
fakebook
Josh, Lauren, Chaoran, Meghan, Roshan

Lets kick off with a fake news/gif !!!


ellentube
NBC NEWS LIVE
THE PRESIDENTIAL DEBATE

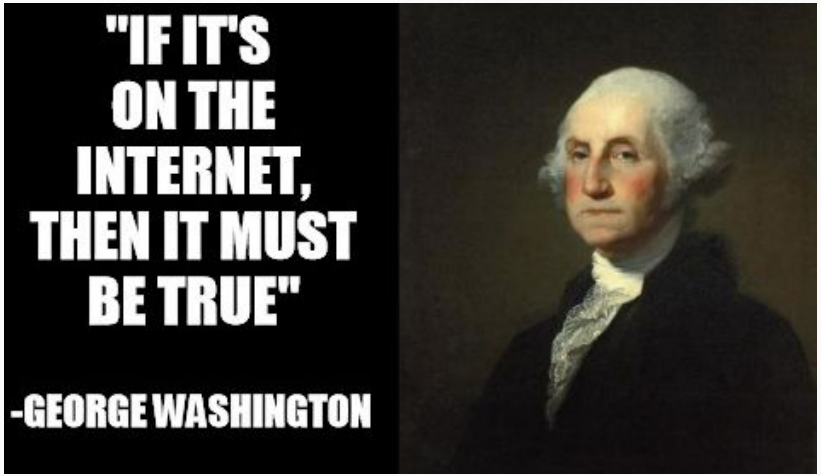
HOW TO SPOT FAKE NEWS.....

Agenda

- Motivation
- Introduction
- Data Overview
- Visualization
- Model Discussion
- Result & Conclusion

Motivation

- Tsunami of fake news on the social media
- Influence people's judgement and decisions
- Unveil the truth
- Predict the flow of fake news



Introduction

kaggle

Who starts and who debunks rumour

POLITIFACT

WINNER OF THE PULITZER PRIZE

Fact checker related to politics

- **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
- **Total 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

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

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

claim_label

Fake quote: Tom Price did not tell cancer patients to die more quickly

page_is_first_citation = true

By Joshua Gillin on Tuesday, March 21st, 2017 at 12:57 p.m.

page_is_first_citation = false

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

About this statement:

date_published

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

Researched by: Joshua Gillin researched_by

Edited by: Aaron Sharockman edited_by

Subjects: Fake news tags

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

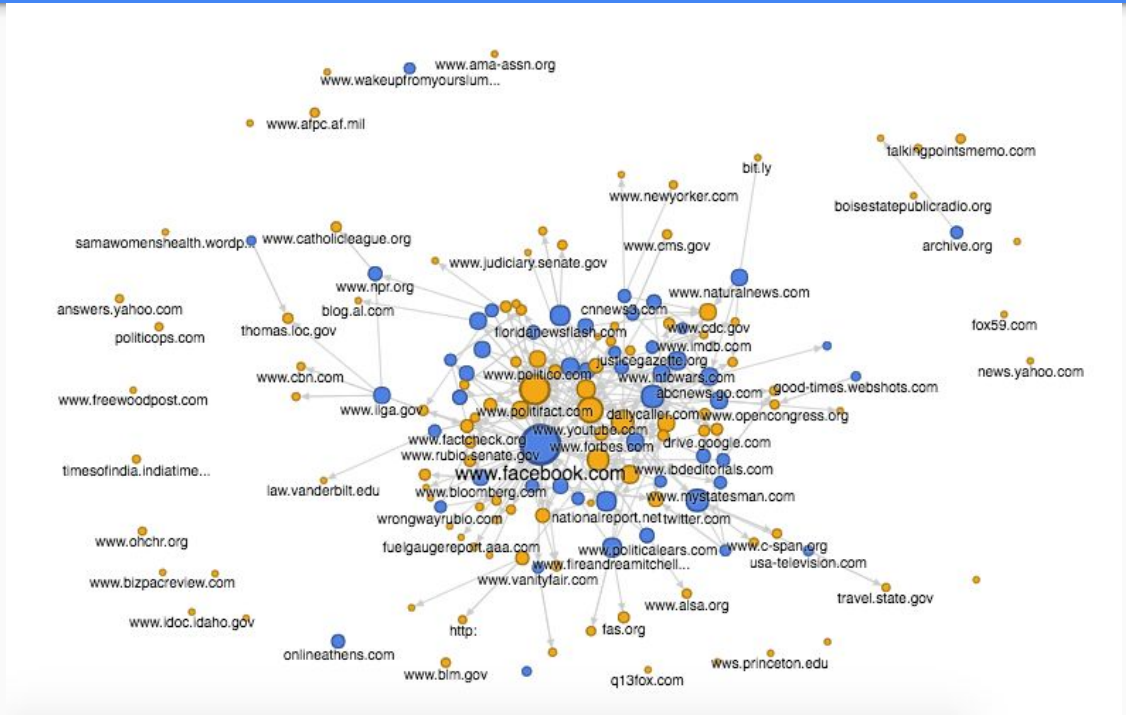
CBS News, "Alan Grayson 'Die Quickly' Comment Prompts Uproar," Sept. 30, 2009

Data Snapshot

	A	B	C	D	E
1	**Source**	**website**	**page_url**	**TRUE**	**FALSE**
2	www.facebook.com	www.facebook.com	42	5	37
3	www.facebook.com	www.politifact.com	22	0	22
4	nationalreport.net	www.whitehouse.gov	14	0	14
5	www.youtube.com	www.youtube.com	9	2	7
6	www.politifact.com	www.politifact.com	8	0	8
7	docs.google.com	docs.google.com	8	0	8
8	www.naturalnews.com	www.cdc.gov	8	0	8

