Fake news is more likely to spread than real news, resulting in more highly connected graphs;
2. Fake news is more likely to spread because it is more novel (“sensational”).

For the first objective, we focused on comparing the clustering coefficients, average maximum degrees, and average path length of our tweet networks; for the second objective, we compared the values between fake and real news across a “sensational” topic likely to evoke strong emotions--either positive or negative--in readers.

Our dataset is taken from a data collection project for fake news research at Arizona State University. The dataset contacts two kinds of data; there is "news content," which includes all of the fake news articles, and "social context," which includes the social engagement of the fake news articles from Twitter. The news content provides the source, headline, body text, and image/video for each listed news article. The social context data provides the user profile fields of the users' basic information, the users' most recent posts on twitter, the list of followers of the user, and the list of people the user was following.

All sensitive user information is anonymized (i.e. hashing of Twitter handles), and for profile features and content, the dataset provides bag-of-word features. Social relationships of users (links) are maintained.

An important question that comes up early in researching anything related to fake news is for an absolute definition of fake news. Instead of tackling this definition ourselves, we decide how fake a piece of news is based on ratings from 2 independent sources - PolitiFact (PF) and BuzzFeed (BF). To be clear, these articles are not published by PF and BF, but are rated by PF and BF as being Fake or Real news (binary). We analyzed 182 articles verified by BuzzFeed and 240 articles verified by PolitiFact; each group of articles was split evenly into real and fake news.

The dataset we are using then provides the features described above for both fake and real news articles based on both PF and BF, giving us 4 distinct datasets; PF Fake News, PF Real News, BF Fake News, BF Real News.

**Approach**

To analyze each tweet network, we first created subgraphs of users for each individual news article in the dataset. We then computed measures of connectivity of the giant component of each subgraph, as graphs often consisted of large numbers of individual, disconnected nodes and then a smaller connected component. Each node describes a user, and each edge describes a connection between users that retweeted the same article. However, it is important to note that two users can independently retweet the same article and still have an edge connecting them in our graph.

Measures of connectivity we focused on were the clustering coefficient (CC), measuring the closeness of nodes in the graph; average path length (APL), measuring the distance between users in the giant component; and the maximum degree of any node in the graph (D), finding the most central user in a graph and seeing if there were many connections between it and other users. We also decided to develop an independent heuristic that was defined as H = CC x D; we hypothesized that as the clustering coefficient and the maximum degree were both higher, a news article would be more likely to be fake news, as this implies a more closely connected network of people spreading news.

We broke each subgraph into categories depending on the analysis we were doing, but generally, they were grouped by verification source and veracity. We then conducted analyses on these sets of values--real news verified by BuzzFeed; fake news verified by BuzzFeed; real news verified by PolitiFact; fake news verified by PolitiFact. We used further subdivisions for analysis based on mentions of Donald Trump, filtering articles by mentions of Trump in the bodies of the articles.

**Observations**

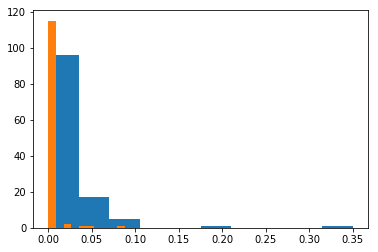
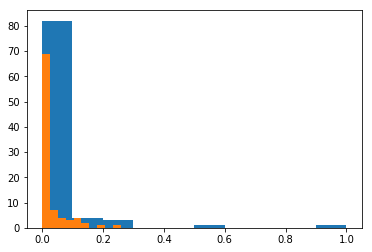
After clustering as described above, we looked at the values of the heuristics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average CC** | **Minimum CC** | **Maximum CC** |
| **Real / BuzzFeed** | 0.02189 | 0.02427 | 0.25903 |
| **Real / PolitiFact** | 0.00181 | 0.02083 | 0.08799 |
| **Fake / BuzzFeed** | 0.04331 | 0.02000 | 1.00000 |
| **Fake / PolitiFact** | 0.01692 | 0.02819 | 0.35000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average APL** | **Minimum APL** | **Maximum APL** |
| **Real / BuzzFeed** | 1.59657 | 1.0 | 5.01138 |
| **Real / PolitiFact** | 0.75629 | 1.0 | 7.14237 |
| **Fake / BuzzFeed** | 1.46901 | 1.0 | 4.50969 |
| **Fake / PolitiFact** | 1.68516 | 1.0 | 6.15864 |

