

Social Network Analysis Project

Chaoran Wei

Josh Wang

Lauren Yu

Meghan Ding

Roshan Kumar

Table of Contents

INTRODUCTION	3
DATA DESCRIPTION	5
FEATURE SELECTION	13
MODELS	14
APPLICATIONS	19
CONCLUSION	21
FUTURE WORKS	22

INTRODUCTION

The motivation of the project comes from solving an increasingly urgent problem in social media: the spread of fake news. Although in the past it has been the purview of journalists and other publishers of traditional news content, the proliferation of user-generated content on social media has made it more difficult to detect and curb the spread of fake news. The nature of online news publication has changed such that traditional fact checking and vetting from potential deception is impossible against the flood arising from content generated from social media. Hence the urgent need for new, data-driven approaches towards fake news detection.

Our project seeks to utilize social network analytics methods to detect fake news and optimize the news ‘flow’ from and to Facebook, arguably the most important social media hub. Given how ill-defined the concept of ‘fake news detection’ can be as a machine learning problem, here we strictly limit our project scope to detecting *intentionally deceptive* news content online. Using a dataset obtained from Politifact.com, which involves information of news claims, source/destination sites and other details, we were able to construct a network of “news flows” with Facebook (our client) as either source or destination. Our objectives are 1) to obtain a confidence score for any news traveling to or from Facebook, and this confidence score is the output of the classification model we constructed 2) optimize the fake news flow without significantly sacrificing real news.

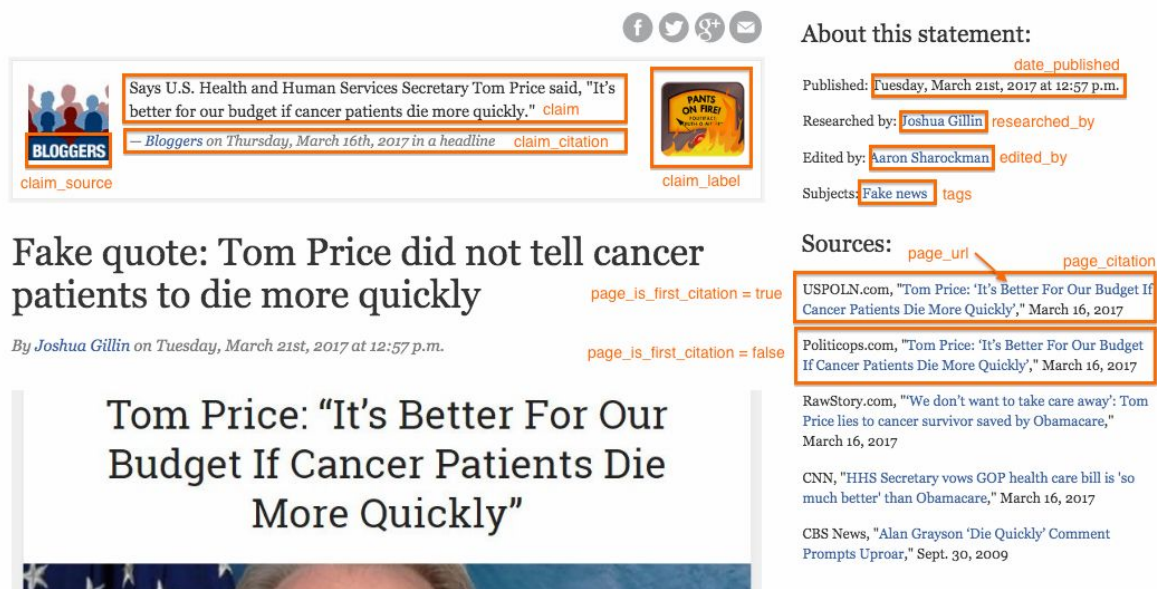
To classify whether fake news will flow from Facebook to another site (or from another site to Facebook), we labelled each source/destination pair as 1 if fake news has ever travelled through this path and 0 otherwise. For predictor features, we noticed that popular veracity assessment methods in the field of fake news detection usually fall into two camps: linguistic cue approaches (with natural language processing methods) and network analysis approaches. We see much promise in both --- and therefore we not only extracted network attributes as predictors, but also included the median polarity score as an indicator of the “emotionally intensity” of the news ‘flow’ for a given source/destination pair. Eventually, our model was able to achieve a rather high AUC score of 0.93.

It is reasonable to assume our results are limited by the data we have available. Because Politifact.com is mainly focused on debunking rumor, thus there’s substantial selection bias in the data --- not biased to the extent that we can’t conduct any classification, but still would have a sizeable impact on our results. With a more comprehensive and representative dataset, we are confident our model would be able to help readers better identify potential fake news.