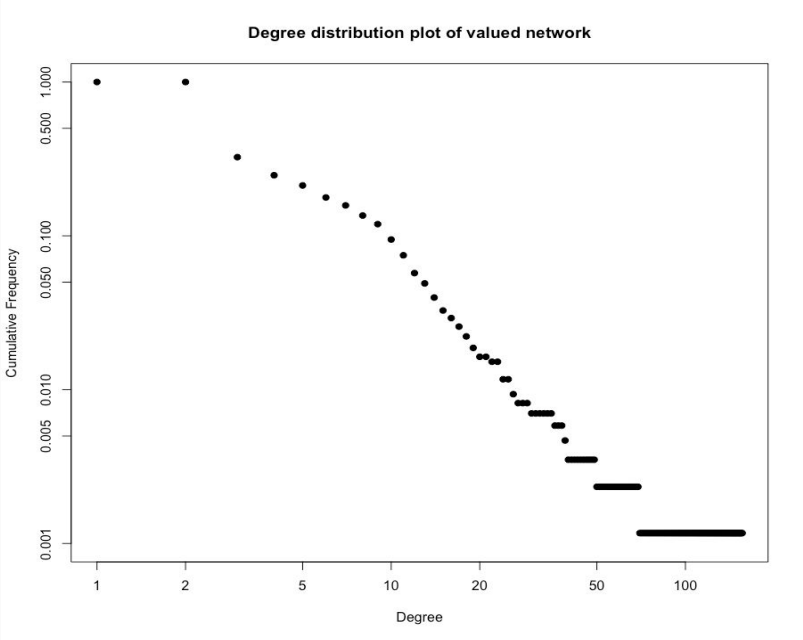
Network Visualization

- **Blue nodes:** fake news sender
- **Yellow nodes:** fake news receiver



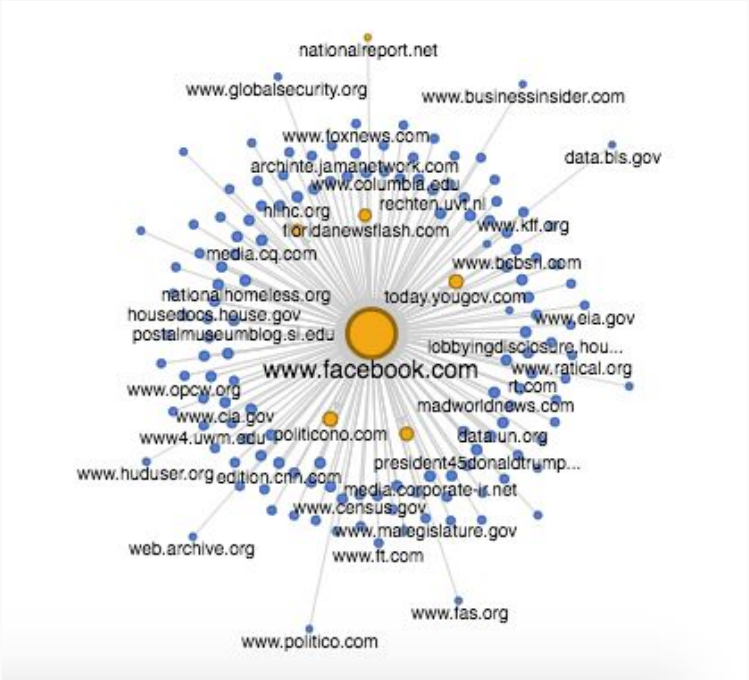
Global Attributes - FAKE News Network

- **No. nodes:** 856
- **No. links:** 1364
- **Network density:** 0.0019
- **Kolmogorov-Smirnov test p-value:** 0.87
- **ACC:** 0.07 (vs sample mean ACC 0.003)
- **CPL:** 4.02 (vs sample mean CPL 5.83)



Global Attributes - FACEBOOK News Network

- **No. nodes:** 175
- **No. links:** 177
- **Network density:** 0.0058
- **Kolmogorov-Smirnov test p-value:** 0.41
- **ACC:** 0.00
- **CPL:** 1.89



Attributes for Links

- **Day of week:** Monday, Tuesday, etc.
- **Most common tags:** elections, religion, economics, etc.
- **Polarity** of news description

Applications

Confidence Score

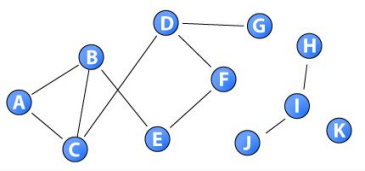
- Indicates how **confident** the news is fake when facebook is a source or a destination.
- This will increase the credibility of facebook.com as a news source.

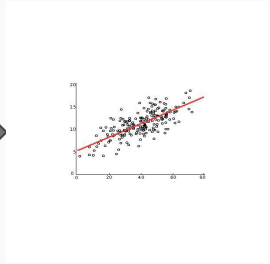
Optimization

- Establish influence of facebook in **truth** network and truth network **ONLY**
- Multi-objective linear programming model

Classification

	News Source	News Destination	authenticity
news1	abc.com	CNN	fake
news2	Fox news	White House	fake
news2	facebook.com	chaoranwei.com	TRUE





Data
- Source
- Destination
- Authenticity
- volume

Network topology
- Mutuality
- Centrality
- Jaccard Similarity

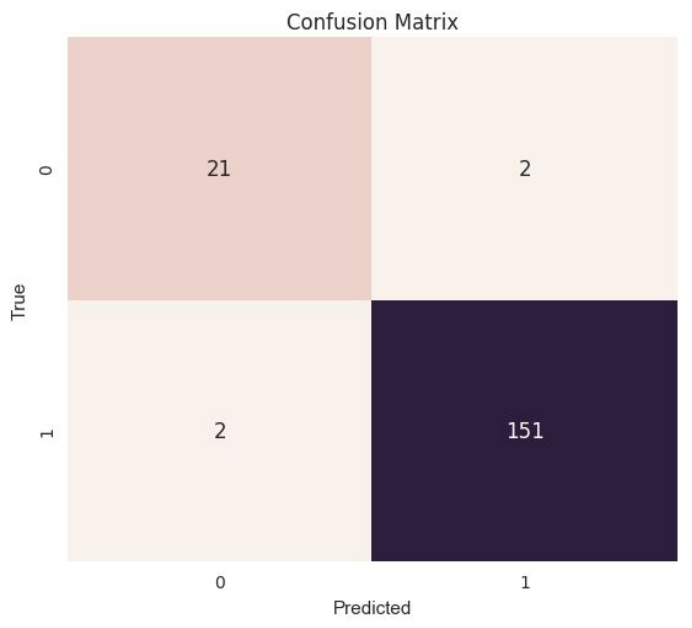
Non-network Features
- Tags
- Day of week
- Polarity score