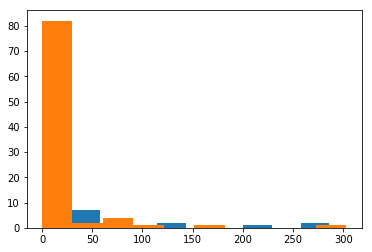
|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average D** | **Minimum D** | **Maximum D** |
| **Real / BuzzFeed** | 13.93406 | 1 | 303 |
| **Real / PolitiFact** | 2.77500 | 1 | 98 |
| **Fake / BuzzFeed** | 18.53846 | 1 | 286 |
| **Fake / PolitiFact** | 15.26667 | 1 | 697 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average H** | **Minimum H** | **Maximum H** |
| **Real / BuzzFeed** | 1.16408 | 0.43684 | 32.54653 |
| **Real / PolitiFact** | 0.05569 | 0.47917 | 3.54396 |
| **Fake / BuzzFeed** | 2.17564 | 0.14000 | 61.5637 |
| **Fake / PolitiFact** | 0.84008 | 0.25375 | 49.61036 |

****

**Distribution of clustering coefficients, BuzzFeed Distribution of clustering coefficients, PolitiFact**

**[orange=real, blue=fake]**

****

**Distribution of maximum degrees, BuzzFeed Distribution of maximum degrees, PolitiFact**

**[orange=real, blue=fake]**

For all of our heuristics, the minimum values did not differ much between real and fake news. This is expected; both fake news and real news can be unappealing to a user, and many of the differences we are looking at are important in how news articles go “viral” and whether fake news articles are more likely to do so than real news articles.

In keeping with this, the maximum values and the averages showed a clear distinction between real and fake news, particularly for our derived heuristic (H) and the clustering coefficient (CC). The averages were also distinct among the maximum degree (D), although the maximum values were not, likely owing to an individual news article’s popularity. The maximum APL of fake news articles was lower than the maximum APL of real news articles, indicating a more closely connected network of fake news spreaders.

For the BuzzFeed dataset, the maximum CC was 4 times higher for fake news than it was for real news, and for the PolitiFact dataset, it was 4.375 times higher. This means that the networks for the fake news were much more closely connected (or closely-knit) that the networks surrounding the spread of real news.

We hypothesize a few reasons for this. For one, this could be because the members of the community that believe and share these fake news stories are naturally more separated from the rest of the world and so are more interconnected. These communities may look more inward rather than outward, and such are much interconnected than the general public (assuming here that the retweeters of real news are more similar to the population of the general public).

Another interesting possibility is that there are fake accounts at work sharing the fake news. These fake accounts (also known as bots) need to appear real to be credible, and in order to appear real, they connect with each other. In this way, a nefarious actor can make a group of bots look more real by friending each other. In that case, the network sharing the fake news would have a higher clustering coefficient, because the fake users will be artificially much more likely to be connected to each other. Without de-anonymized data from the users, however, it is very hard for us to tell whether the users are real or not.

Fake news also tends to have higher maximum degree on average. This could be for a variety of reasons. It could be indicative of biases found in filter bubbles and echo chambers, where the group may share fake news articles with each other if they reinforce the group’s bias. A large degree could indicate a prominent figure or news source that many people follow and retweet. A higher degree could also indicate an interesting, memable, or satirical fake news article that even if fake is spread for a good laugh.

The disparity between degree of real and fake news (12.5) is larger in PolitiFact, which could indicate that fake news classified by this news source is harder to distinguish as fake than fake news from BuzzFeed (disparity 4.6), where this disparity is lower. This difference could also indicate polarity. More extreme groups, prominent figures, or news sources with more bias may chose to spread fake news aligned with their agenda.