DATA DESCRIPTION

Data Source

We worked with the kaggle data set “Who starts and who debunks data”. This dataset contains three files: One file is a collection of all webpages cited in Emergent.info, the second is a collection of webpages cited in Snopes.com, and the third is a similar collection from Politifact.com. The webpages were often cited because they had played a role in the rumor-spread chain --- which might be source, sharer or debunker. Of the three files, we mainly focused on politifact.com. Politifact.com follows a well-structured format in reporting and documenting rumors. Here is a snapshot of what the website looks like. There is a sidebar on the right side of each page that lists all of the sources cited within the page. The top link is the likeliest to be the original source of the rumor. For this link, `page_is_first_citation` is set to true.



claim_source: Says U.S. Health and Human Services Secretary Tom Price said, "It's better for our budget if cancer patients die more quickly." **claim**

claim_citation: — Bloggers on Thursday, March 16th, 2017 in a headline

claim_label: PANTS ON FIRE!

About this statement:

date_published: Published: Tuesday, March 21st, 2017 at 12:57 p.m.

researched_by: Researched by: Joshua Gillin

edited_by: Edited by: Aaron Sharockman

tags: Subjects: Fake news

Sources:

page_url: **page_citation**

USPOLN.com, "Tom Price: 'It's Better For Our Budget If Cancer Patients Die More Quickly'," March 16, 2017

Politicops.com, "Tom Price: 'It's Better For Our Budget If Cancer Patients Die More Quickly'," March 16, 2017

RawStory.com, "'We don't want to take care away': Tom Price lies to cancer survivor saved by Obamacare," March 16, 2017

CNN, "HHS Secretary vows GOP health care bill is 'so much better' than Obamacare," March 16, 2017

CBS News, "Alan Grayson 'Die Quickly' Comment Prompts Uproar," Sept. 30, 2009

page_is_first_citation = true

page_is_first_citation = false

Tom Price: "It's Better For Our Budget If Cancer Patients Die More Quickly"

Inferential Statistics

We used three inferential statistics methods to explore and gain a deeper understanding of the fake news network - ERGM, Power Law test, and Small-world test. We will talk about each in the following.

First, we ran ERGM (Exponential Random Graph Models) on our network. Due to the size of our network, we are not able to test many hypotheses. Below are three hypotheses we tested:

1. The edges are not random. That is, the sources do not arbitrarily post news on the destination sites.
2. The edge between two nodes is mutual. That is, if the website A posts news on website B, then website B is more likely to post news on website A.
3. Transitivity. That is, if website A posts news on website B, and website B posts news on website C, then it is likely that website A posts news on website C.

Below shows the result of ERGM:

```

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % p-value
edges      -4.78057    0.06288      0 <1e-04 ***
mutual       0.92478    0.64626      0  0.152
transitive   -Inf     0.00000      0 <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Null Deviance: 42213 on 30450 degrees of freedom
Residual Deviance:   NaN on 30447 degrees of freedom

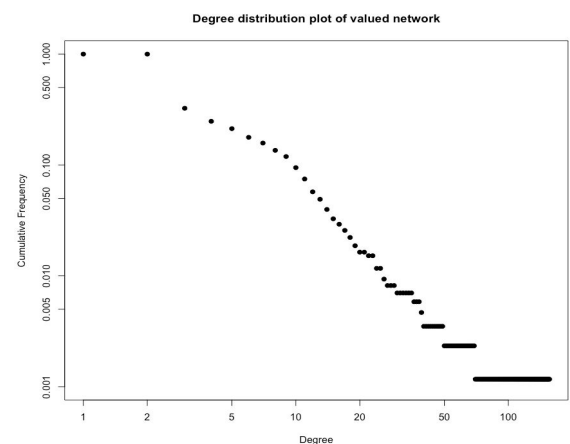
AIC: NaN    BIC: NaN    (Smaller is better.)

```

As we can see from the above result, both edges and transitivity is significant. That is, transitivity and nonrandom edge properties exist in our network. However, the high p-value in mutuality indicates that if website A posts news on website B, website B is *not* more likely to post news on website A. The AIC and BIC scores are NaN. This indicates the non-convergence of our ERGM model. Therefore, the above analysis might not be valid.

We also used Kolmogorov-Smirnov test to test whether the degree distribution of nodes in the fake news network follows the Power Law distribution. The p-value of the test is 0.87, which means we fail to reject the hypothesis that the fake news network follows the power-law distribution.

We then perform two-sided student-t tests to test the hypotheses whether the fake network shares similar network characteristic matrices with a small-world network. We first simulated 100 random networks with the same numbers of nodes and links as the fake news network does, which gave the mean sample Average Clustering Coefficient (ACC) and Characteristic Path Length (CPL). We tested the following two hypotheses:



Hypothesis 1. Fake news network ACC = mean sample ACC (observed fake-news network is random network) vs. Fake news network ACC > mean sample ACC (observed fake-news network is a small-world network).

Hypothesis 2. Fake news network CPL = sample mean CPL (observed fake-news network is random network) vs. Fake news network CPL = sample mean CPL (observed fake-news network is a small-world network).

Both tests output p-values close to zero, suggesting that the fake news network is indeed a small-world network.

Data Transformation and Feature Engineering

The raw dataset has 2,923 rows of news reported to Politifact.com, with each row representing a piece of news record including its origin, currently associated website and authenticity. Our data transformation and feature extraction from this original dataset is completed in two parts, network feature extraction, and news-based feature extraction.

