Fake Twitter Accounts Exploration

Yuzhou Tu

Content

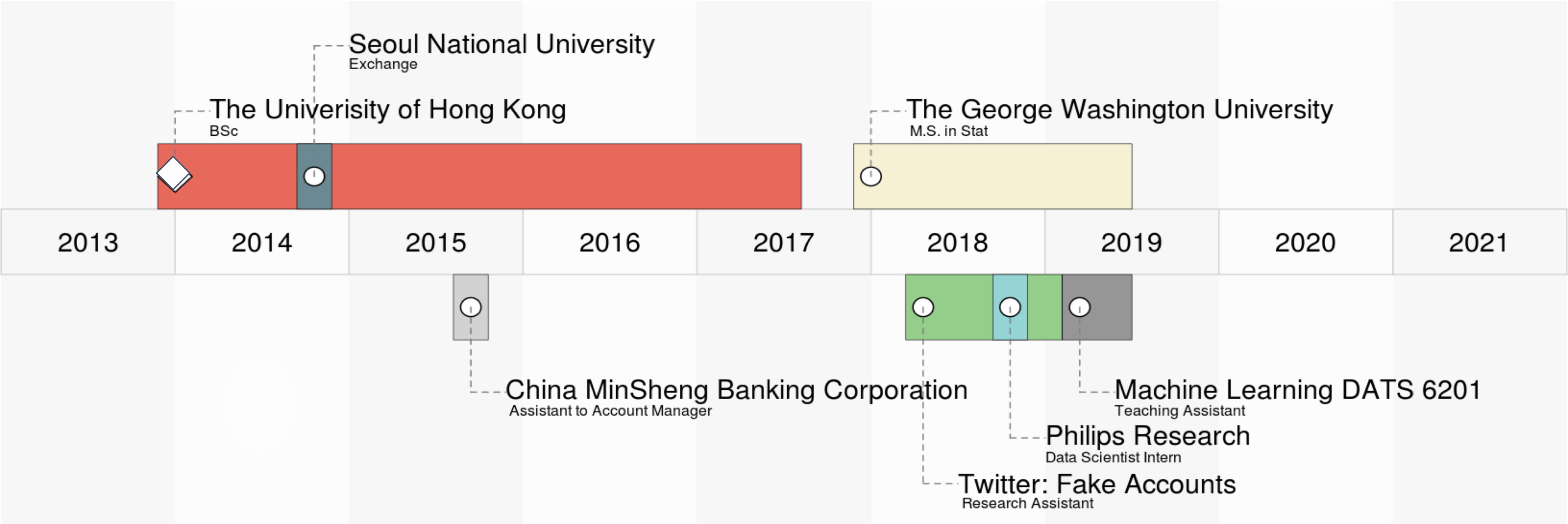
1 Background Information

2 Solutions
- Text Mining
- Network Analysis

3 Results
- Shiny Dashboard

Yuzhou Tu

- Master student @ GWU
- https://www.linkedin.com/in/yuzhou-tu/
- yuzhou_tu@gwu.edu



1

Background

Client

Matthew Hindman

Title: Associate Professor of Media and Public Affairs
Office: MPA 411
Phone: 202-994-5158
Email: hindman@gwu.edu

Areas of Expertise

Internet politics; political communication; online campaigning; the politics of search engines.

Client

Disinformation, **'Fake News'** and Influence Campaigns on **Twitter**

KNIGHT FOUNDATION

OCTOBER 2018

1 Background

Is Trump's campaign boosted by social media ?

Many researchers and main stream medias suspect Trump's campaign in 2016 were manipulated by social medias.

In 2016, Trump's campaign hires Cambridge Analytica to help him win the vote. They develop profiles of people and classify them, then made many fake accounts to show specific fake news to them to influence their choices.

News links:

Shane, S., & Goel, V. (2017, September 06). Fake Russian Facebook Accounts Bought $100,000 in Political Ads. Retrieved from https://www.nytimes.com/2017/09/06/technology/facebook-russian-political-ads.html

The 50,000 automated accounts the company determined had ties to Russia sent more than 2 million election-related tweets between September 1, & 15, N. (n.d.). Russian bots retweeted Trump almost half a million times in final weeks of 2016 campaign. Retrieved from http://money.cnn.com/2018/01/27/technology/business/russian-twitter-bots-election-2016/index.html

1 Data

Number of Twitter Accounts in Each Group

Group	Number
Trump Support	3762
Conservative	3536
Hard Conservative	1399
Other	1951
Intl Right|Anti-Islam	959
Russia	414
Total	12021

1 Data

name	class	sample
id	numeric	9.69147E+17
terrt_url	character	https://twitter.com/davidpwil/status/969147368993157120
created_at	character	Thu Mar 01 09:48:33 +0000 2018
parsed_created_at	time	43160.40872
user_screen_name	character	davidpwil
text	character	All businesses need this payment system - https://t.co/aTBz51aITI
tweet_type	character	original/retweet/quote
coordinates	character	85.310121 27.716710
hashtags	character	America
media	character	https://twitter.com/Norms_Nonsense/status/958294429516705793/photo/1
urls	character	https://vimeo.com/165213948
favorite_count	numeric	22
in_reply_to_screen_name	character	NewsRepublicans
in_reply_to_status_id	numeric	7.95074E+17
in_reply_to_user_id	numeric	7.64637E+17
lang	character	en
place	character	Luling, LA

1 Data

name	class	sample
retweet_count	numeric	7
retweet_or_qoute_id	numeric	9.69657E+17
retweet_or_qoute_screen_name	character	celal2023
retweet_or_qoute_user_id	numeric	249207900
source	character	Twitter Web Client
user_id	numeric	25924436
user_created_at	character	Mon Mar 23 01:20:04 +0000 2009
user_default_profile_image	character	FALSE
user_description	character	What we do in life echoes in eternity. \M/ #Indy500 @IndyCar @F1
user_favourites_count	numeric	67212
user_followers_count	numeric	1942
user_friends_count	numeric	506
user_listed_count	numeric	122
user_location	character	Florida
user_name	character	DavidPWil -Christian
user_statuses_count	numeric	57128
user_time_zone	character	Eastern Time (US & Canada)
user_urls	character	http://ifyoucouldknow.info
user_verified	character	FALSE

1 Main Objectives & Solutions

Analyze the tweets sent by fake accounts

- Text Mining
- Topic Modeling

Find out the connection between fake accounts

- Network Analysis

Track the the newest behavior

- Shiny Dashboard

1 Key Challenges

Challenges

No Existed Database
Information about twitter accounts are stored in multiple csv files

Massive Data
The dataset contains more than 3 million tweets

Lack of data visualization
Data is in text format, special data visualization is required

Unstructured Data
Tweet data contains a large number of uninformative words and symbols