Predictive Modeling
- Gradient Boosting
- Random Forest

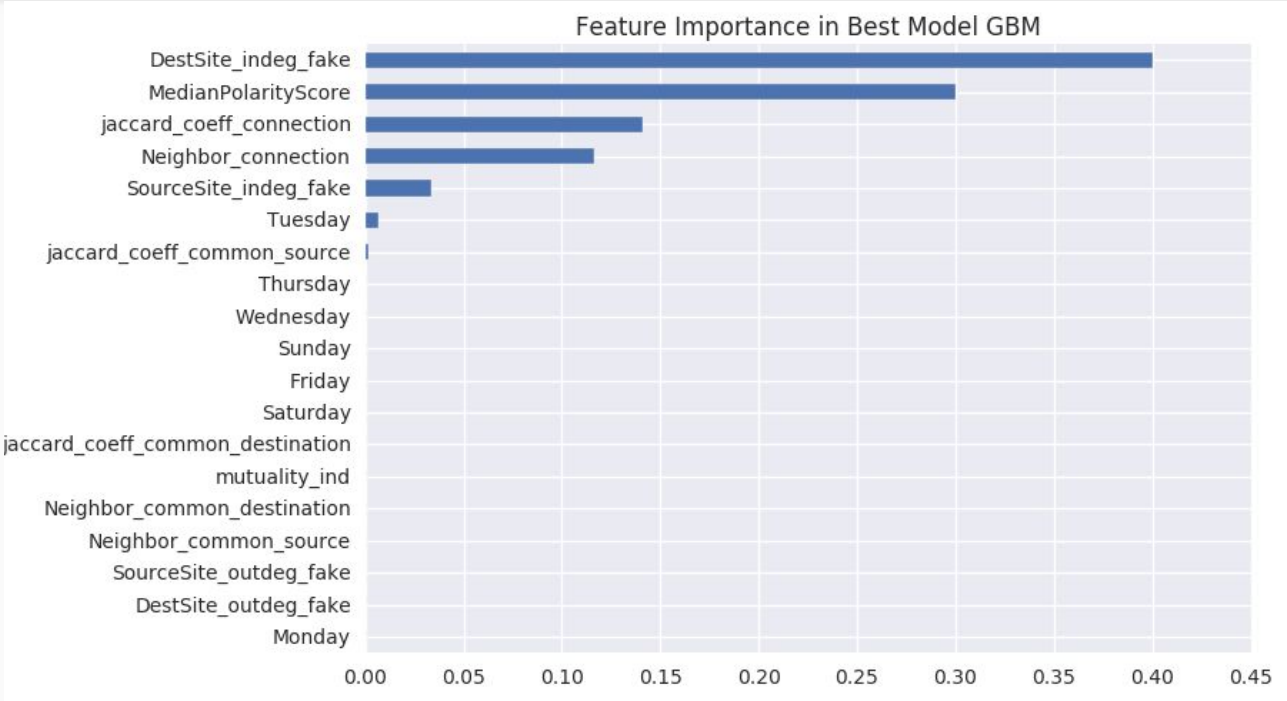
Cross validation
- AUC score

Model Output

- **Best model:** Gradient Boosted Tree (**AUC score:** 0.93)
- Output probability as the **confidence** score.



Feature Importance



Network Optimization Motivation

Supposedly, facebook is our client.

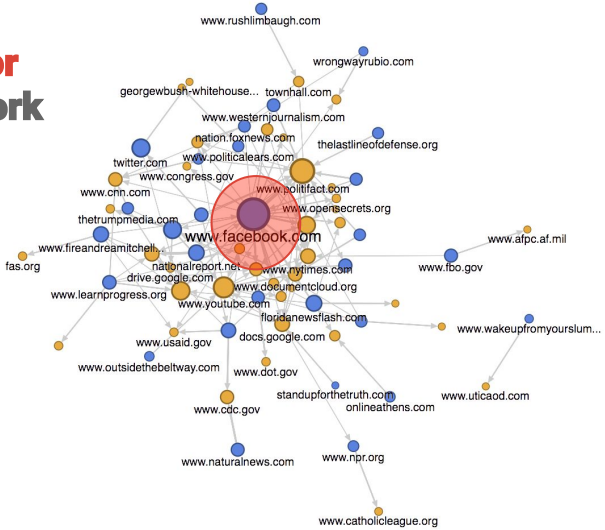
Hub in both rumor and truth network.

Q: Central ONLY in TRUTH network?

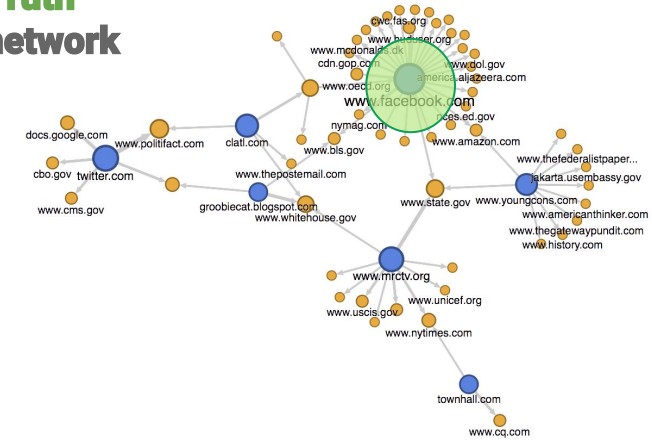
A: Optimizing degree centrality.

1. Computational advantage
2. Easy to formulate the problem

Rumor network



Truth network



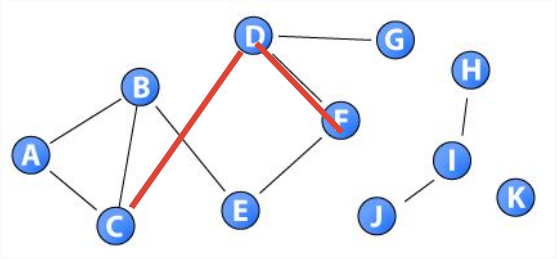
Network Optimization Formulation

Optimizing **Degree** Centrality

= Optimizing **Flow of News**

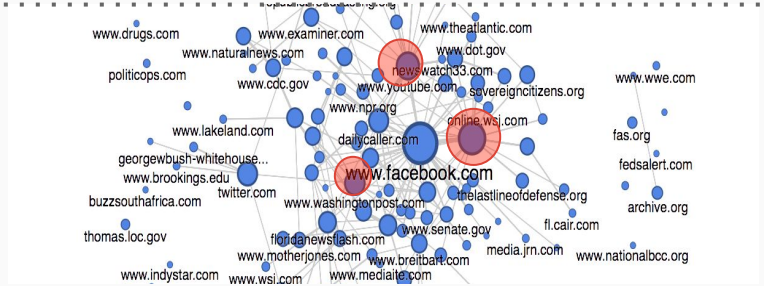
Assumptions on:

- Distribution of news flow
- Levels of control(optimization)
- Constraints on control
- Weights of truth and rumor in objective



LEVELS OF CONTROL

Gate links specifically, instead of removing a node entirely.