In network feature extraction, we transformed the raw dataset into an initial *directed* network dataset of 1452 rows and 5 columns:

- Source: The website where a piece of news originally comes from.
- Destination: The website where a piece of news lands on. Each (Source, destination) pair represents a path that a news record might travel through.
- Volume: The total number of news records that has travelled through the above-mentioned path.
- True: The total number of *real* news records that has travelled through the above-mentioned path.
- Fake: The total number of *fake* news records that has travelled through the above-mentioned path. Fake + True = Volume.
- Label: 1 if at least one fake news exists on this path between source and destination pair, and 0 otherwise. This variable will be used as response variable in the classification model.

Here we identify the websites as nodes, (source, destination) tuples as links and the number of news records traveling in each link as its corresponding value. There are 905 distinct nodes (websites) in the total network. Among the total number of paths existing in the network, 1,364 paths are labelled as 'fake' (because fake news has travelled through them) and 88 are labelled as 'real' (because fake news has never travelled through them).

Using this initial directed dataset, we then generated the following network-based features for each source and destination pair:

- Indegree centrality/outdegree centrality in the fake news network for source website: number of news shared to or from the source website.
- Indegree centrality/outdegree centrality in the fake news network for destination website: number of news shared to or from the destination website.
- Betweenness centrality: centrality measure of nodes in the fake news network based on shortest paths.
- Eigen-centrality - Measure of the influence of a node in the fake news network. It assigns relative scores to all websites in the network based on the concept that connections to high-scoring websites (nodes) contribute more to the score of the node in question than equal connections to low-scoring nodes.
- Mutuality: in the full network if there is a link from node A to node B and vice versa, then we created a categorical variable which indicates 1 if the link exist and 0 if it does not.
- Common Neighbours: number of common websites node A and B are connected to, regardless of the connection type (as source or destination).
- Common Neighbours in source - number of common 'sources' node A and node B have in the fake news network.
- Common Neighbours in destination - number of common 'destination' node A and node B have in the fake news network.
- Jaccard coefficient for connection - Set $\{A\}$ to be the set of websites connected to A (regardless of connection type) in the fake news network and similarly $\{B\}$ for B. Then Jaccard Coefficient for connection would be: $\frac{|\{A\} \cap \{B\}|}{|\{A\} \cup \{B\}|}$
- Jaccard coefficient for source - $\frac{|\{A\} \cap \{B\}|}{|\{A\} \cup \{B\}|}$ where $\{A\}$ refers to all the 'sources' A have and similarly $\{B\}$ for B.
- Jaccard coefficient for destination - $\frac{|\{A\} \cap \{B\}|}{|\{A\} \cup \{B\}|}$ where $\{A\}$ refers to all the 'destinations' A have and similarly $\{B\}$ for B.

We also extracted linguistic features and temporal features from the news in the network. For linguistic features, we used 'nltk' library in Python to generate the Median Polarity Score, which is the median of all the sentiment polarity scores calculated for all the claims (cited news title) traveled through the path between each source and destination pair.

In addition, we also aggregated the available tags from the raw data. As a result, we have the following variables:

- Health Care Tags: the number of news with 'Healthcare' tag on the path between each source and destination pair.
- History Tags: the number of news with 'History' tag on the path between each source and destination pair.
- Elections: the number of news with 'Elections' tag on the path between each source and destination pair.
- Military: the number of news with 'Military' tag on the path between each source and destination pair.
- Religion: the number of news with 'Religion' tag on the path between each source and destination pair.

For temporal features, we aggregated the following features:

- Day of the week: seven variable "Monday", "Tuesday", ..., "Sunday" specifying number of news posted on each day of the week for each source and destination pair.

Descriptive Statistics

Here we list some descriptive statistics relevant to our data, which help us gain intuitive understanding of the data we collected. Numbers of nodes and edges for the fake news and true news networks are shown below.

	Number of Nodes	Number of Edges	Network Density	Network Degree Centrality	Network Betweenness Centrality	Network Closeness Centrality	Network Eigen Centrality
Fake News Network	856	1364	0.001863694	0.08888342	0.03434465	0.0008192955	0.9766616
True News Network	88	99	0.01293103	0.1671291	0.007437006	0.01044297	0.9440524
Facebook News Network	175	177	0.005812808	0.5085877	0.0498638	0.1145333	0.9327255

Truth and rumor network entangled, and that's what makes the prediction and optimization works complex but important. To check how the existing path in the two networks resembles with each other, we perform **Quadratic Assignment Procedure**(QAP) with 100 runs of simulation. As shown below, all the 100 simulations yielded smaller correlations of paths than those in real observations, which suggests paths between the two networks are significantly correlated.

```
> qapoutput =
qaptest(list(g_fake_binary_net,g_real_binary_net),gcor,g1=1,g2=2, reps=100)
> qapoutput$testval
[1] 1
> summary(qapoutput)
```

QAP Test Results

```
Estimated p-values:
p(f(perm) >= f(d)): 0
p(f(perm) <= f(d)): 1
```

Limitations of Data

The data is extracted from a third party news-investigation website with information contributed by online volunteer fact-checkers, resulting in potential selection bias and imbalanced weights between fake and true news. The total number of '0's in the label variable (binary response variable for classification) consists of only 6.7% in the entire dataset, but upsampling for true news portion should be unnecessary here since Sci-kit learn implementation of machine learning models such as random forest will take care of this imbalanced data problem. Furthermore, if we only take the subset where Facebook is either source or destination, the total number of 0's jumps up to 13%, which should be sufficient for classification model training.