2

Solutions

2 Data Preparation

Original Text:
[1] " RT @marklevinshow: Had they wanted to save children<U+2019>s lives, schools would<U+2019>ve been secured and children made safe long ago. Anyone who believes ridding us of our best defense of weapons has anything to do with saving children<U+2019>s lives is delusional. @NRA #NRA #2A #MolonLabe #MakeAmericaGreatAgain https://t.co/3AkGx3dl49”

After Cleaning:
want save children live school wouldv secur children made safe long ago anyone believ rid us best defens weapon anyth save children live delusional nra molonlab makeamericagreatagain

- Remove Retweet: RT @...
- Remove URLs
- Remove Punctuation Marks
- Remove Numbers
- Remove Unicode: <U+2019>
- Remove Stopwords: “i”, "me”, "my”, "myself”, "we”, …
- Keep Word Stem Only

2 Data Preparation

TermDocumentMatrix:

```
                Docs
Terms            148 149 150
  emailimessagefb  0   0   0
  emanuel          0   0   0
  embarrass        0   0   0
  embarrassmenthi  0   0   0
```

Word Frequencies

DocumentTermMatrix:

```
     Terms
Docs  emailimessagefb emanuel embarrass embarrassmenthi
  148               0       0         0               0
  149               0       0         0               0
  150               0       0         0               0
```

Topic Modeling

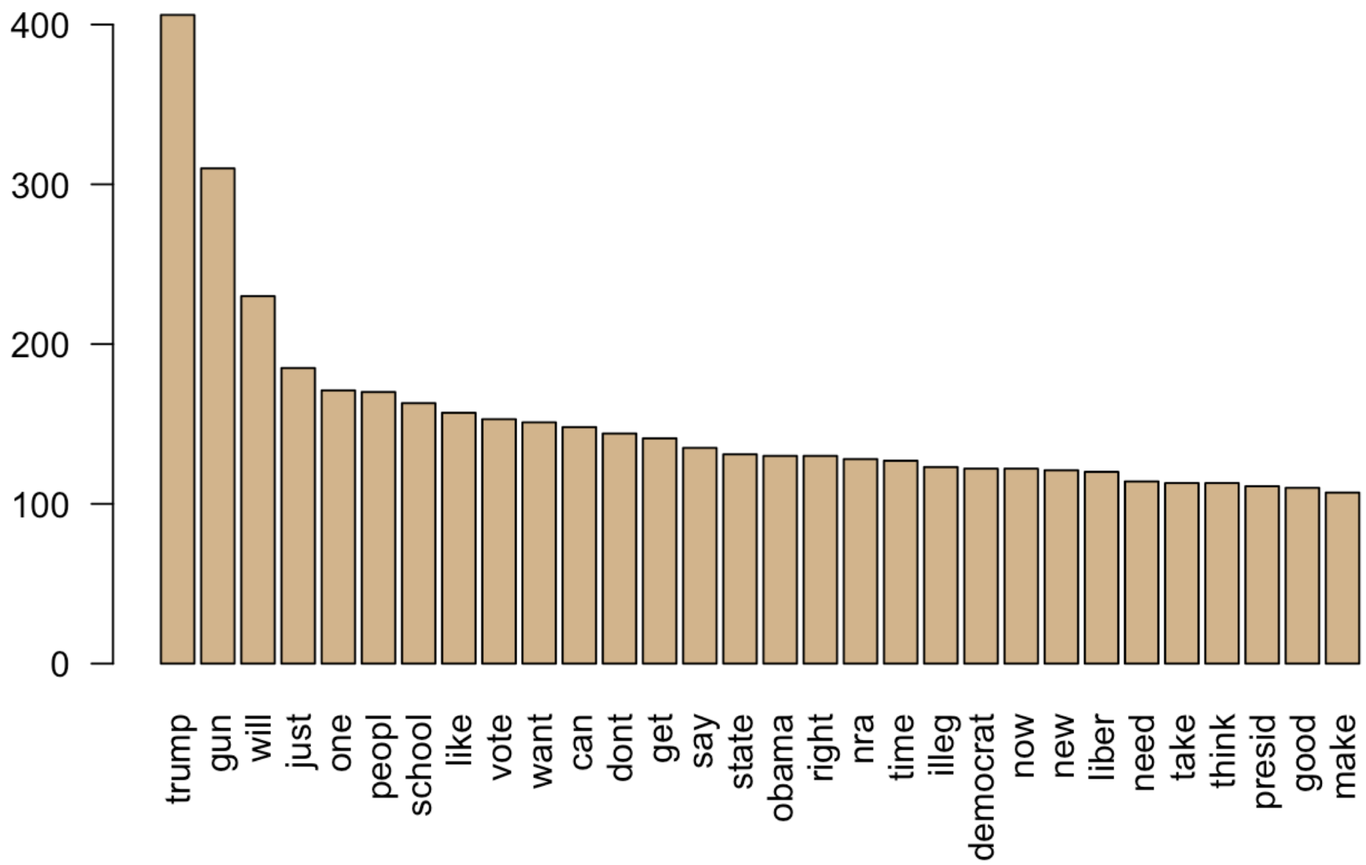
Tidy Text:

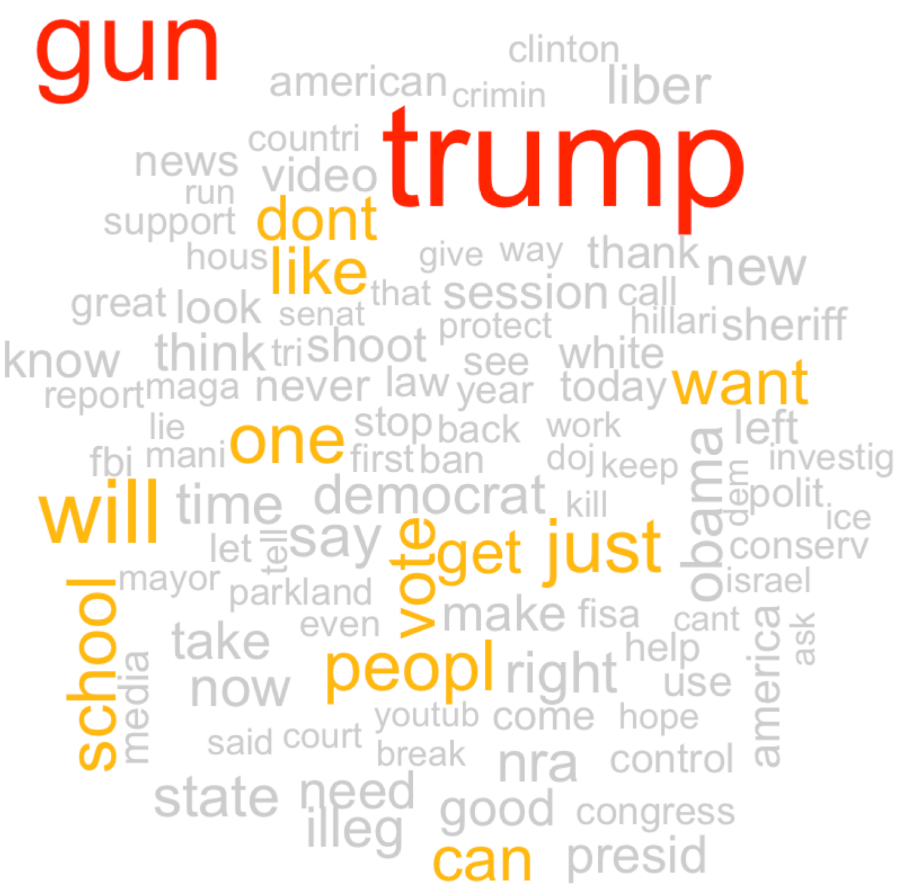
id <dbl>	word <chr>
9.691474e+17	businesses
9.691474e+17	payment
9.691474e+17	system
9.694061e+17	freak
9.694061e+17	america
9.694061e+17	celebrating
9.694061e+17	mike
9.694061e+17	huckabee
9.694061e+17	forced
9.694061e+17	intolerantcommie

