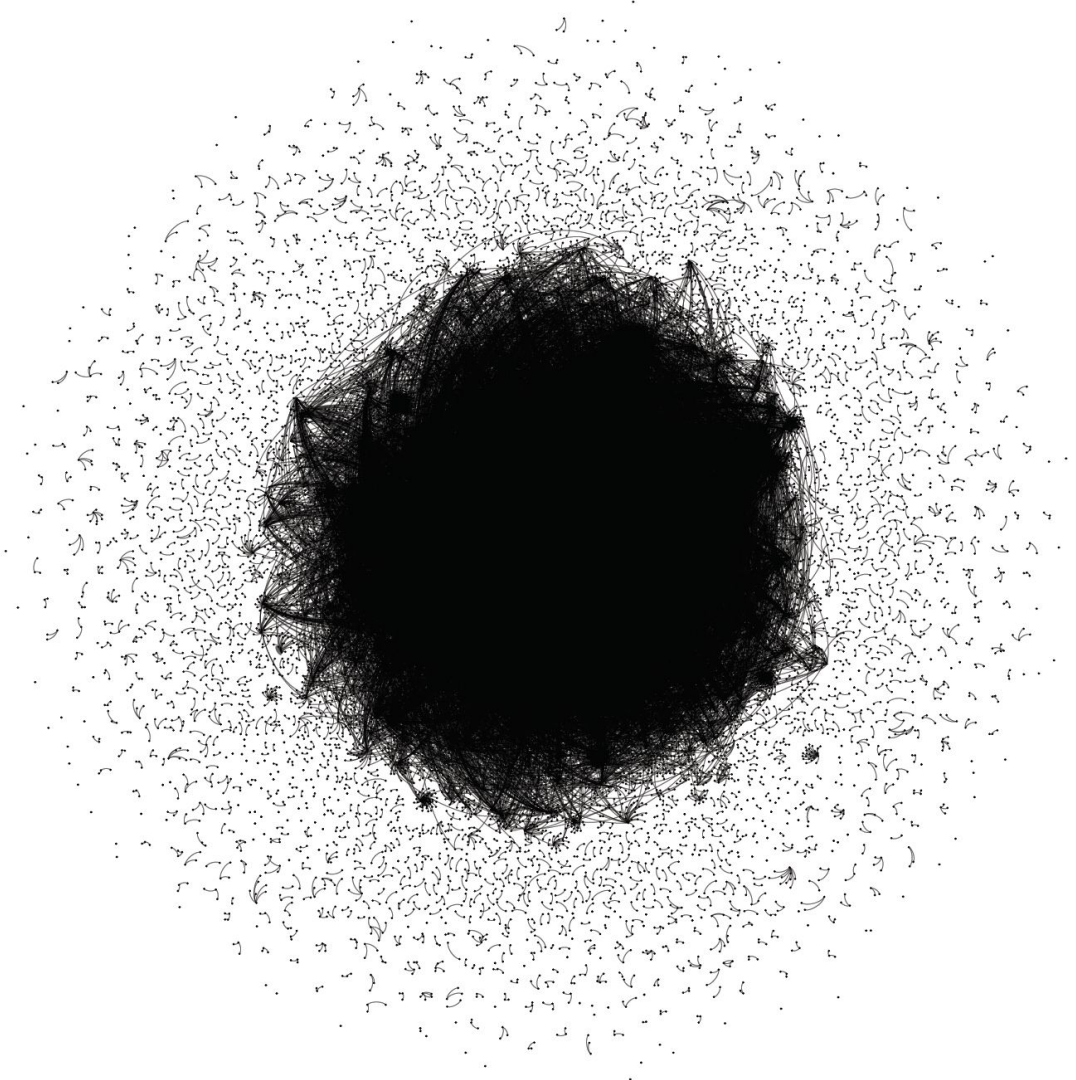
Network Analysis of the 2016 Democratic Convention

