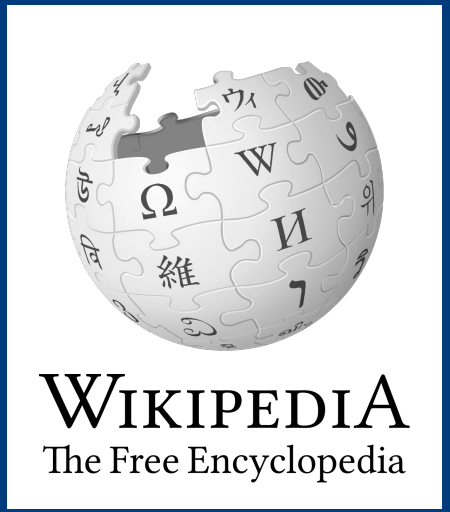
Leveraging Wikipedia Data To Assess Public Awareness of U.S. Hate Groups



Final Project of Data Science and Public Policy

Group 5: Michelle Strayer, Louis Jarvers, Liam Kearney, Jiaxi Liu, Leia Song

Louis Jarvers
(MPA '21)

Liam Kearney
(MIA '22)

Jiaxi Liu
(Stats '21)

Jiage Song
(Stats '21)

Michelle Strayer
(QMSS/Stats '21)

Agenda

1. Motivation & Problem
2. Data Source and Extraction
3. Analysis methods
    a. Social Network Analysis
    b. Natural Language Processing
    c. Times Series Analysis
4. Findings
5. Policy Implication & Next Steps

1. Motivation

Key Question: Visualizing and measuring public knowledge of U.S. hate groups

Wikipedia content as proxy for public attention

Why does this topic matter?

- The rise of right wing extremism and white supremacist groups
- Rise of anti-immigration hate groups with waves of immigration
- Trends in anti-Asian, antisemitic, and racist hate crimes
- Importance of understanding public knowledge in prevention and policy

Roberts, NurPhoto/Getty Images. 2020; Tyler Prince/214 Digital, 2021.

NATIONAL SECURITY

FBI Report: Bias-Motivated Killings At Record High Amid Nationwide Rise In Hate Crime

November 16, 2020 · 5:39 PM ET

Hate crimes in US reach highest level in more than a decade

By MICHAEL BALSAMO November 16, 2020

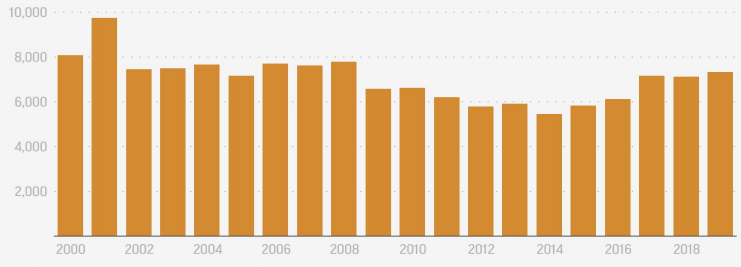
FOR IMMEDIATE RELEASE Wednesday, March 31, 2021

Maryland U.S. Attorney's Office and FBI Baltimore Field Office Condemn Acts of Violence and Discrimination Against Asian Americans and Pacific Islanders, Urge Reporting of Hate Crimes

Hate crimes rise in US

Incidents of hate crimes in the United States – including vandalism, intimidation, assault and murder – rose in 2019 to levels not seen in over a decade, according to federal data.