\*A note: We calculated the average path length, which measures the distances between the nodes of the graphs. However, due to the varying sizes of our data, this piece of data is less conclusive than the others. We found that degree and clustering coefficient were much better predictors of the veracity of the news and focused our analysis on that.

**Further Analysis: Trump**

To analyze our second objective, focusing on shock value, we decided to distinguish between articles containing the word “Trump” and articles not mentioning him, since he is a very polarizing topic, and news surrounding him often carries large amounts of shock value. We hypothesized that there would be a smaller divergence between real and fake articles mentioning Trump than those not mentioning him.

Here are the clustering coefficients, average path lengths, and derived heuristics of our data, organized by the occurrence of “Trump” in Politifact and Buzzfeed articles respectively.

**Articles including “Trump”:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average CC** | **Minimum CC** | **Maximum CC** |
| **Real / BuzzFeed** | 0.02429 | 0.02427 | 0.25903 |
| **Real / PolitiFact** | 0.00294 | 0.02083 | 0.08799 |
| **Fake / BuzzFeed** | 0.04430 | 0.02000 | 1.00000 |
| **Fake / PolitiFact** | 0.01731 | 0.02819 | 0.35000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average APL** | **Minimum APL** | **Maximum APL** |
| **Real / BuzzFeed** | 1.5677 | 1 | 4.48512 |
| **Real / PolitiFact** | 0.8804 | 1 | 3.74004 |
| **Fake / BuzzFeed** | 1.4599 | 1 | 3.975996 |
| **Fake / PolitiFact** | 1.4373 | 1 | 6.058395 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average H** | **Minimum H** | **Maximum H** |
| **Real / BuzzFeed** | 0.98180 | 0.4368 | 14.2404 |
| **Real / PolitiFact** | 0.09031 | 0.47916667 | 3.543956 |
| **Fake / BuzzFeed** | 3.32599 | 1.5307 | 35.055188 |
| **Fake / PolitiFact** | 0.17471 | 0.3888889 | 1.750000 |

**Other Articles:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average CC** | **Minimum CC** | **Maximum CC** |
| **Real / BuzzFeed** | 0.01837 | 0.03310 | 0.19048 |
| **Real / PolitiFact** | 0 | 0 | 0 |
| **Fake / BuzzFeed** | 0.042717 | 0.02000 | 1.00000 |
| **Fake / PolitiFact** | 0.016669 | 0.02819 | 0.18383838 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average APL** | **Minimum APL** | **Maximum APL** |
| **Real / BuzzFeed** | 1.6386 | 1 | 5.0113 |
| **Real / PolitiFact** | 0.5567 | 1 | 7.1424 |
| **Fake / BuzzFeed** | 1.4744 | 1 | 4.5097 |
| **Fake / PolitiFact** | 1.8447 | 1 | 6.1586 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Veracity / Verification Source** | **Average H** | **Minimum H** | **Maximum H** |
| **Real / BuzzFeed** | 1.22220 | 0.49660 | 32.54653 |
| **Real / PolitiFact** | 0 | 0 | 0 |
| **Fake / BuzzFeed** | 2.01747 | 0.14000 | 61.56371 |
| **Fake / PolitiFact** | 1.26847 | 0.25375 | 49.61036 |

In actuality, although the relationships between real and fake news from our overall analysis hold, the divergence between average values of our heuristics for fake and real news is greater among Trump-based news than among other news. This is likely due to the higher potential for more explosive fake news associated with Donald Trump; additionally, the articles that BuzzFeed, in particular, analyzed were taken from right before the 2016 election, and many of the fake news articles during that time were intended to stir discomfort for those voting in the election. Therefore, fake news during that time may have been intended to be especially divisive.