FEATURE SELECTION

Feature Selection for Classification Model:

The classification model is aimed at predicting probability of fake news existence in between a source and destination pair.

- Response variable: 'Label' of 1 and 0 for the prior existence of fake news between a source and destination pair.
- Predictors: They include network-related features and link-related features including linguistic variables and temporal variables.
 - Selected network-related features include: source website indegree/outdegree centrality in the fake news network, destination website indegree/outdegree centrality in the fake news network, mutuality score of source and destination pair, number of common neighbors, number of common destinations, number of common sources, jaccard coefficients of common neighbors (sources and destinations), jaccard coefficient of common destinations, jaccard coefficient of common sources.
 - Selected linguistic variable: Median Polarity Score of source and destination pairs.
 - Selected temporal variables are the seven variables "Monday" to "Sunday" specifying the number of news posted on each day of the week.
 - Aggregated news tags: Health Care Tags, History Tags, Elections, Military, Religion
 - Noted that we have excluded the linguistic variables recording number of news being tagged for topics in "Elections" "Military" etc. This is because the goal of our classification model is to predict the existence of fake news in a path, instead of the truthfulness/fakeness of a specific news, and thus we choose to include path-related variables instead of news-related variables.

Feature Selection for Optimization Model:

The optimization model is aimed at maximizing the control on flows of fake news in the news network while minimizing the impact on flows of true news, by tuning a hyperparameter used to calculate the loss function of fake news confined by a realistic constraint.

The input features are source and destination pair, volume or total number of news traveling from source to destination, number of fake news traveling from source to destination, number of true news traveling from source to destination. The hyperparameter (λ) is the relative weight of a fake news compared to a true news. By randomly pruning a number of news links existing in the news network, we will be able to calculate the amount of fake news loss (calculated by weighted fake news removed from the network

- weighted true news removed from the network). The model, when pruning links, is confined by the constraint *maxLink*, which specifies the maximum number of links allowed to be pruned. We then trained the model to find the optimal pair of hyperparameter value and *maxLink* value, such that the specified model will work to eliminate the maximal amount of fake news flow while reasonably preserve majority of the true news flow.

MODELS

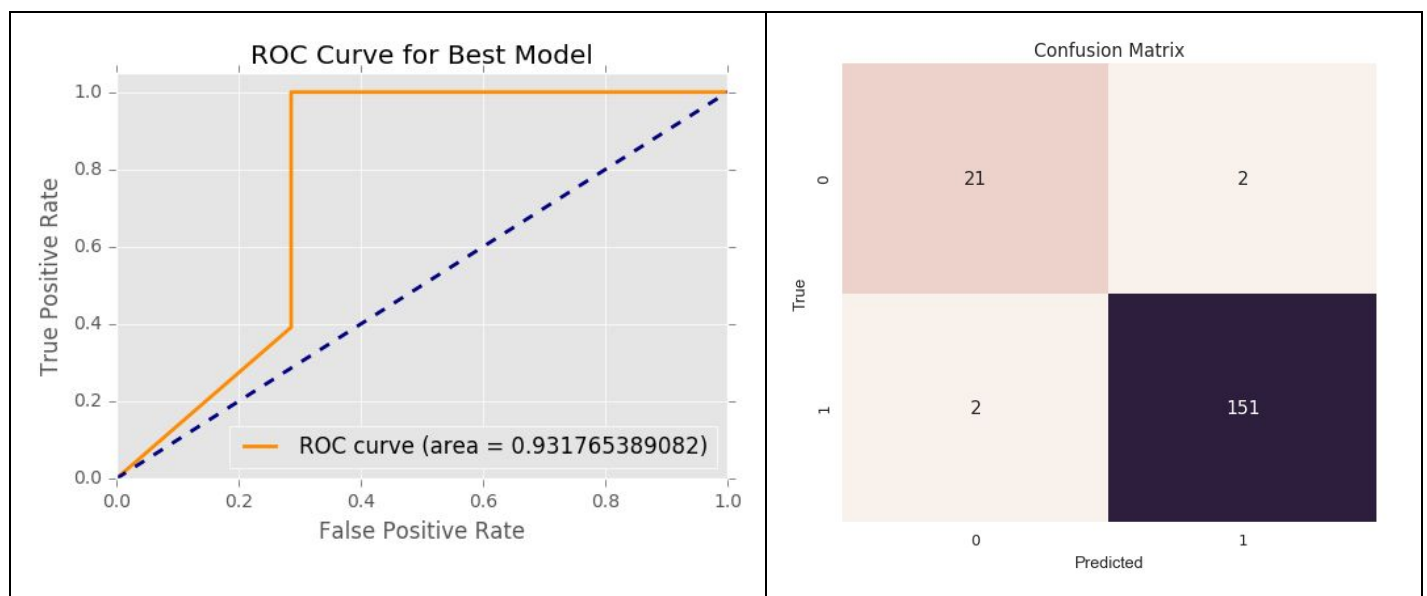
Classification Model

As previously mentioned, we took the subset data where Facebook is either source or destination. This leaves us with 176 rows, where 13% of the labels are 0. Thus we do not have to worry about class imbalance. We also noted that of the 176 rows, 94% of the cases have Facebook as the destination.