Sentiment Analysis
Bigram Analysis

2 Text Mining

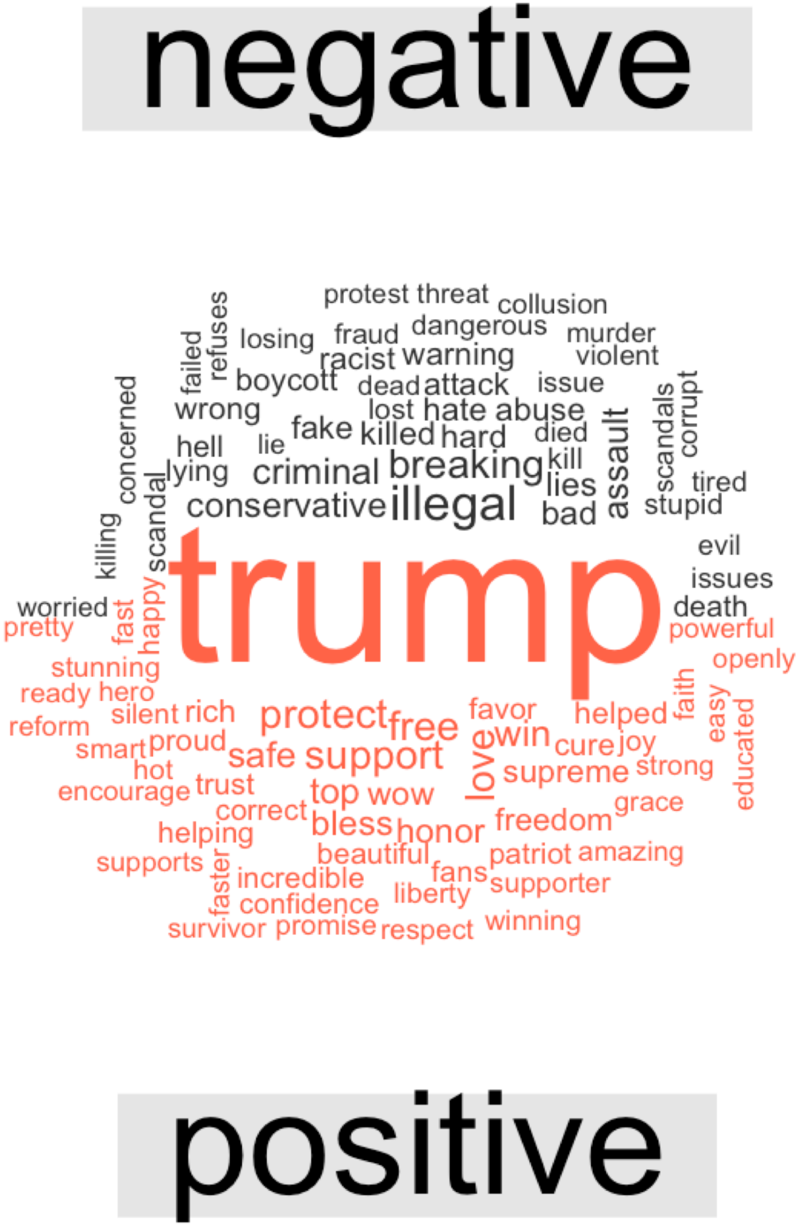
Word Frequencies



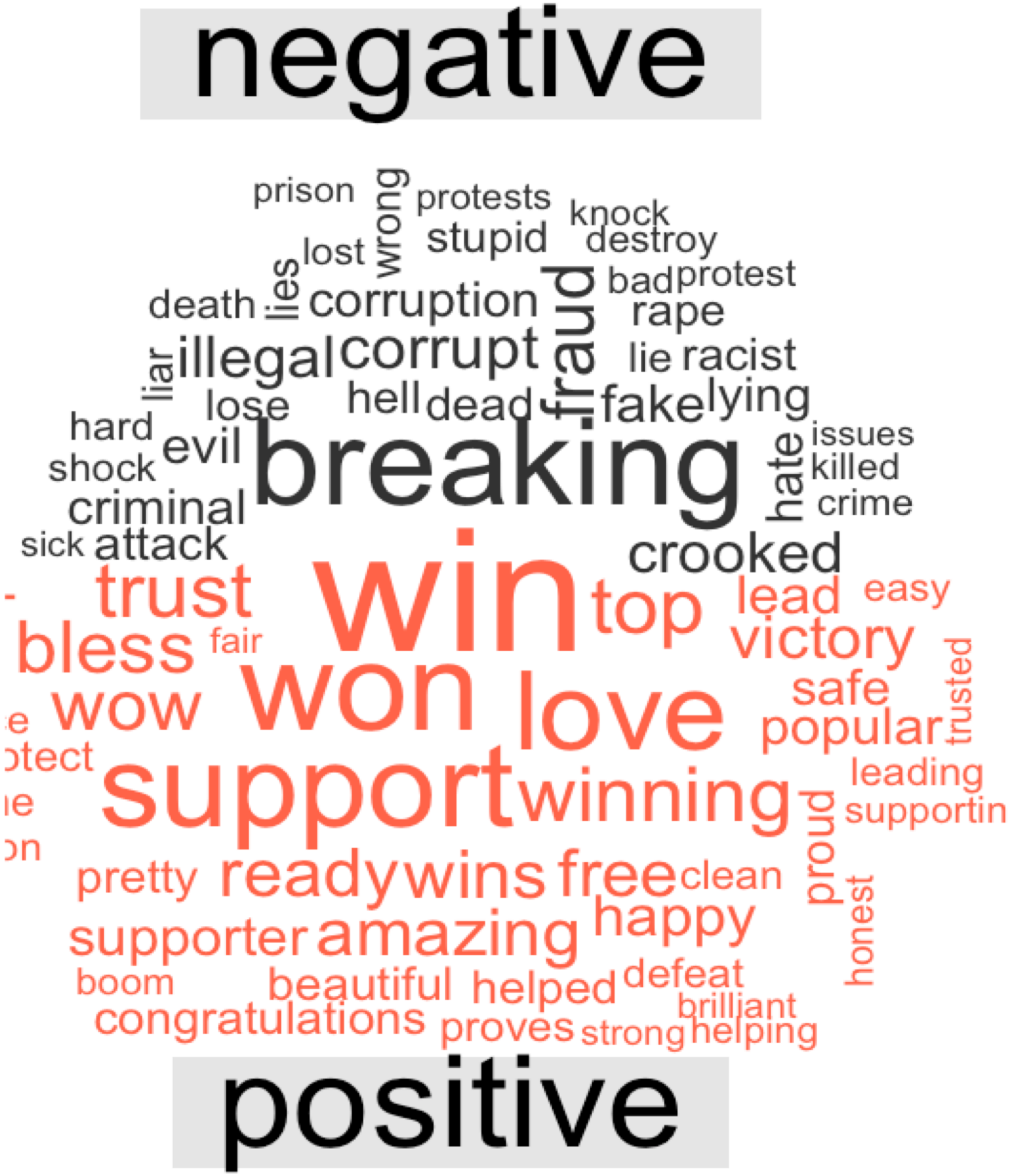


Text Mining
Sentiment Analysis

Remove ‘trump’



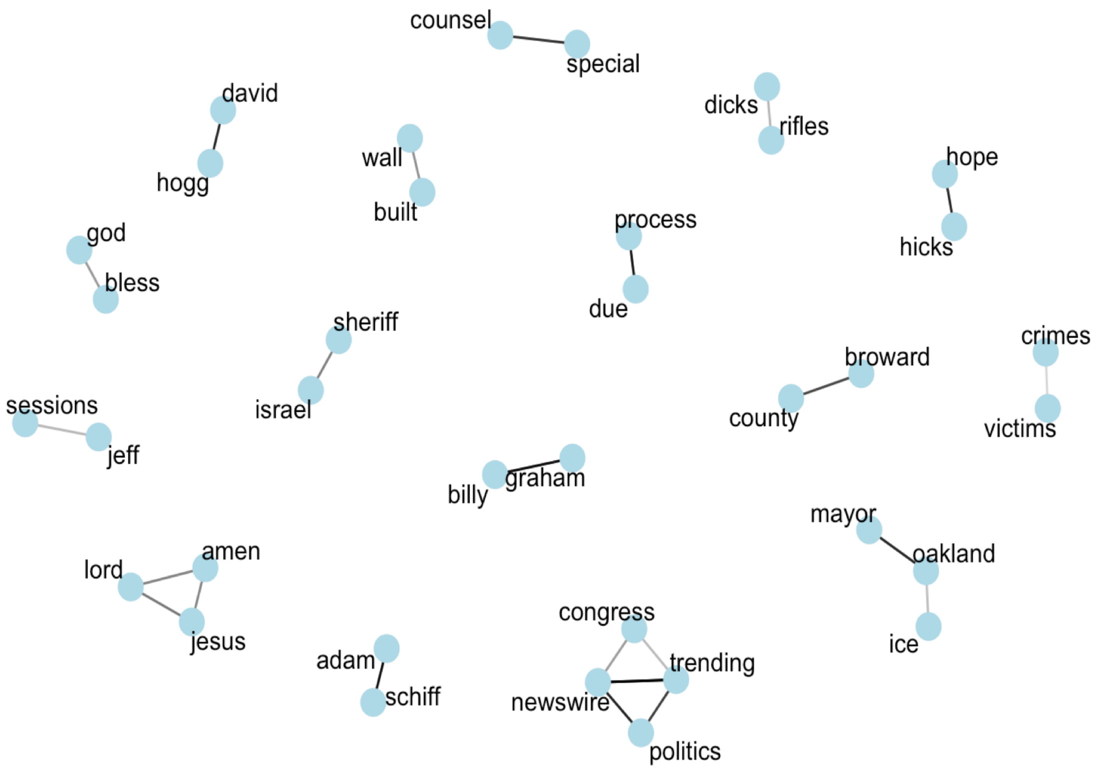
Text Mining
Sentiment Analysis



2 Text Mining

Bigrams

item1 <chr>	item2 <chr>	correlation <dbl>
newswire	trending	0.9327743
trending	newswire	0.9327743
graham	billy	0.8205177
billy	graham	0.8205177
schiff	adam	0.8172636