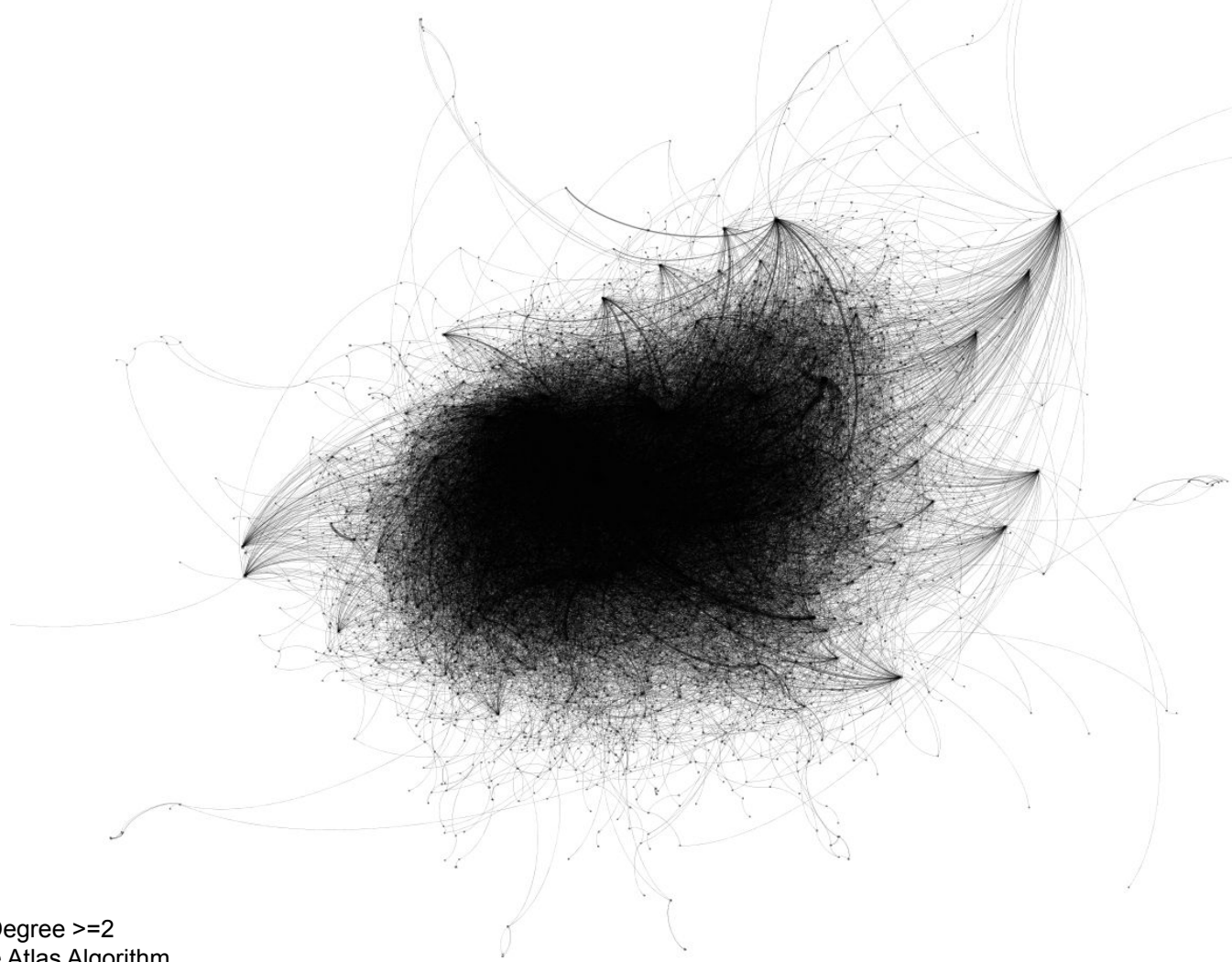
Identifying political communities on twitter using graph theory and natural language processing

Introduction to Network Analysis

- Gephi Software
- Nodes = Users
- Edges = Tweets
- Nodes are weighted by the sum of their edges
- Data:
  - 130,000 tweets about the 2016 DNC convention
  - 8,329 Nodes
  - 34,143 Edges
  - User metadata for each tweet