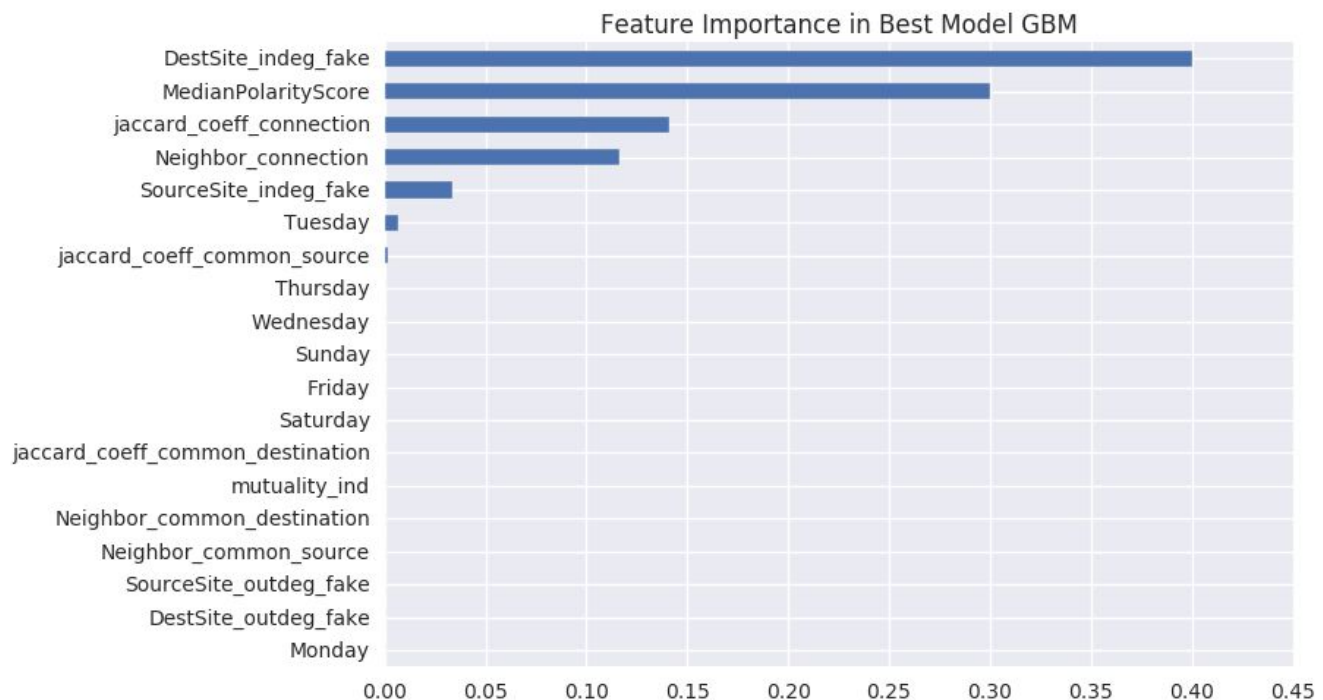
For classification algorithms, we used 5-fold cross validation to choose from the following models with corresponding sets of parameters. The following table also details the best AUC score for that given model.

Model	Parameters	Best AUC Score
Random Forest	Number of trees: [30,50], by steps of 5 Maximum depth of tree: [2,5]	0.912267
Gradient Boosted Machine (GBM)	Number of trees: [30,50], by steps of 5 Maximum depth of tree: [2,5] Learning rate: 10 values with equal intervals in the range of [0.1,0.3]	0.931765
Logistic Regression	Penalty (Regularization): L1, L2 C (Inverse of regularization strength): 5 equally spaced values in the range of [0.1,1]	0.753361

According to model output, the best model is GBM with 30 estimators, tree max depth of 4 and a learning rate of 0.29. Please find below the confusion matrix and ROC curve.



Please also find below a bar chart showing feature importances.



As we can see, the most important feature to detect fake news is the destination website indegree for fake news. This indicates that the number of news that points to the destination website is the best predictor for the fakeness of the news. The second most important feature is the median of polarity score for all the news in a certain source-destination path. This is very intuitive because fake news almost always has an emotive intention, which would result in lack of objectiveness. The third most important feature is the Jaccard coefficient (in terms of generation connection) for the specified path. The Jaccard coefficient of two nodes can be defined as the number of common connections between two nodes divided by the sum of number of connections of two nodes. Therefore, Jaccard coefficient is a similarity measure of two nodes in a network. The presence of Jaccard coefficient as the third most important features for our predictive model indicates that similar nodes are more likely to post fake news to each other, although we need to test this hypothesis to determine if this statement is true.