adam	schiff	0.8172636
process	due	0.7989087
due	process	0.7989087
newswire	politics	0.7727941
politics	newswire	0.7727941



	Has word Y	No word Y	Total
Has word X	n11	n10	n1·
No word X	n01	n00	n0·
Total	n·1	n·0	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}}$$

2 Text Mining

Topic Modeling (LDA) – Algorithm

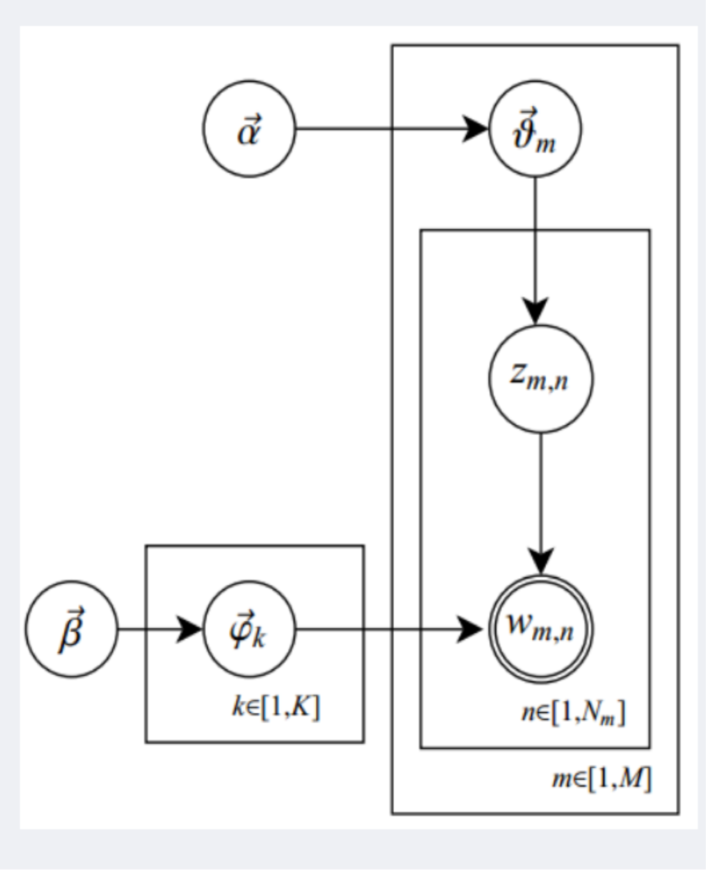
Assumptions of Latent Dirichlet Allocation(LDA):

Every document is a mixture of topics.