CONSTRAINTS

Cannot afford to censor/check **all the websites/news** at the same time

maximize

$$\lambda * \#RumorOutDegree - \#TruthOutDegree$$

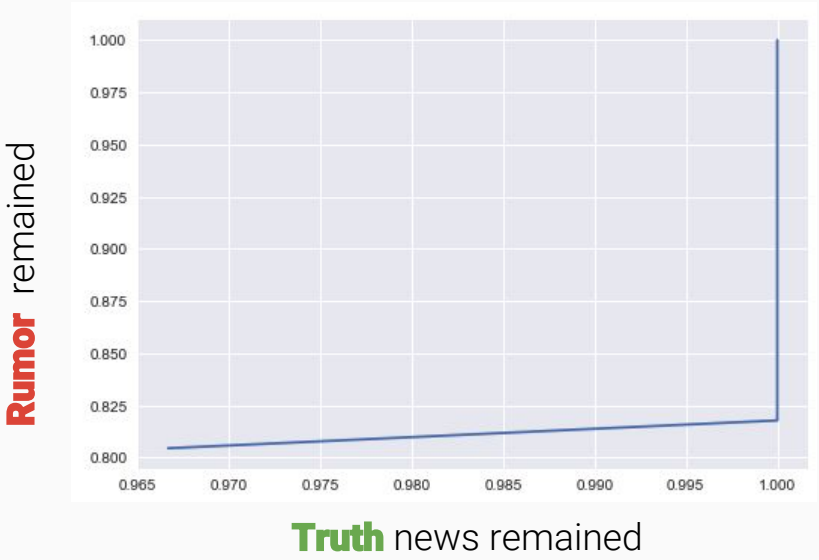
MULTIOBJECTIVE

Relative importance between debunking rumors and keeping truth flows

Network Optimization Step 1

Scenario: **Very limited control**

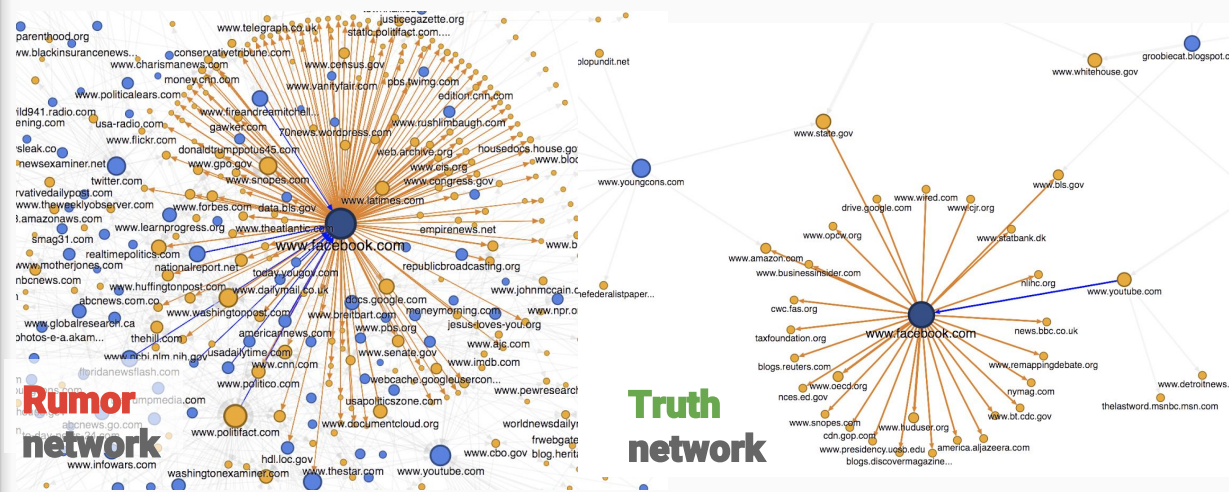
Set maxControl = 5
lambda : 0 to 10

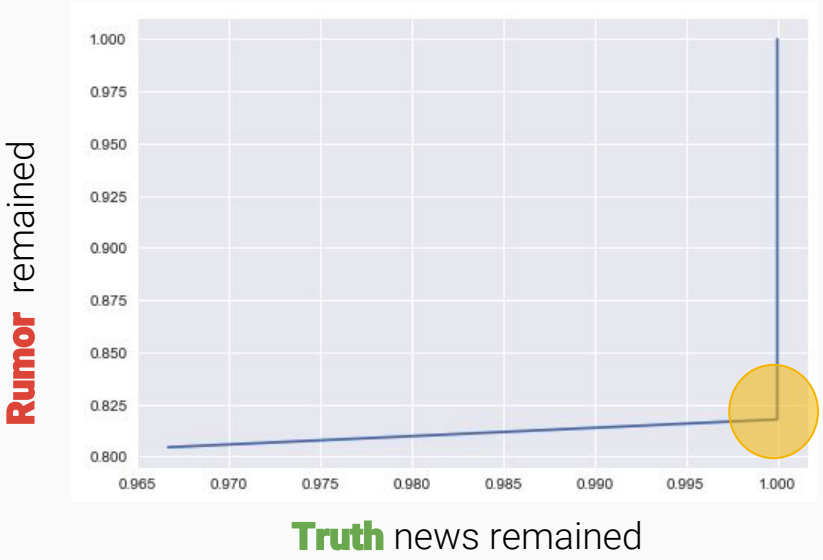


Network Optimization Step 1

Scenario: **Very limited control**

Set maxControl = 5
Set lambda =0.02, which means putting relatively high weight on **truth.**