8,329 Nodes (Twitter Users)
34,143 Edges (Tweets)



Filter: Node Degree >=2
Layout: Force Atlas Algorithm

18 (27.57%)
29 (13.54%)
70 (7.81%)
25 (6.01%)
46 (5.79%)

Labeling Communities:

1. What words do members of the community use most frequently in conversation?
2. What words do members of the community use to describe themselves?
3. Who are the top 5 most influential users in the community?
4. What terms/words are most significant in defining the community (TF-IDF)

18 (27.57%)
29 (13.54%)
70 (7.81%)
25 (6.01%)
46 (5.79%)

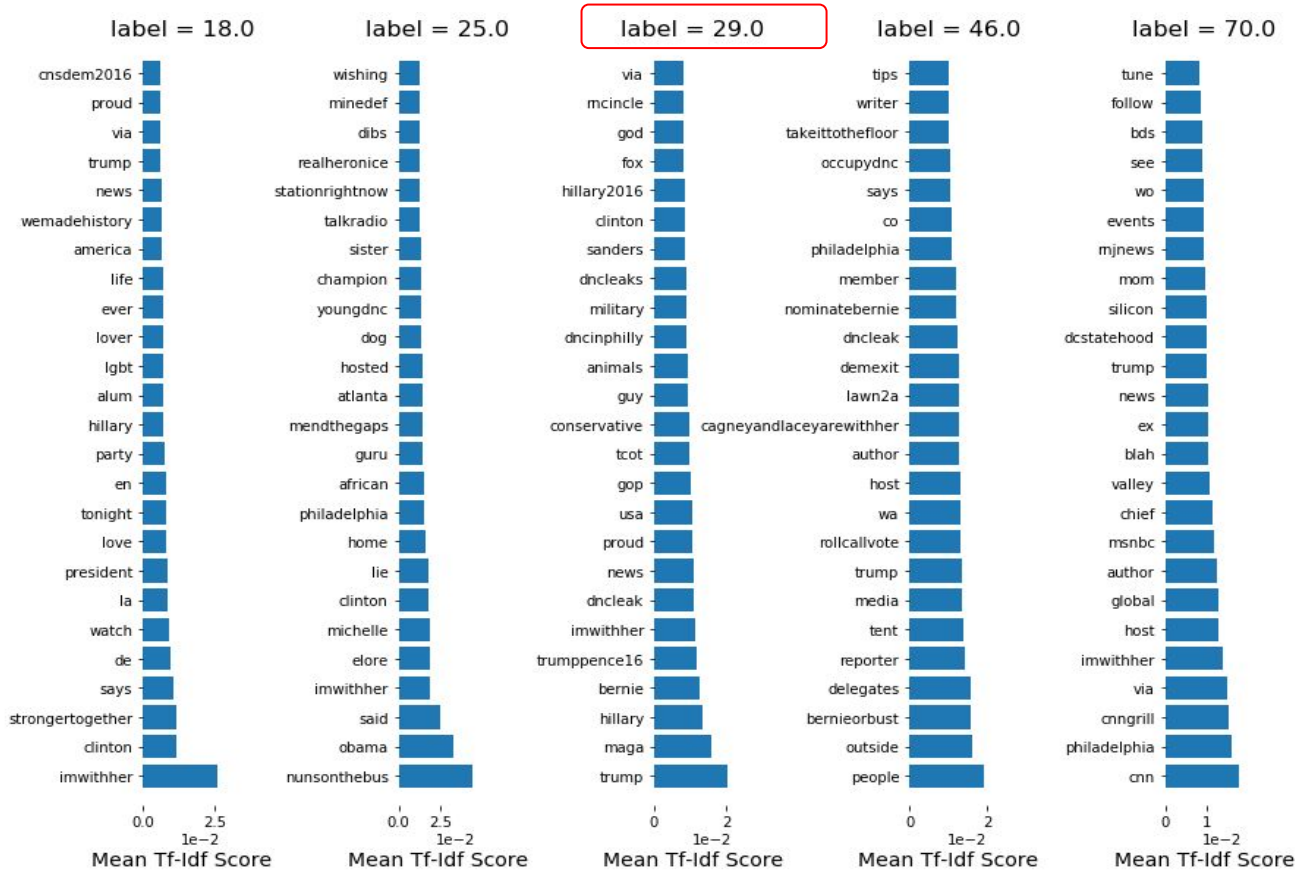
What does community #29 talk about most?