Every topic is a mixture of words.

// topic plate
for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim \mathrm{Dir}(\vec{\beta})$
// document plate:
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim \mathrm{Dir}(\vec{\alpha})$
  sample document length $N_m \sim \mathrm{Poiss}(\xi)$
  // word plate:
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n} \sim \mathrm{Mult}(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim \mathrm{Mult}(\vec{\varphi}_{z_{m,n}})$

Fig. 7. Generative model for latent Dirichlet allocation.



2 Text Mining

Topic Modeling (LDA) - Example

Assumptions of Latent Dirichlet Allocation(LDA):

Every document is a mixture of topics.

Every topic is a mixture of words.

Document 1: I had a peanut butter sandwich for breakfast.
Document 2: I like to eat almonds, peanuts and walnuts.
Document 3: My neighbor got a little dog yesterday.
Document 4: Cats and dogs are mortal enemies.
Document 5: You mustn’t feed peanuts to your dog.

Topic 1: 30% peanuts, 15% almonds, 10% breakfast…
Topic 2: 20% dogs, 10% cats, 5% peanuts…

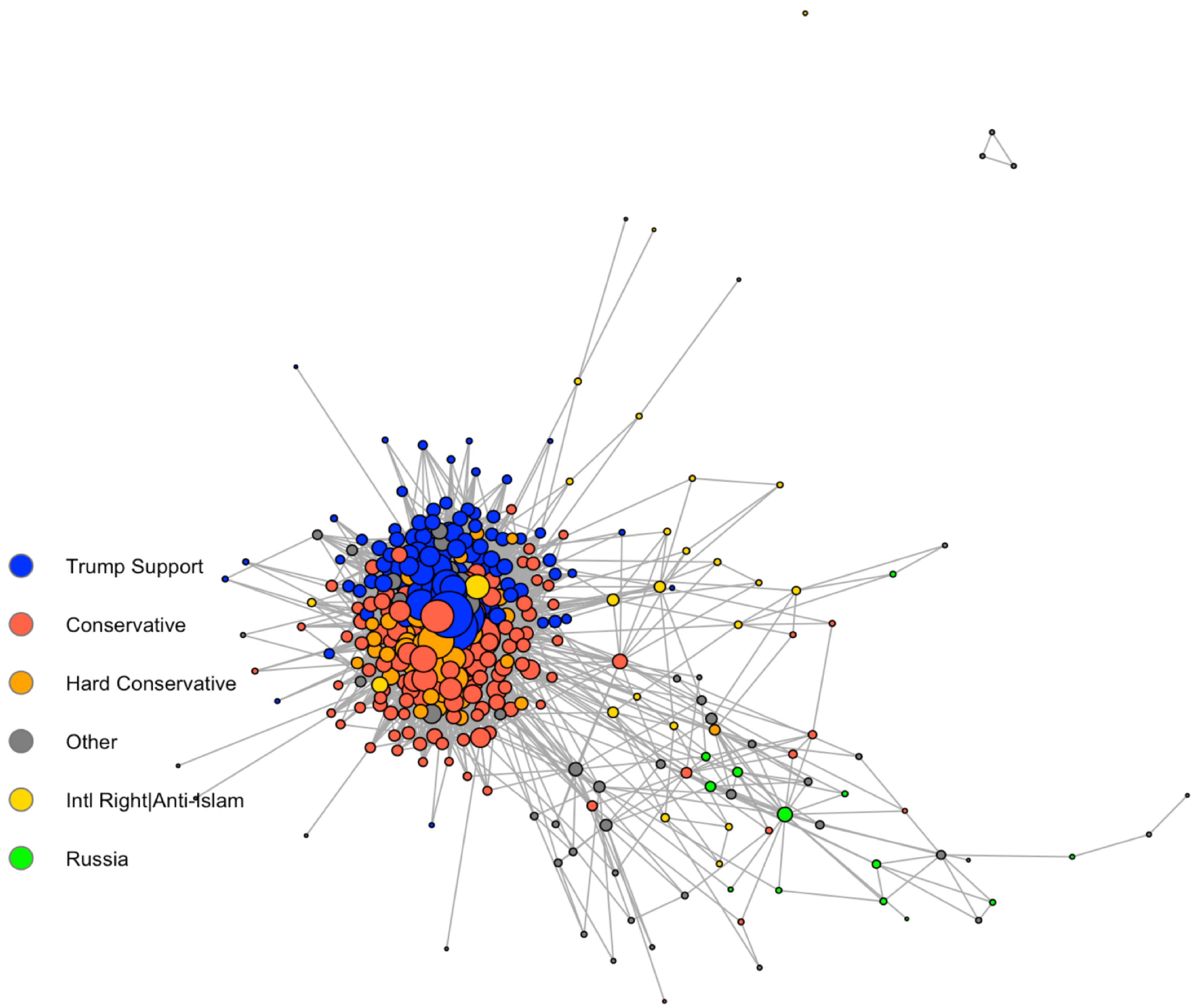
Document 1: 100% Topic 1
Document 2: 100% Topic 1
Document 3 : 100% Topic 2
Document 4: 100% Topic2
Document 5:   70% Topic 1, 30% Topic 2

2 Topic Modeling

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	die	trump	trump	people	в	de	syria	de	school	の
2	de	fbi	maga	dont	di	la	war	la	gun	を
3	der	obama	obama	time	и	el	russia	le	nra	に
4	und	president	news	im	che	en	news	les	shooting	が
5	van	democrats	video	love	il	se	uk	à	guns	は
6	ist	memo	tcot	country	la	los	eu	des	florida	と
7	das	hillary	hillary	god	ve	por	israel	en	trump	で
8	het	clinton	america	youre	bir	para	world	pour	sheriff	た
9	en	people	clinton	day	на	del	russian	pas	cnn	て
10	een	america	president	hes	è	da	government	une	parkland	も