Incidents reported to the FBI since 2000:

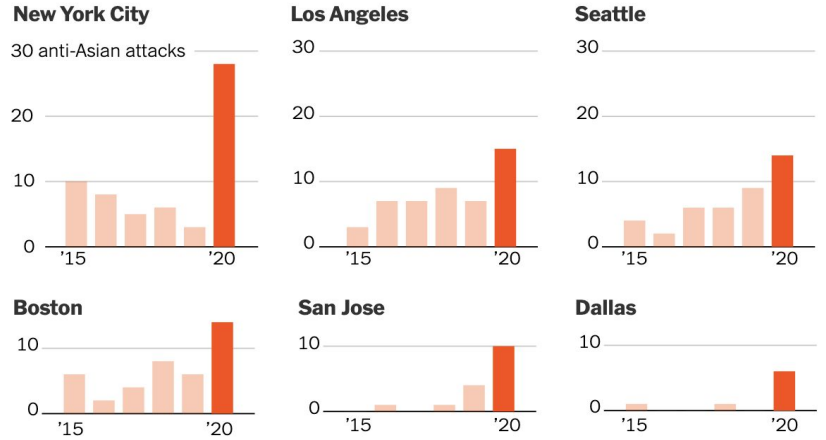


Includes incidents based on a victim's race, religion, sexual orientation or other factors.

Source: FBI

AP

Anti-Asian hate crimes reported to the police



Source: B. Levin, K. Grisham, A. Venolia and G. Levin, Center for the Study of Hate and Extremism at California State University, San Bernardino ▪ Note: Data is based on each city's police departments. Los Angeles data is for only the Los Angeles Police Department. The 2015 number for San Jose is not available.

Michael Balsamo, "Hate crimes in US reach highest level in more than a decade," *AP*, 16 November 2020; Hannah Allam, "FBI Report: Bias-Motivated Killings At Record High Amid Nationwide Rise In Hate Crime," *NPR*, 16 November 2020; "Maryland U.S. Attorney's Office and FBI Baltimore Field Office Condemn[...]," The United States Attorney's Office, District of Maryland, 31 March 2021.

Policy Context

Many countries' anti-radicalization and countering violent extremism (CVE) policies use education as a focal point.

Some of that is about **upward socioeconomic mobility** through education.

Some is about **educating people about specific tools**, ie. making sure citizens are aware of the ways extremist groups use the internet to radicalize.

Suggestions that **open dialogue** may frustrate recruitment and radicalization.

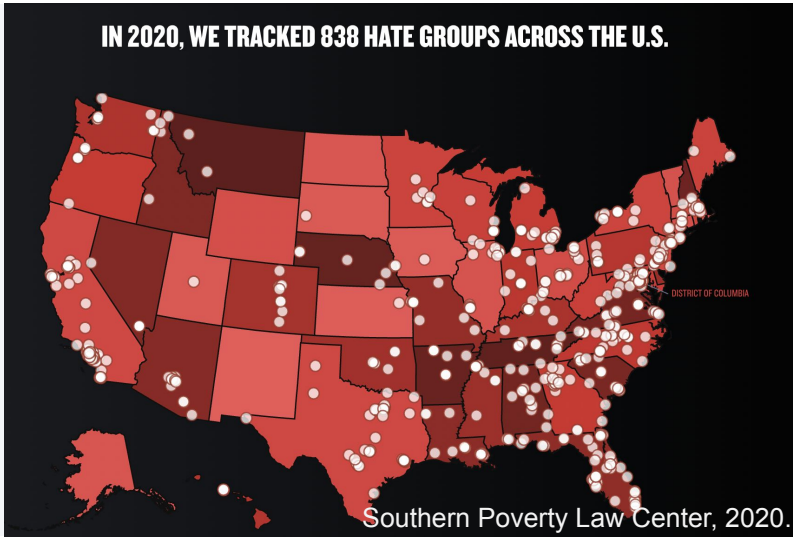
Getty/Nathan Howard, 2020.