Optimization for Facebook News Network:

The past few years have seen the rise of the trend that more people use Facebook as news source (which is also what Facebook is expecting and pushing forward). Unfortunately, this comes with the downside that Facebook's also spread rumors and has even been deemed and denounced as hub of rumors recently. One of the big question here is how could Facebook keep its influence as a news sources but curb the spreading of rumors from Facebook. To put it using social network analysis

terminologies, we hope that, by tuning the outgoing links Facebook sustained in the network, we could control the centrality of Facebook in rumor networks, but keep its centrality in truth network intact.

Here we formulate the optimization framework. First, by saying optimizing centrality, we chose to optimize out-degree centrality for the following reasons: (1) compared with other metrics (for instance, eigen-value centrality) it's much easier to compute; (2) Since we'll never remove a website from the network (we'll be optimizing at link level), the number of nodes will be fixed and optimizing degree centrality degenerates to optimizing the exact news flow in the network.

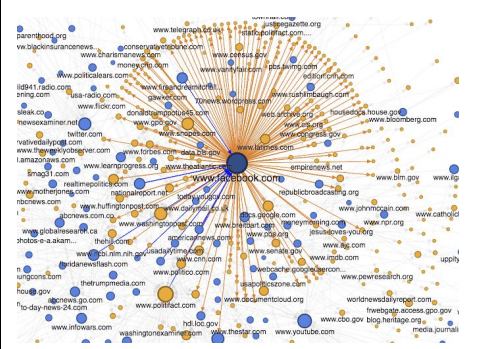
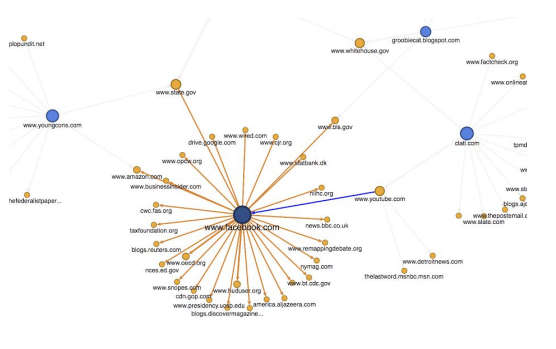
Besides, we imposed a constraint on the number of links we can play with, because it's not affordable for Facebook to monitor to all the websites that reposted their news or articles, or take actions immediately and simultaneously. For this constraint, maxLink, the values we used in our test were 5 (which is very limited), 10, 20, and 50 (about $\frac{1}{3}$ of all the outgoing links of Facebook).

Most importantly, in this multiobjective framework, relative weight between truth and rumor should be decided beforehand as a hyperparameter (lambda in our case).

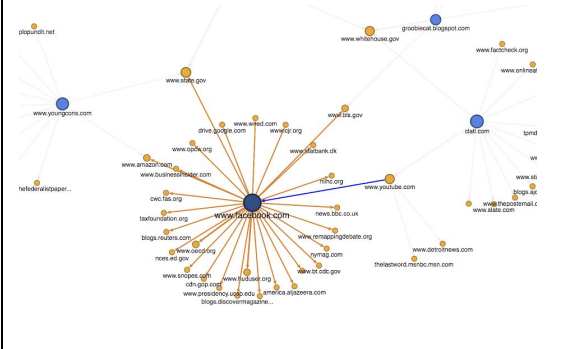
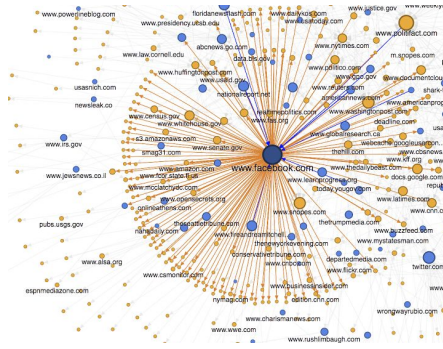
$$\text{maximize } \lambda \cdot \text{RumorOutDegree} - \text{TruthOutDegree}$$