2 Network Analysis - Followers

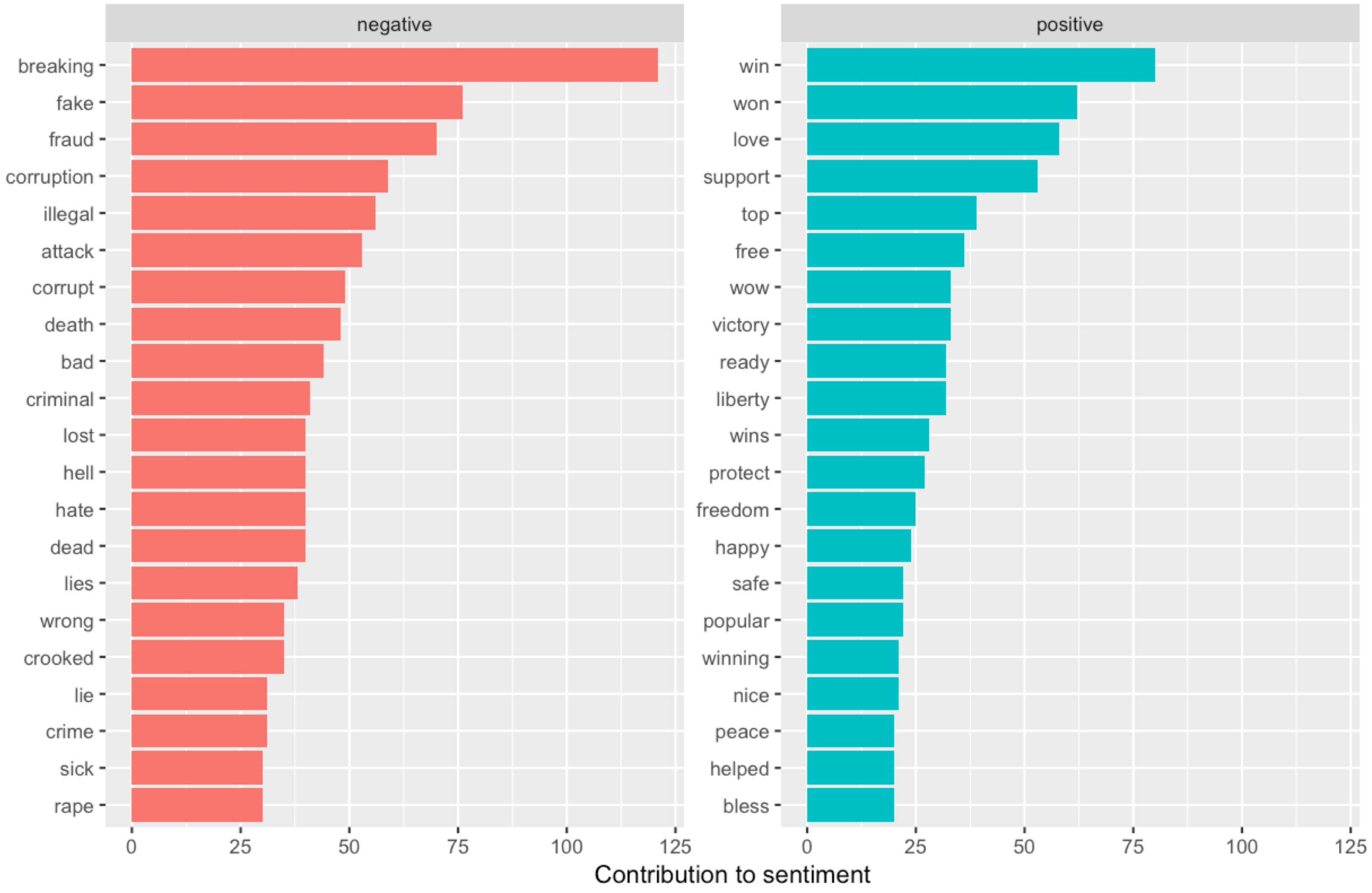
- 500 random selected accounts
- Connected if followed
- Colored by the group they belong to