The Southern Poverty Law Center (SPLC)

List of organizations designated by the Southern Poverty Law Center as hate groups

From Wikipedia, the free encyclopedia

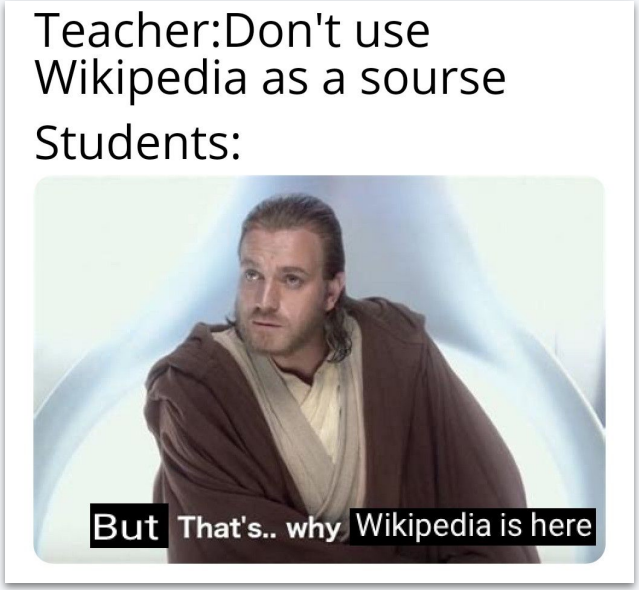
- SPLC is a non-profit legal advocacy organization that mostly deals in civil rights and public-interest law.
- They maintain case files, datasets, and a number of other resources on extremist and hate groups active in the United States.



2.1 Data Source: Wikipedia

About Wikipedia

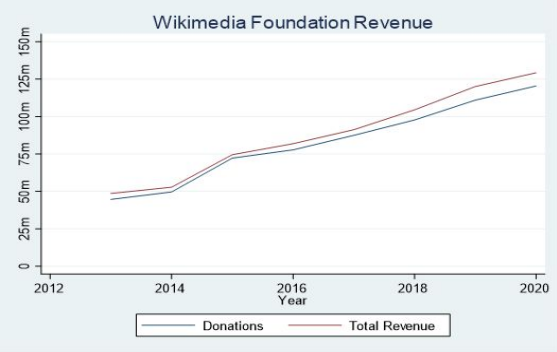
1. Wikipedia was launched on **January 15, 2001**, by Jimmy Wales and Larry Sanger; Sanger coined its name as a portmanteau of **"wiki" and "encyclopedia"**
2. **Wikipedia** is a free, multilingual open-collaborative online encyclopedia
3. The English Wikipedia, with **6.3 million articles** as of April 2021, is the largest of the 321 language editions.
4. **Created and maintained** by a community of volunteer contributors using a wiki-based editing system.
5. The **largest general reference** work on the Internet.

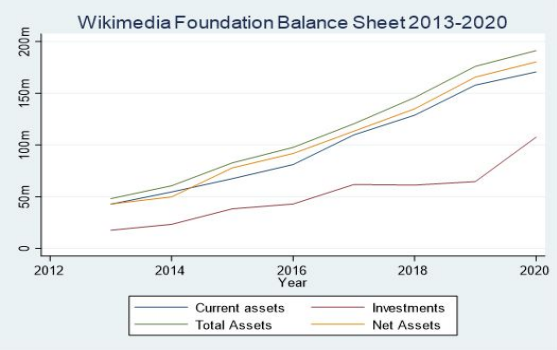


Wikipedia Editors

1. Majority of them are **male**, living in **global north.**
2. Many people **volunteer** to share their knowledge.
3. Counties with **high religious** adherence have a **low level of Wikipedia-editing activity** & counties with **low religious** adherence have **high levels of editing**.
4. **Older articles** from the early years of Wikipedia **leaned Democratic**, whereas those created more **recently** held more **balance**.
5. Articles did not change their bias significantly due to revision, but rather that over time newer articles containing opposite points of view were responsible for **centering the average overall**.