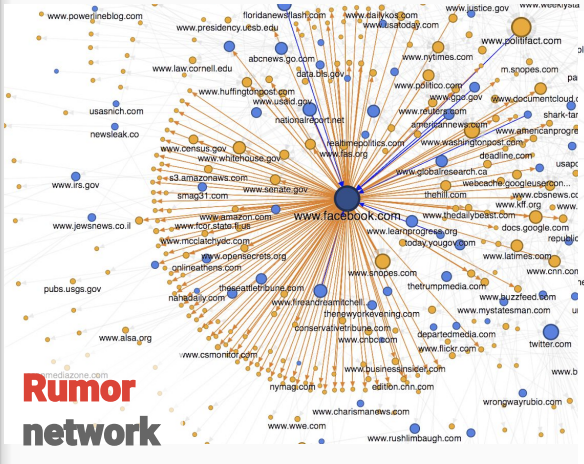


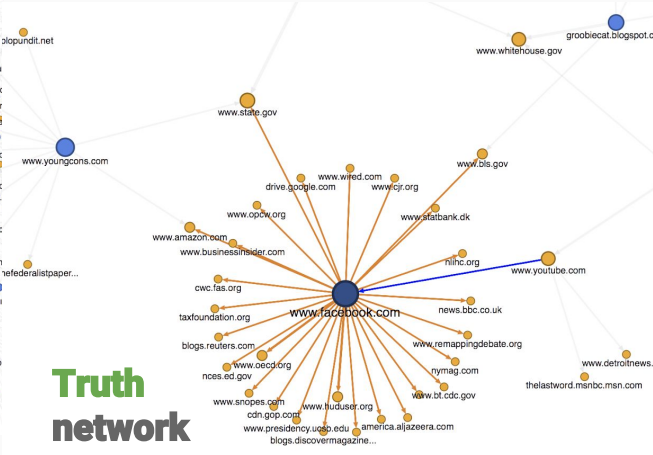


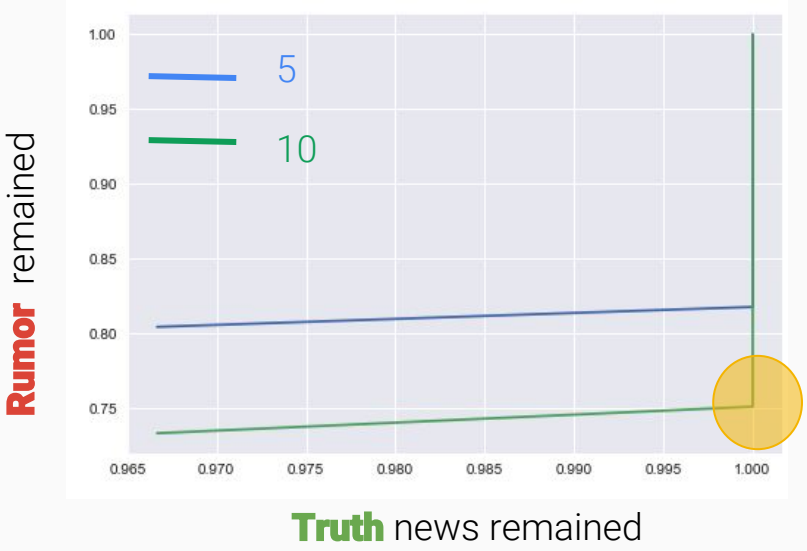
Network Optimization Step 2

Scenario: Compare 5 with 10

Set **maxControl = 10**





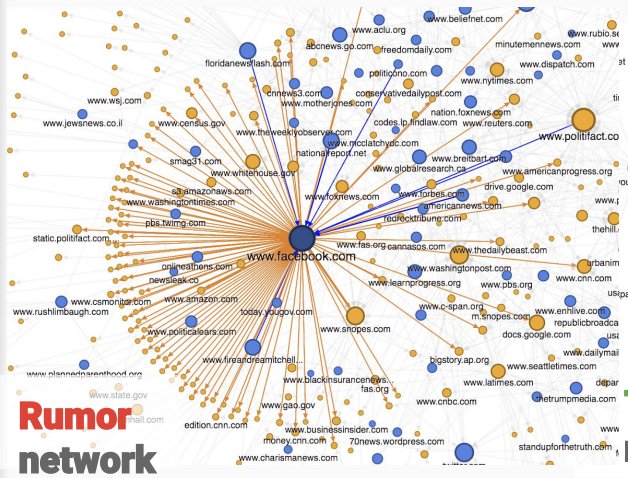


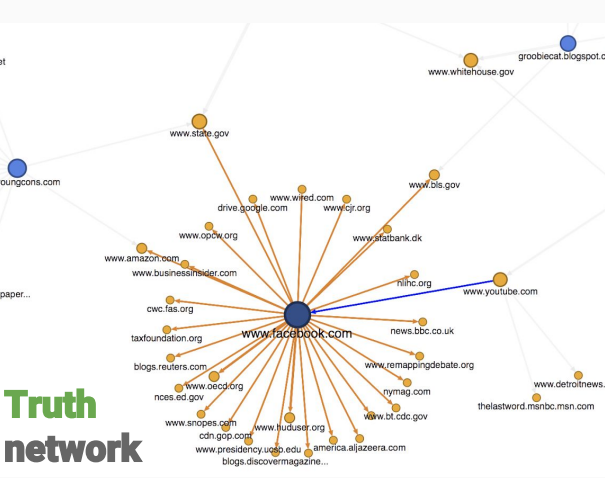
Network Optimization Step 3

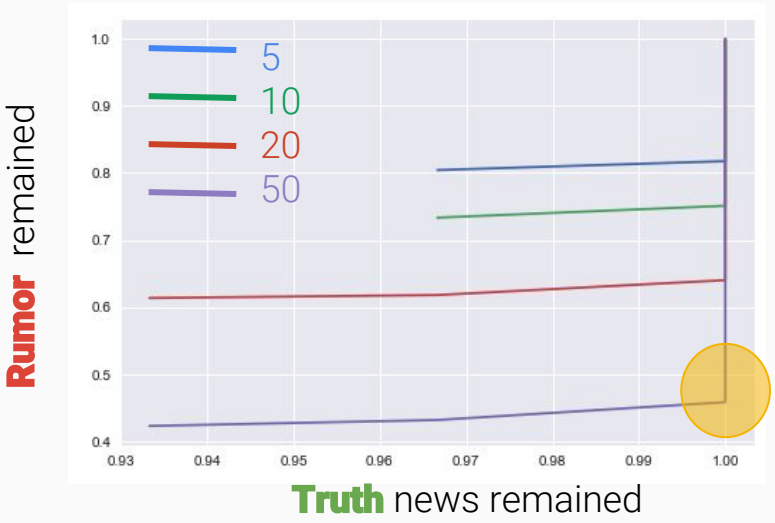
Scenario: test larger control threshold

Set maxControl = 20

Set maxControl = 50 (about ⅓ of all links)