3

Results

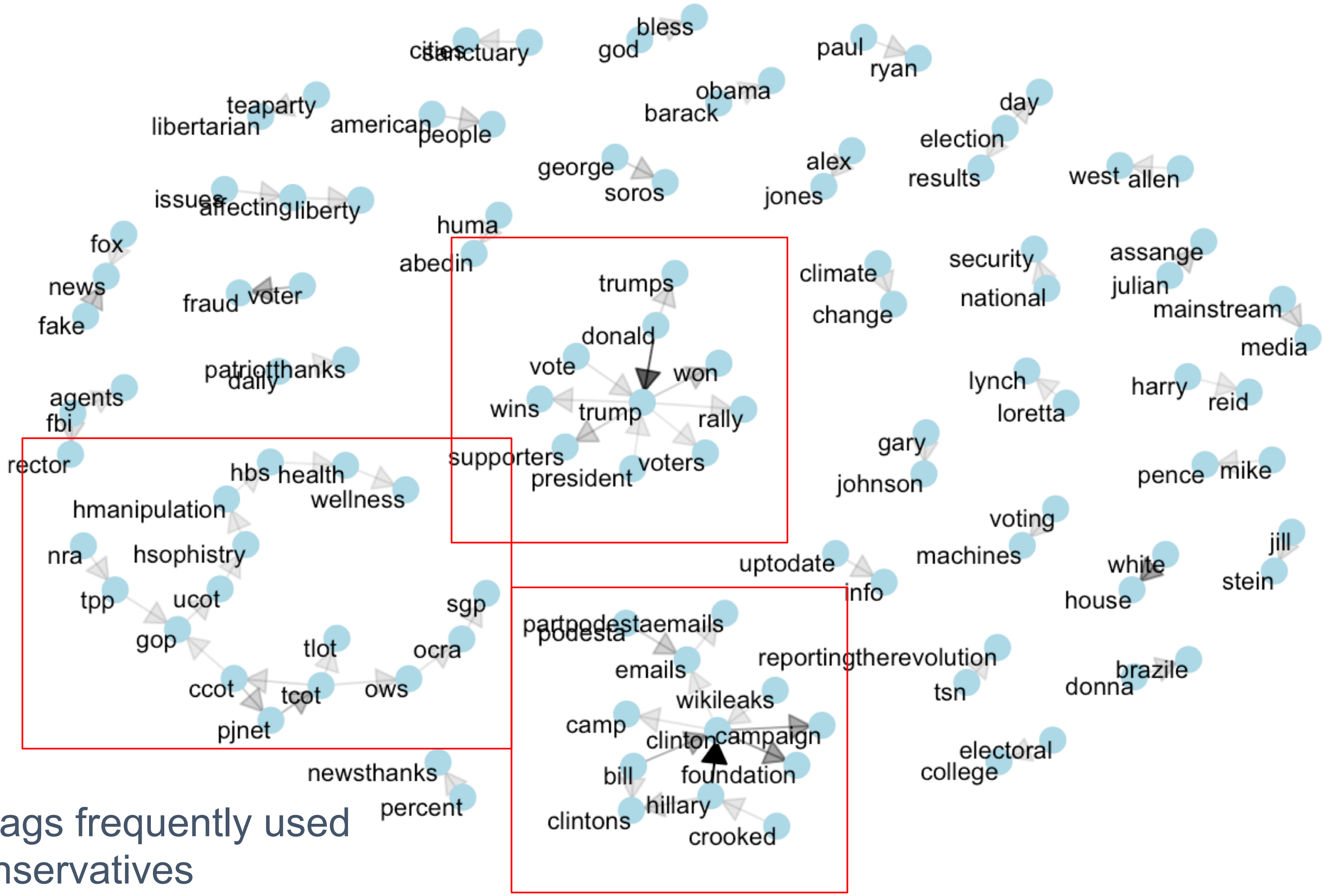
3 Text Mining
Sentiment Analysis



more negative words

3 Text Mining

Bigrams



Hashtags frequently used by conservatives

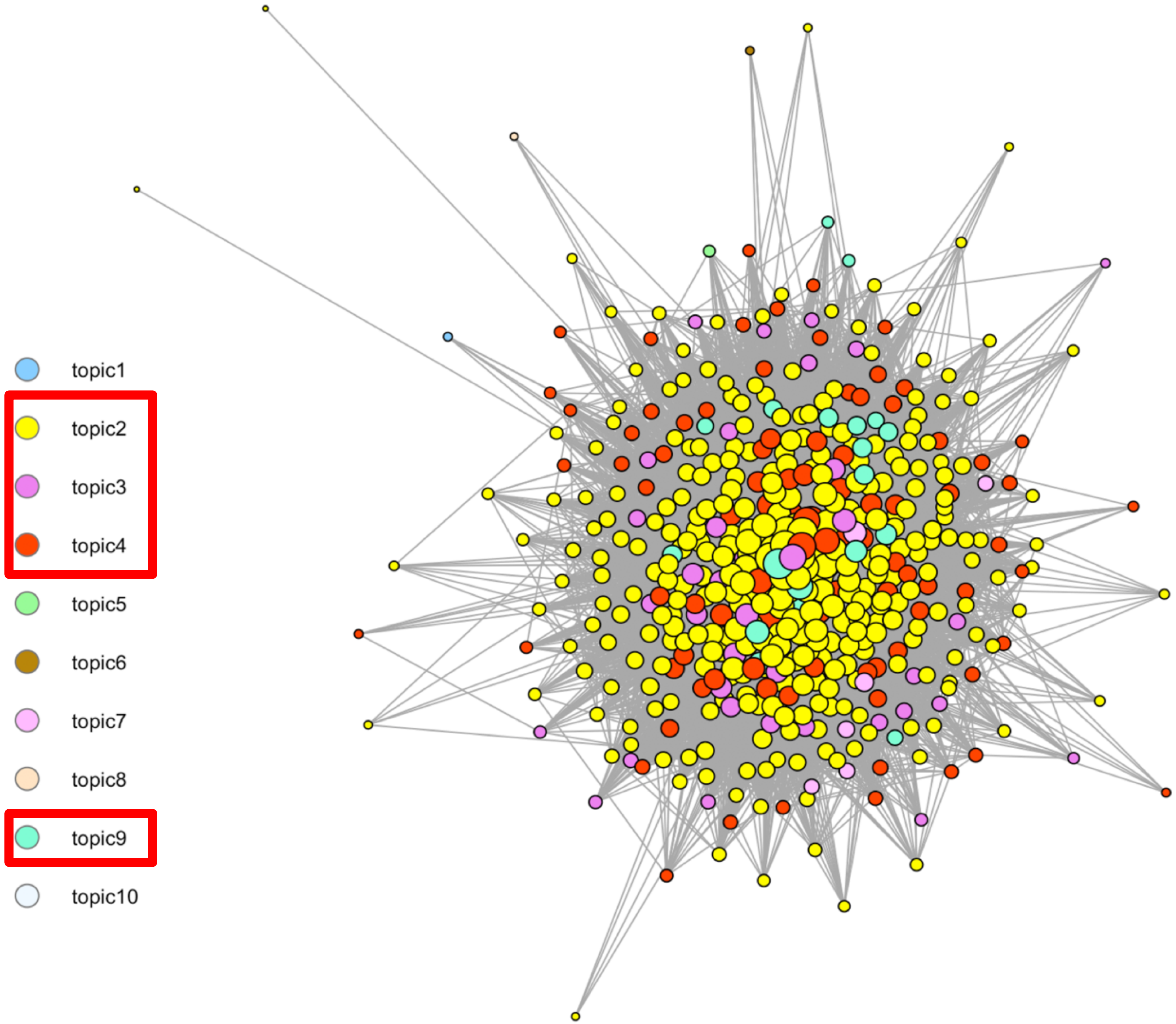
Trump:

Hillary

3 Topic Modeling

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	die	trump	trump	people	в	de	syria	de	school	の
2	de	fbi	maga	dont	di	la	war	la	gun	を
3	der	obama	obama	time	и	el	russia	le	nra	に
4	und	president	news	im	che	en	news	les	shooting	が
5	van	democrats	video	love	il	se	uk	à	guns	は
6	ist	memo	tcot	country	la	los	eu	des	florida	と
7	das	hillary	hillary	god	ve	por	israel	en	trump	で
8	het	clinton	america	youre	bir	para	world	pour	sheriff	た
9	en	people	clinton	day	на	del	russian	pas	cnn	て
10	een	america	president	hes	è	da	government	une	parkland	も