Wikipedia Funding





Wikimedia endowment, est. 2016
$62m ($100m target to 2026)

$5 million+

Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin

$2 million+

Amazon
Google.org
Musk Foundation
George Soros

$1 million+

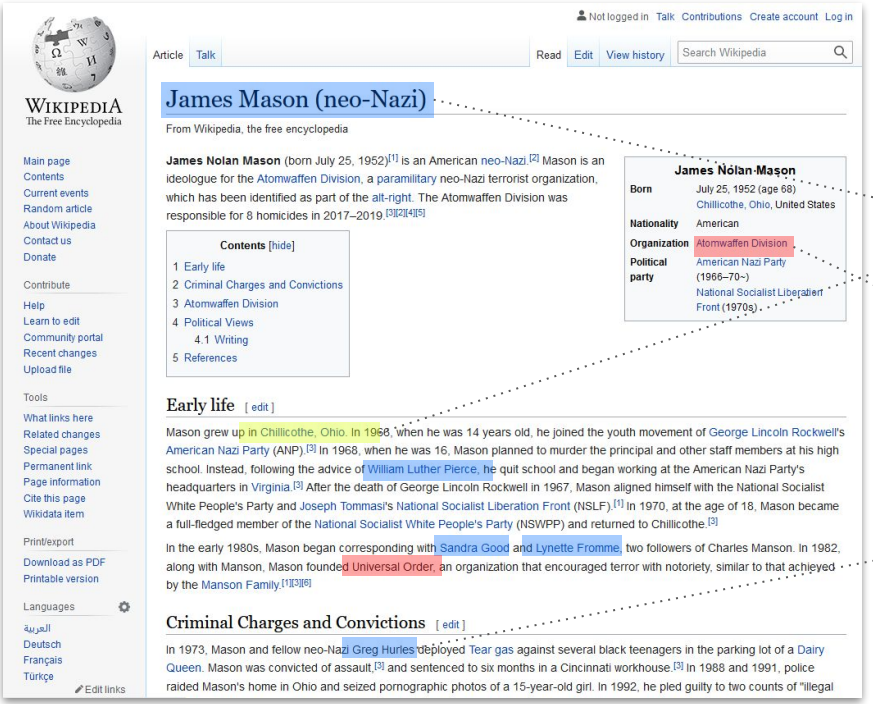
Craig Newmark Philanthropies
Facebook
Jim Pacha*

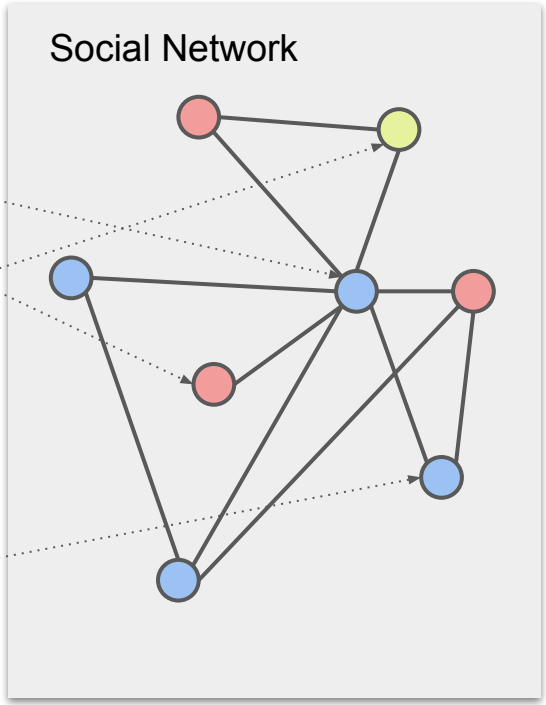
Problems and Confounding Variables

1. **Data scope** - articles only or all revisions of all articles?
2. **Historical figures** - filtering historic figures, entity recognition of date of birth?
3. **Edit bias** - editing rules, accuracy contingent on the editors
4. **Selection bias** - due to data availability (luckily, this is specifically what we want to study)
5. **Equivalence fallacy** - articles on living people or active groups are more active than those on deceased people or inactive groups
6. **Edit history** - limitations in measuring growth, works for small projects, but large projects may reach a saturation point