Results of Optimization Model

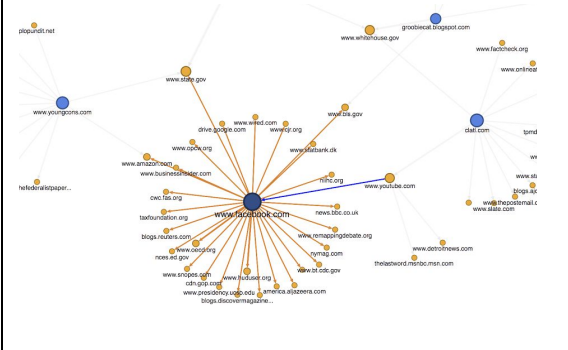
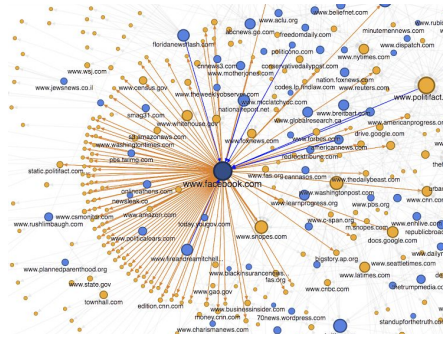
Attached below are network charts and performance metrics of our model. It can be identified that: (1) (links colored in orange are outgoing links we care about) by increasing the number of maxLink (but set Lambda low), we managed to keep truth network intact while we took down more than 50% of rumors (though we should keep in mind that the data set is not a balanced sample); (2) when we set Lambda larger, the optimization program will start pruning major rumor paths in rumor networks while the truth network wouldn't suffer too much; (3) betweenness and eigen-centrality has been improved for rumor network.

parameters	rumor	truth
maxLink:5 Lambda:0.02		

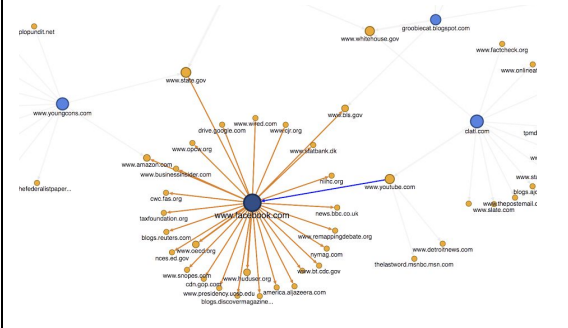
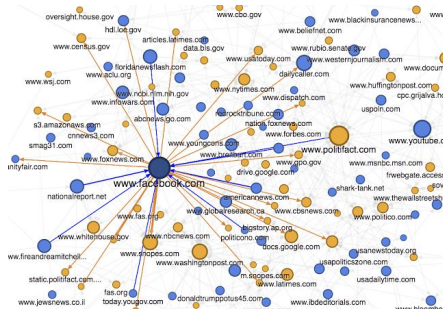
maxLink:10
Lambda:0.02



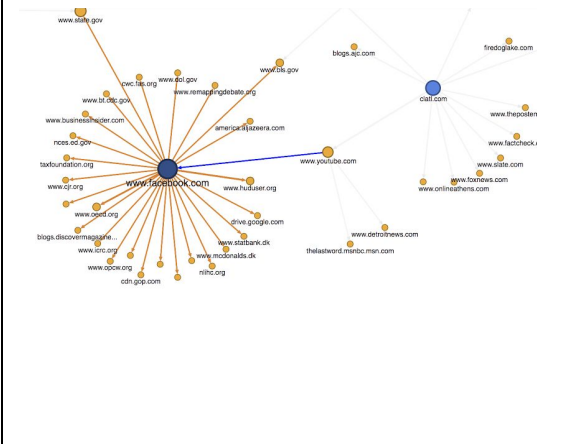
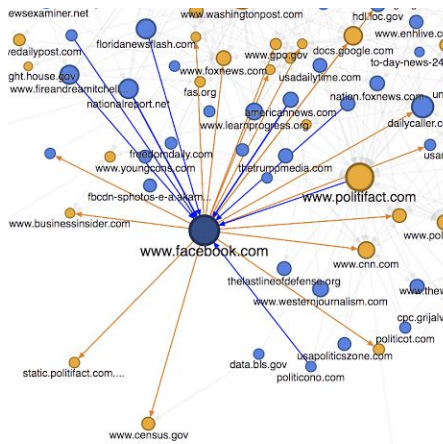
maxLink:50
Lambda:0.02



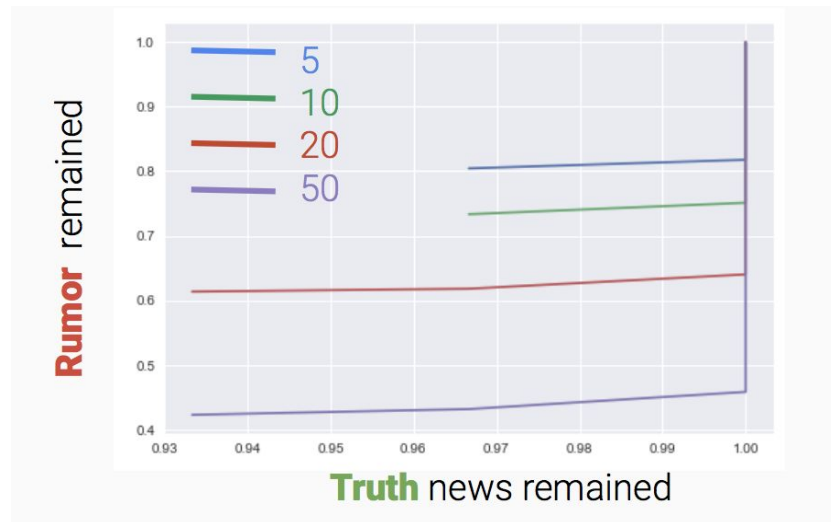
maxLink:50
Lambda:0.02
(showing only top
200 nodes)