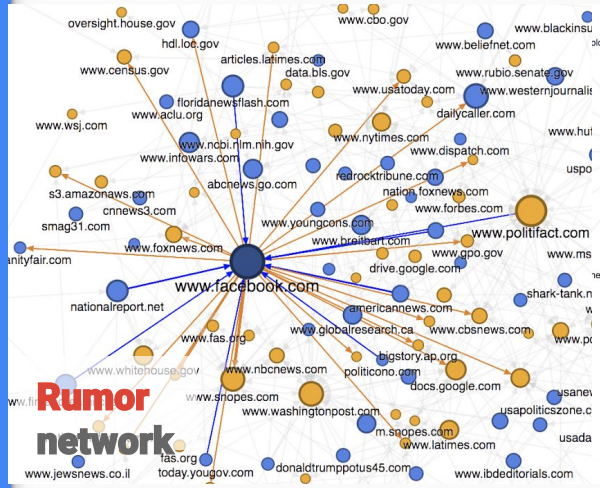


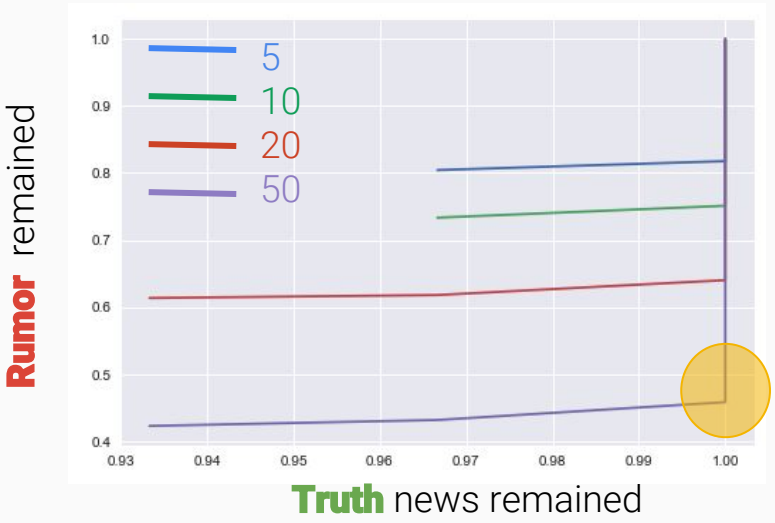




Network Optimization Step 4

Scenario: Draw only the **top 200 nodes** with highest weights.



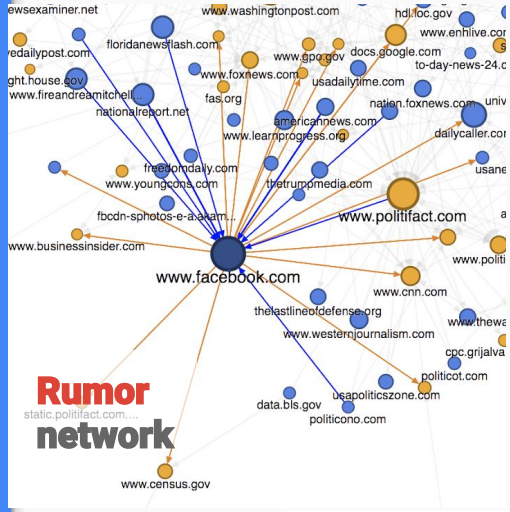


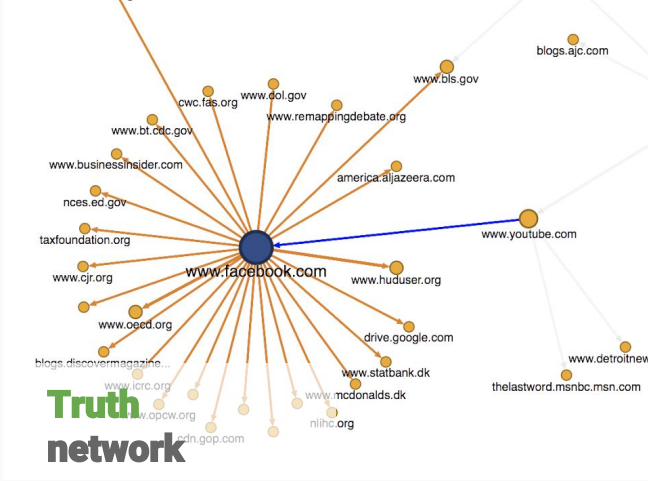
Network Optimization Step 4

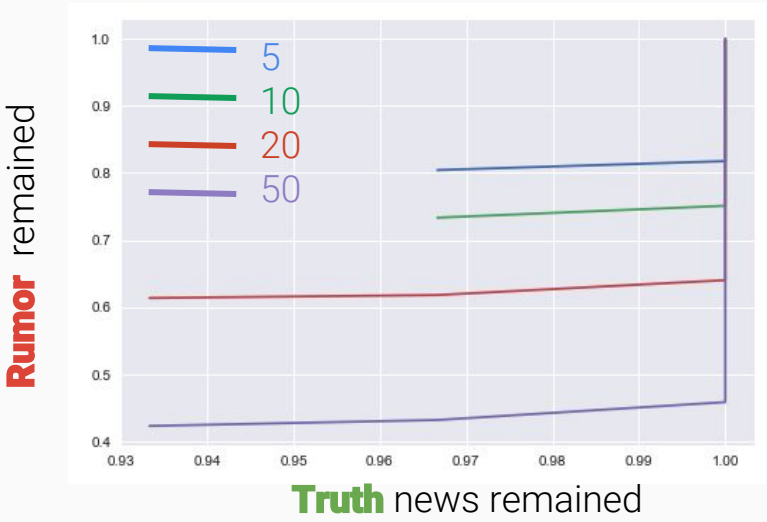
Scenario: Draw only the top 200 nodes with highest weights.

Then set lambda higher (0.02 -> 10).

Effectively **cut major rumor paths**, but won't do too much harm to circulation of truth.