The headlines of the articles with the maximum heuristic values in each category confirm this:

|  |
| --- |
| **TRUMP:** |
| Real headline from BuzzFeed with max heuristic = 14.240407424677981:  *National poll: Clinton leads Trump by 6* |
| Fake headline from BuzzFeed with max heuristic = 35.05518836417083:  *Surgeon General warns: Drinking every time Trump lies during debate could result in acute alcohol poisoning* |
| Real headline from PolitiFact with max heuristic = 3.543956299157183:  *Road to 270: CNN's general election map* |
| Fake headline from PolitiFact with max heuristic = 1.7500000000000002:  *Democrat Maxine Waters Has Shown Up To Only 10% Of Congressional Meetings For 35 YEARS* |

|  |
| --- |
| **OTHER:** |
| Real headline from BuzzFeed with max heuristic = 32.54653365426408:  *How to watch the first presidential debate* |
| Fake headline from BuzzFeed with max heuristic = 61.56371603134698:  *Federal Agents Make Massive Discovery at Southern Border… ISIS Is Here ⋆ Freedom Daily* |
| Real headline from PolitiFact with max heuristic = 0  *None available* |
| Fake headline from PolitiFact with max heuristic = 49.610360378112716:  *Australia Becomes First Country To Begin Microchipping Its Public* |

Most of the fake news headlines had higher derived heuristic values. Interestingly, this data on degrees demonstrates the effect of clickbait. The more outrageous or satirical the title of the fake news article, the higher the degree. For instance, “Surgeon General warns: Drinking every time Trump lies during debate could result in acute alcohol poisoning” (degree 219) is clearly satirical and “Australia Becomes First Country To Begin Microchipping Its Public” (degree 697) is a ridiculous title, and both have very high degrees. The closer the title sounds to a real article, the lower the degree disparity between the two. For instance, “Cavuto Just Exposed Lester Holt's Lies During Debate” and “How to watch the first presidential debate” have small degree disparity (just 17!), but one is real and one is fake. From the title of each, it is not immediately obvious that either is a fake news article. Finally, there may be many reasons for low degree headlines. Perhaps the headline was not relevant to the times, or it was too obviously clickbait, or it just wasn’t interesting and didn’t stand out from other similar articles. These headlines may have gotten lost in the noise of other more appealing headlines that harnessed better clickbait technique. For instance, “When does early voting start in every state?” (degree 23) is a very bland title, and probably did not make many outgoing connections in our news network. In the table above, we calculate the heuristics, which match the degree almost exactly.

**Conclusion**

Using the data points we collected on each news network, we can inform users to detect or calculate probabilities of fake news. By using our derived heuristic to represent some scale of the fakeness of the news. In the future, this data could also be used to train neural networks.

Our derived heuristic (H = CC x D) ended up being a relatively accurate distinguisher between real and fake news; in the future, these heuristic values could be used to predict whether something is true or false news. Additionally, these heuristic values provide important context that could be used to prevent the spread of fake news; the fact that fake news networks tend to be more highly clustered than real news networks can be valuable to social networks trying to prevent the spread of misinformation by providing highly clustered networks with more outside information. Many social media sites also function as news sites for their users and filter news for users based on the data they provide the network; users also filter news by choosing what to click. If we can change the filtering of information, we can change the spread of news, allowing there to be lower chances of echo chambers spreading the same news over and over again.

Further research into the matter could include filtering by topics in a more fine-grained manner and using larger datasets to increase the accuracy of our news set; we could also conduct sentiment analysis on news articles that are classified as real and fake to see how the sentiments differ and how news with different sentiments spreads through social networks. Additionally, we could look further into the community structures of our networks to analyze how users with more followers and less followers spread news.

Works Cited

Kai Shu, et al. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD*

*Explorations Newsletter*, vol. 19, no. 1, June 2017, pp. 22–36. *EBSCOhost*, doi: 10.1145/3137597.3137600.

Vosoughi, Soroush, et al. “The Spread of True and False News Online.” *SCIENCE*, vol. 359, no.

6380, p. 1146–+. *EBSCOhost*, doi:10.1126/science.aap9559. Accessed 13 Dec. 2018.