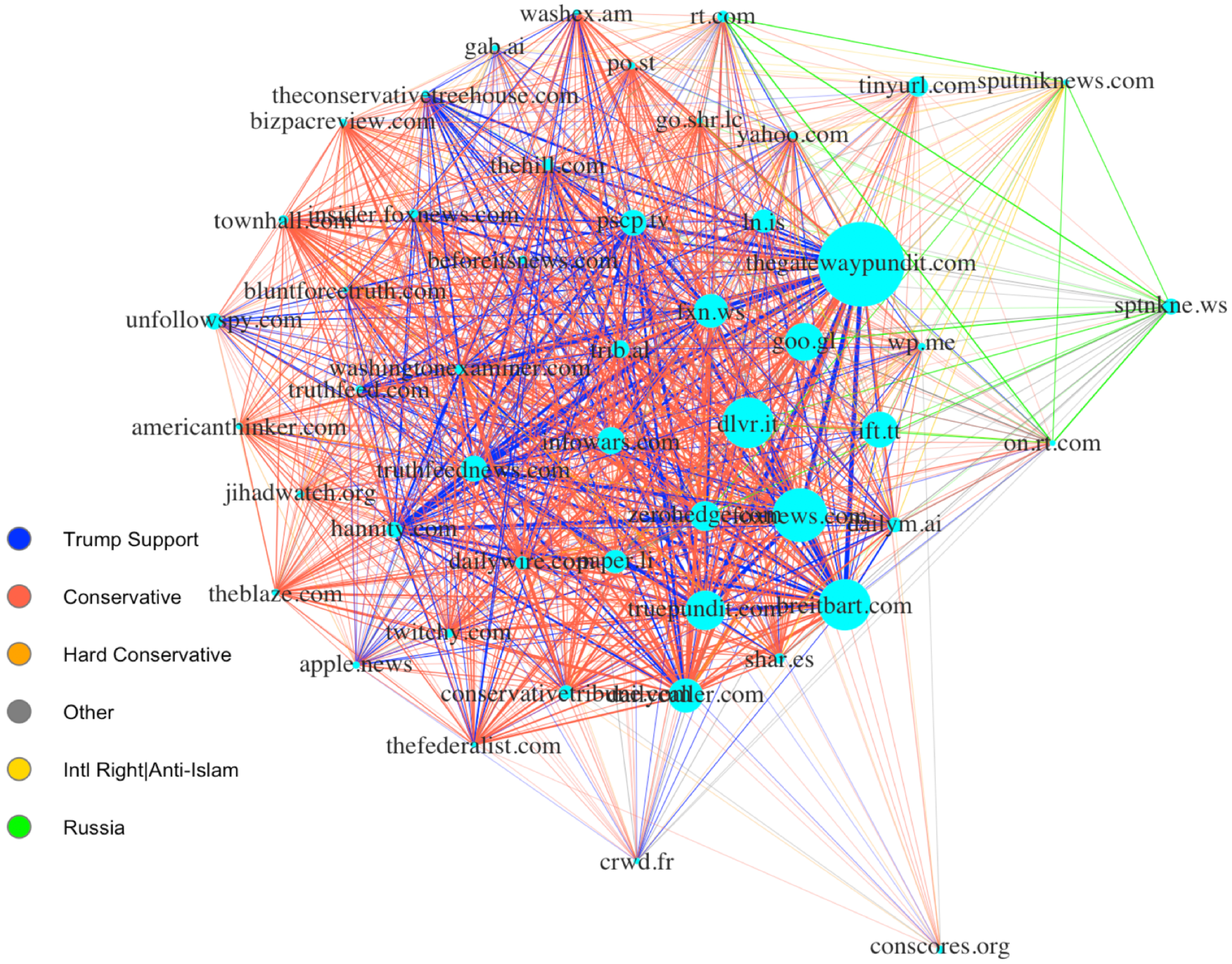
Network Analysis – Topics



- 500 random selected accounts from database
- Connected if followed
- Colored by the topic they tweet (from topic modeling)

Fake accounts discuss about 4 major topics are closely connected

3 Network Analysis - URLs



- Each point is a website quote by fake accounts
- Connected if two websites are quoted by same account
- Lines are colored by groups

Fake accounts from ’Trump Support’ and ‘Conservative’ groups are quoting the same websites

3 Conclusion

Very similar but abnormal activities among groups of fake accounts

Try to support Trump

High probability of being controlled by robot

Use Shiny dashboard to Keep tracking their activities

Notice their strategies in advance

Use URLs or friends network to category or identify new fake accounts

3 Feedback

Department Newsletter, Fall 2018

STATS STUDENTS AID GW RESEARCHERS

The department invites researchers from other GW departments to present statistical problems they encounter in their research to the Statistical Consulting course for assistance. Student teams often solve these problems with guidance from course instructors.

For example, students in Professor **Judy Wang**'s spring 2018 class worked with Dr. **Matthew Hindman**, professor of media and public affairs, to study latent features of fake Twitter accounts using text mining, network analysis and other techniques. They developed a Shiny dashboard for dynamic analysis and visualization. Professor Hindman noted that the "students were incredibly helpful" and "In less than two months, they helped put together a toolkit tracking online disinformation in real time, applying cutting-edge text mining and network analysis techniques. They dramatically sped up my research agenda." Another team worked on a research project with Associate Professor of Religion **Irene Oh** on religion and maternal health.

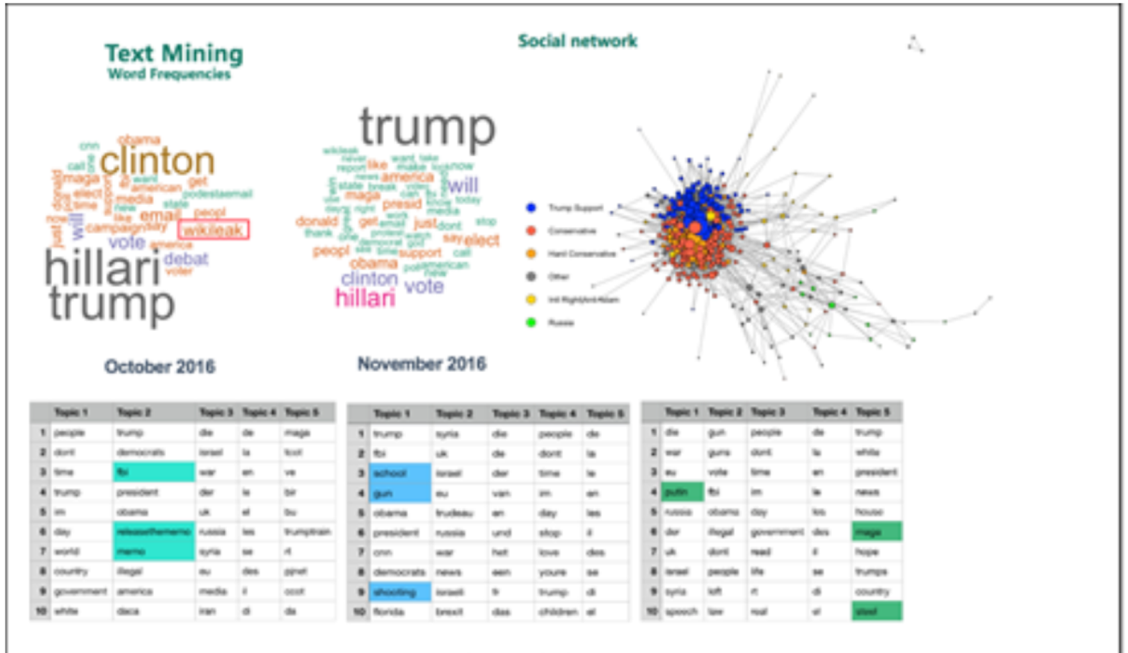


Table 2. Top 10 words for 5 Topics in Jan 2018

3 Feedback

Comment from Client

I was most impressed by their hard work, and their ability to quickly get up to speed on difficult, even cutting-edge topics. This was a challenging project that involved lots of different pieces that had to come together: data wrangling, database deployment, software development, network analysis, matrix factorization, and text mining. The fact that the team was able ultimately to put together a tool kit for **near-real-time analysis of Twitter data with a working dashboard** in just a month and a half is *extremely* impressive.

THANKS