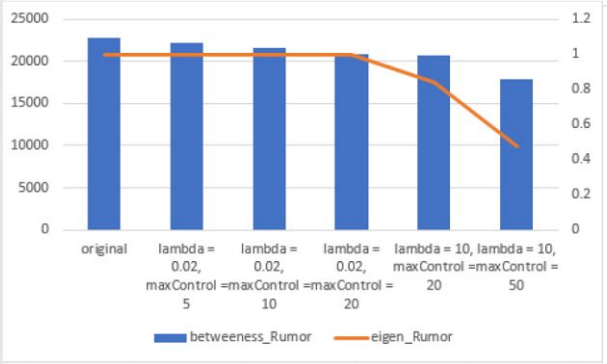


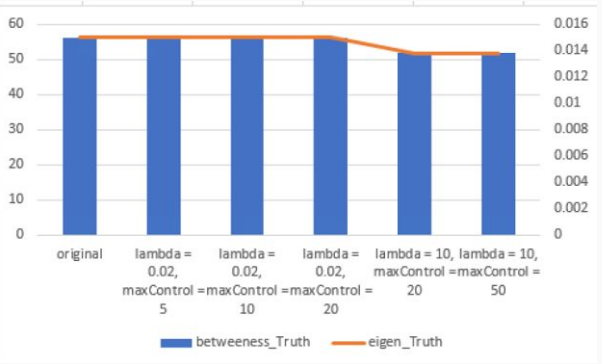
Network Optimization Optimized Centrality

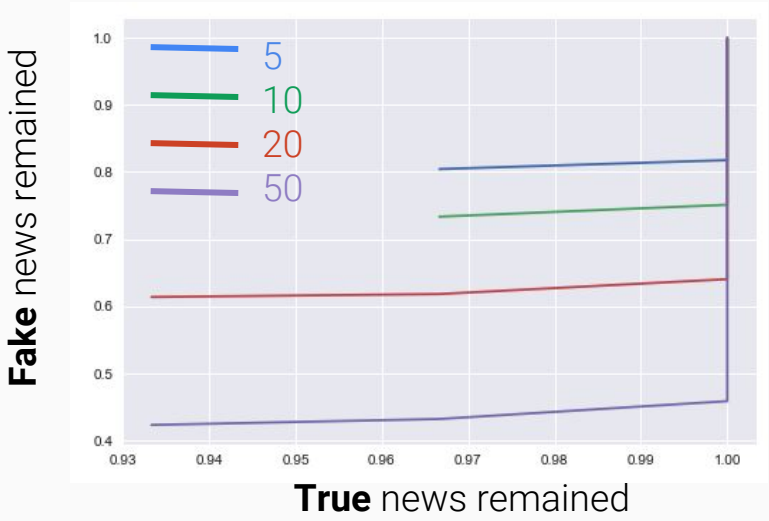
When running out-degree optimization, **Betweenness** and **Eigen Centrality** are optimized simultaneously.

Results:

In the sense of centrality, facebook is still being **central in truth network**, but **much less crucial** in the rumor networks.