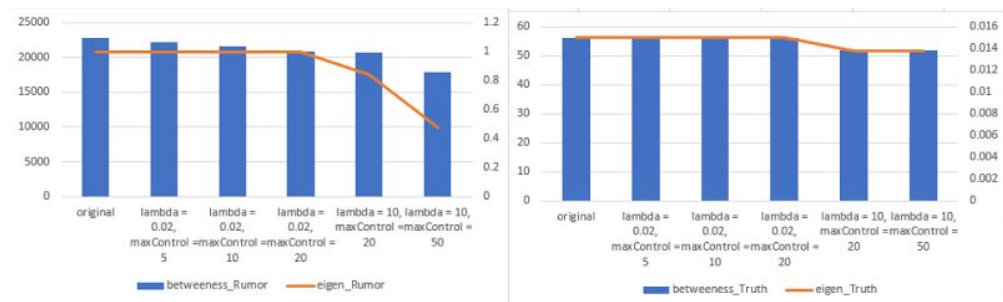
maxLink:50
Lambda:10
(showing only top
200 nodes)



Rumors and truth remained when using different maxLink number (Facebook as source)



Betweenness and eigen-centrality are improved at the same time.



APPLICATIONS

Confidence Score Feature

One possible application for our predictive model is the confidence score feature on Facebook. Our classification model attempts to predict the probability that a piece of news coming from or arriving at Facebook is fake. Facebook can interpret this probability as a confidence score, with the scale of 0 to 1, on how confidently Facebook can assume this news is fake. For example, if a piece of news from Facebook is posted on whitehouse.com, then our predictive model will enable Facebook to report that the confidence score is 96, meaning that this piece of news is very trustable. This confidence score feature will enable Facebook to provide us with the chance to decide by ourselves whether we want to trust this piece of news after we see the confidence score. This way we can improve Facebook's credibility as a news source.

Removal of major fake news paths with optimization model

In sense of centrality, our multiobjective optimization framework successfully brought down centrality of Facebook in rumor network while keeping its centrality in truth network. In reality, this framework can be applied in the following approaches: (1) as it more and more presents itself as a news source, Facebook could officially establish collaboration ties with only a few selected media/websites, so by differentiating its ties to different websites Facebook can navigate its users to more reliable or authenticate paths/websites that our optimization framework suggests, (2) monitor the paths that our framework suggests to cut off, because most rumors are debunked only after they caused enough hype and has already dragged Facebook's name through mud, and checking the paths that tend to share fake news and responding immediately (removing the rumor from Facebook, warning the website to stop intentionally repost sensational fake news, etc.) would help curbing rumors' circulations before they go viral.

CONCLUSION

The client for our project can be any social network site which wants to understand the flow of true and fake news cited from or arriving at their website. For this particular project we focused mainly on politics related data from kaggle, and we focus on all the news with facebook.com as the source. Therefore, Facebook is our project for this particular application. The data was collected from Politifact website.

We tried to achieve two things in this project. First, establish the influence of Facebook in truth network and truth network only which requires multi-objective linear programming model. By running out-degree optimization, we demonstrated that Betweenness and Eigenvector Centrality can be optimized at the same time on rumor network while the metrics of the truth network were not affected too much. Besides, rumors flowing out of Facebook can be effectively controlled by gating just a limited amount of certain major rumor paths.

In the Second part of the project we wanted to find how confident the news is fake when Facebook is a source or a destination. The same model can also be applied to other websites. From the classification model we can conclude that most important features for this confidence source would be the destination website indegree for fake news, median of polarity score for a certain path(source-destination pair), and the Jaccard coefficient (in terms of generation connection) for the specified path.

FUTURE WORKS

Linguistic features

In this project, we utilized median polarity score for all the news flow for a given source-destination path because we assume that news descriptions that are too positive or too negative in tone indicate higher probability of being fake. As a future suggestion, we can add other linguistic features of the news description as the attributes of a link for our predictive models. Below list some candidates for some more linguistic features:

1. Sentiment score: similar to polarity score, but more directly involved with the emotions embedded by the news description.
2. Topic modeling: We can run topic modeling algorithms such as Latent Dirichlet Allocation (LDA) as an alternative for the tags in the news.

Dynamic network features

The fake news network we are working with is static. By making the network static, we are making an assumption that the nature of destination website does not change. That is, the the real/fake news ratio is constant, and the destination website does not change their policy regarding the fake news at all. But this assumption cannot hold for reality. Therefore, we want to incorporate some dynamic network features to our model so that the predictive model will adjust to the change of destination website.

Selection bias in dataset

Our dataset from polifact.com consists mostly of fake news, with only a small percentage of real news. However, in the real world, the real/fake news ratio is different from the ratio reflected by the dataset. Therefore, the classification model will learn the real/fake news ratio from our dataset, instead of the actual real/fake news. This will result in the inaccurate probability of fake news and affect the confidence score accordingly. To resolve this problem, we need a more fair and representative dataset to reflect the fake/real news ratio in reality.