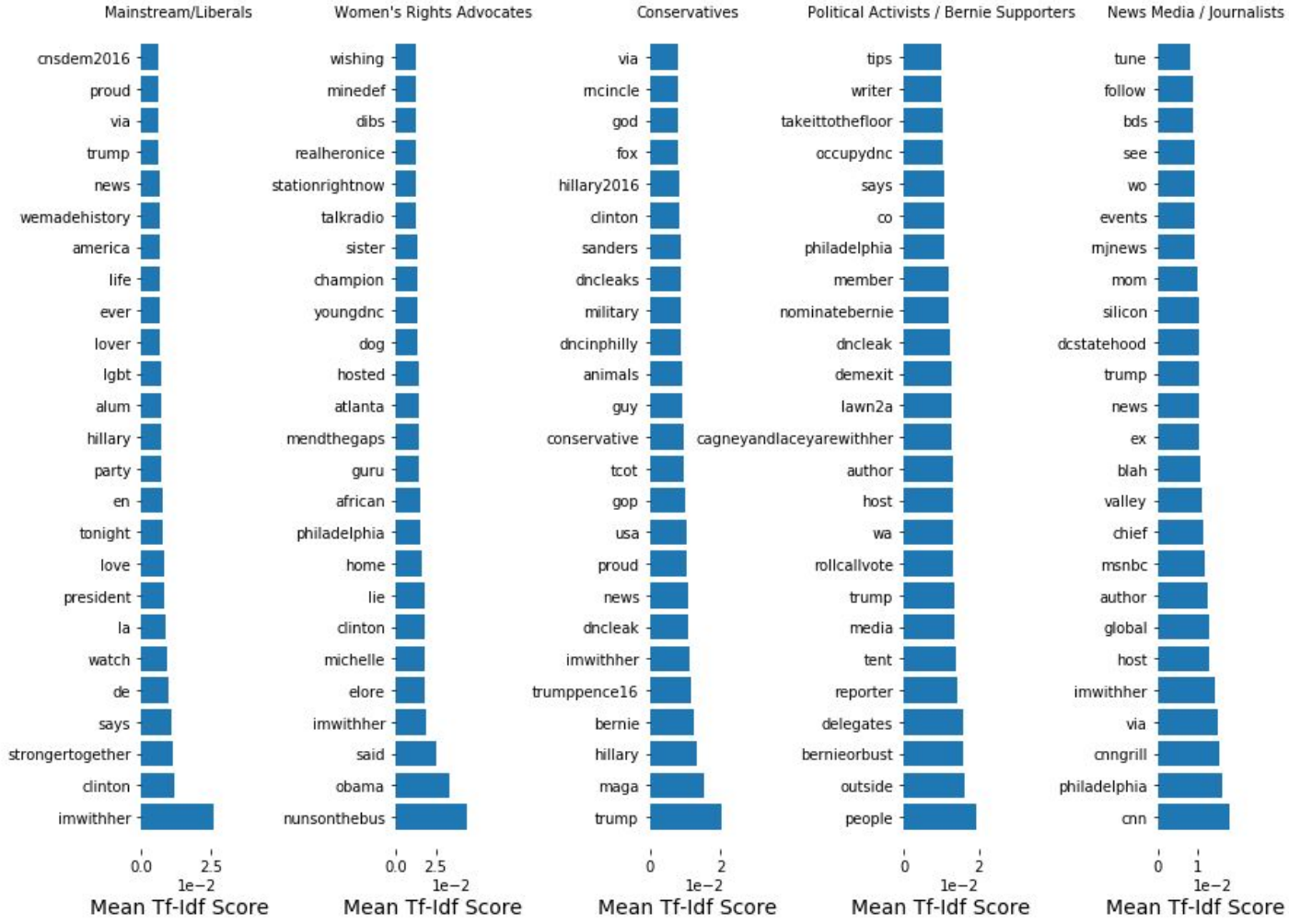
How do they describe themselves?