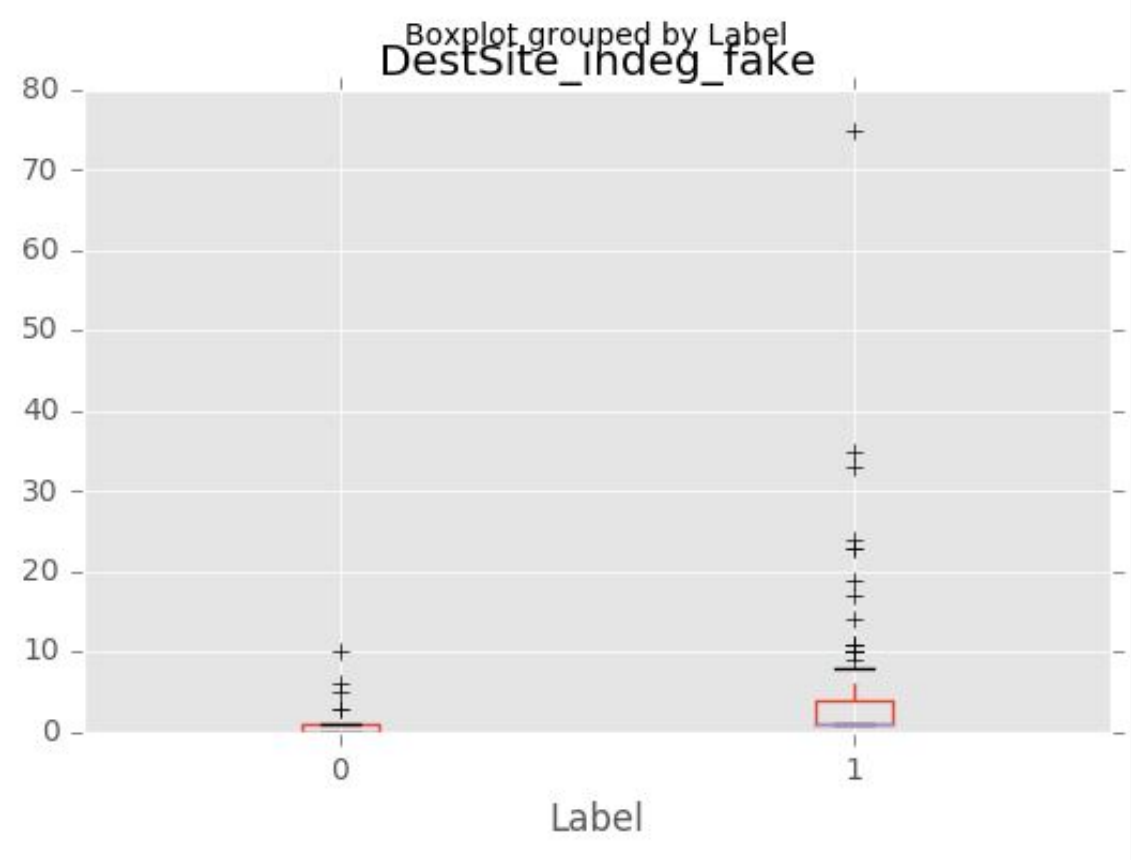




Results & Conclusion

Key features for classification

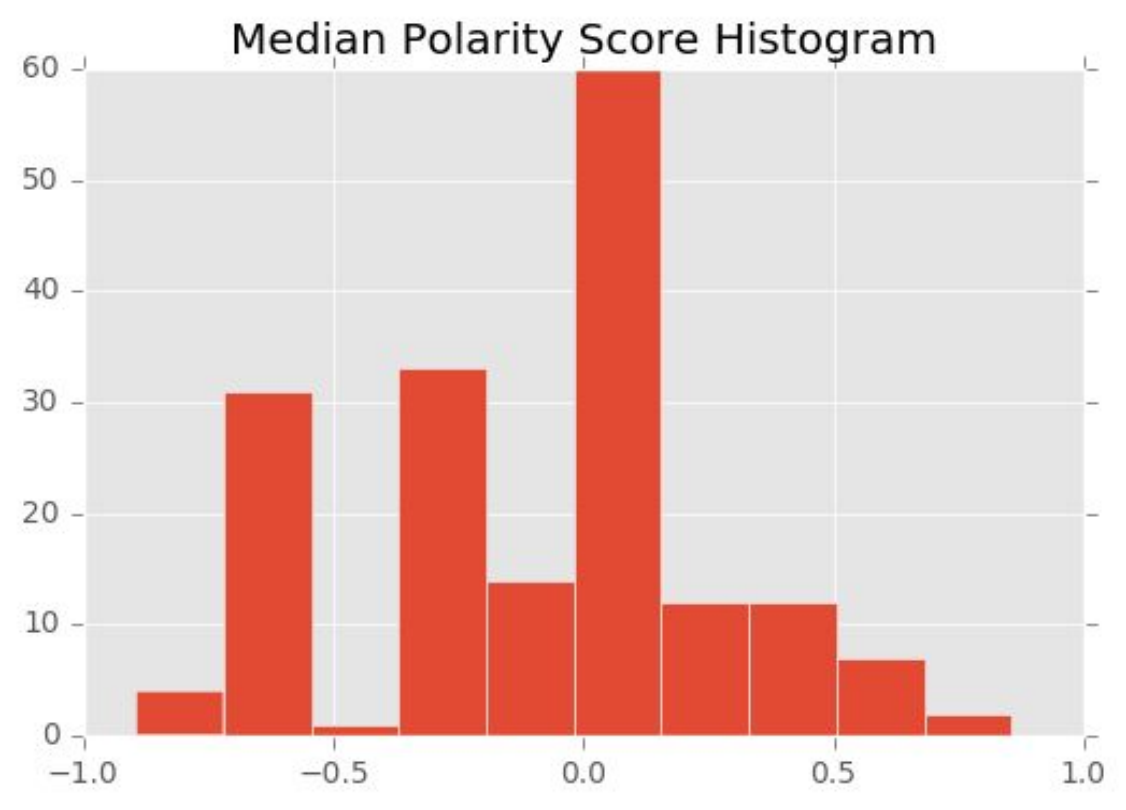
a. **Destination site's in-degree in fake network**
b. Median polarity score
c. Jaccard Coeff for Connection



Results & Conclusion

Key features for classification

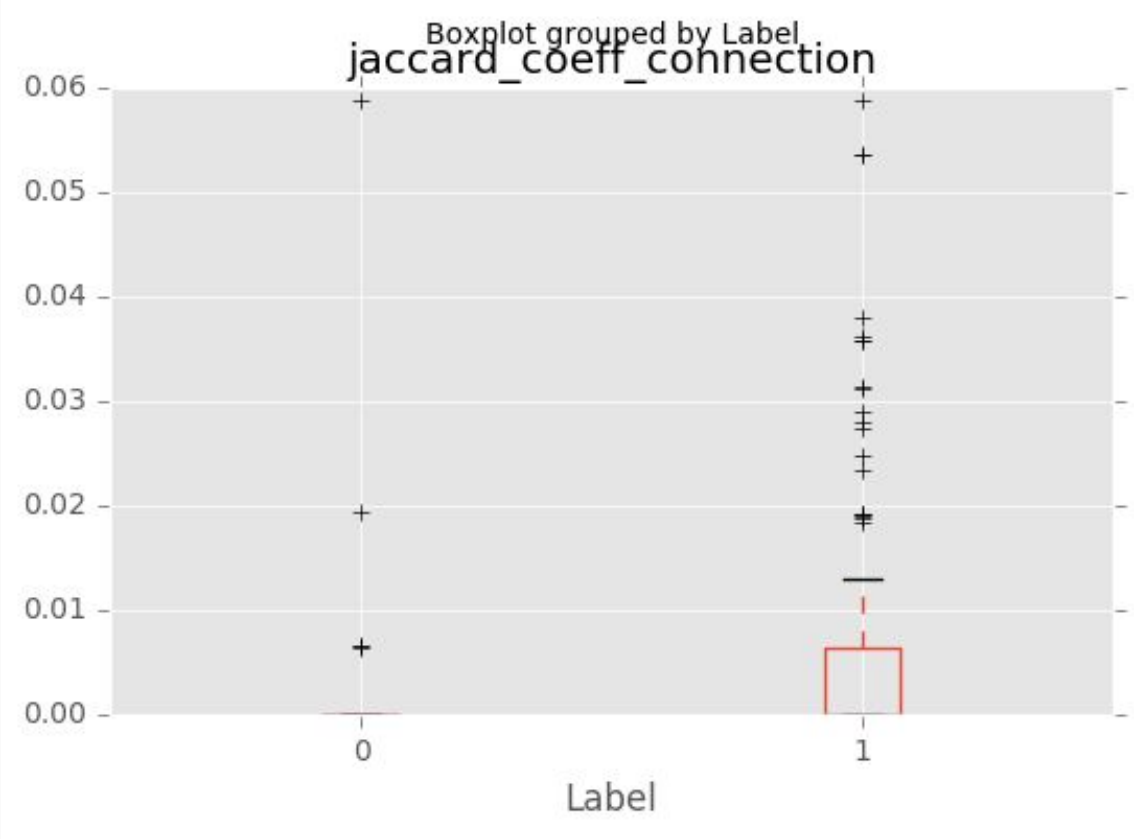
a. Destination site's in-degree in fake network
b. Median polarity score
c. Jaccard Coeff for Connection



Results & Conclusion

Key features for classification

a. Destination site's in-degree in fake network
b. Median polarity score
c. Jaccard Coeff for Connection



Next Steps

- Linguistic Features
- Temporal Features
- Selection Bias

Thanks for your attention!