Who are the top users in community #29?

user_screen_name	user_location	user_description
LouDobbs	New York, NY	lou dobbs tonight fox business network pm eastern
DineshDSouza	United States	behind modern-day socialism evil stop `` united states socialism '' available pre-order today
brithume	Southwest Florida	sr. analyst fox news channel arguments welcome name callers verbal abusers blocked
anamariecox	Minneapolis, MN	ever made world better perfect she/her outright mental defective dog exley cats luke leia milowda gonya leva x
wsbtv	Atlanta, GA	metro atlanta 1 source breaking news weather traffic coverage count

What are the most important features (characteristic words) for the community?




Mainstream/Liberals
cnsdem2016
proud
via
trump
news
wemadehistory
america
life
ever
lover
lgbt
alum
hillary
party
en
tonight
love
president
la
watch
de
says
strongertogether
clinton
imwithher
0.0
2.5
1e−2
Mean Tf-Idf Score
Women's Rights Advocates
wishing
minedef
dlbs
realheronice
stationrightnow
talkradio
sister
champion
youngdnc
dog
hosted
atlanta
mendthegaps
guru
african
philadelphia
home
lie
clinton
michelle
elore
imwithher
said
obama
nunsonthebus
0.0
2.5
1e−2
Mean Tf-Idf Score
Conservatives
via
rncincle
god
fox
hillary2016
clinton
sanders
dncleaks
military
dncinphilly
animals
guy
conservative
tcot
gop
usa
proud
news
dncleak
imwithher
trumppence16
bernie
hillary
maga
trump
0
2
1e−2
Mean Tf-Idf Score
Political Activists / Bernie Supporters
tips
writer
takeittothefloor
occupydnc
says
co
philadelphia
member
nominatebernie
dncleak
demexit
lawn2a
cagneyandlaceyarewithher
author
host
wa
rollcallvote
trump
media
tent
reporter
delegates
bernieorbust
outside
people
0
2
1e−2
Mean Tf-Idf Score
News Media / Journalists
tune
follow
bds
see
wo
events
mjnews
mom
silicon
dcstatehood
trump
news
ex
blah
valley
chief
msnbc
author
global
host
imwithher
via
cnngrill
philadelphia
cnn
0
1
1e−2
Mean Tf-Idf Score

18 (27.57%)
29 (13.54%)
70 (7.81%)
25 (6.01%)
46 (5.79%)

Predicting User Community: Accuracy

For each user, combine their description with the text of all their tweets. “Vectorize” the text data and fit models using the Scikit Learn module.

- Logistic Regression Model: 0.47
- Random Forest Model: 0.46
- Stochastic Gradient Descent Model: 0.48
- Naive Bayes Model: 0.45

Limitations:

- Only 902 twitter users in final dataset
- Multiple classification with 5 classes
- Gephi can only handle so much data at a time.

Future Work:

- Perform sentiment analysis on each tweet (positive or negative)
- Reduce the number of classes
- Run the network analysis on a virtual machine with more memory and computational power.

Thank You.

GitHub
github.com/jakemull13

LinkedIn
linkedin.com/jacob.c.mullins

Email
jakemull13@gmail.com

speaking
obama
donald
great
thank
trump
msnbc
cnn
see
people
love
time
night
sanders
clinton
good
hillary
fb
america
via
bernie
today
imwithher
tcot
philadelphia
global
ex
conservative
want
love
alum
author
rt
host
chief
proud
mom
new
kid
right
reporter
fan
tweets
former
cnn
native
com
endorsement
world
correspondent