2.2 Data Extraction

Turning Wikipedia Data into a Social Network





Edge / Mentioning

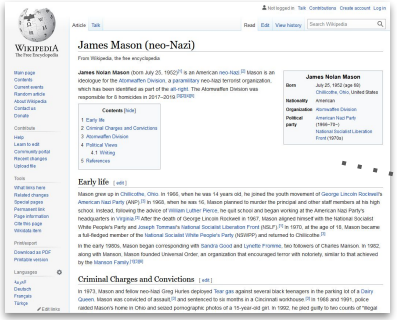
Person

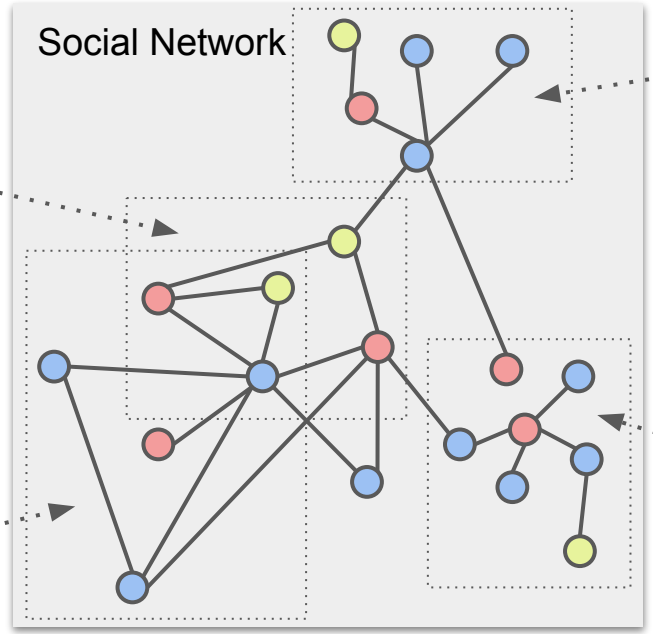
Organization

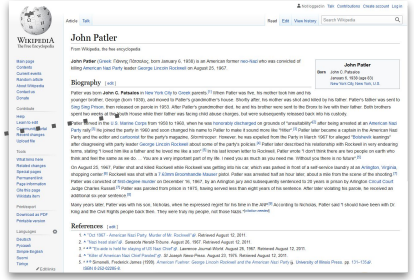
Place

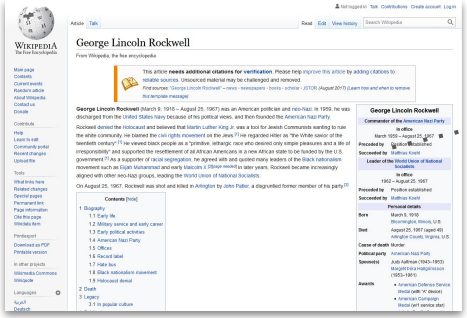
...

Plotting a Network from Multiple Articles









Wikipedia: API, HTML parsing and SNA visualization

API Search

- Identify relevant articles with keyword or category search
- Prompts pageid to parse content

API Collect

- Article details (incl. content)
- Revisions (incl. content)
- Page views

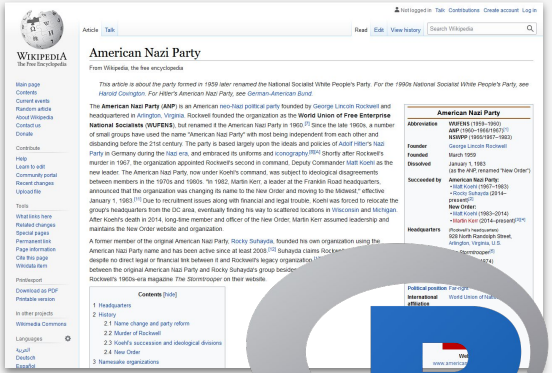
```
wikiscrape_byid = function(pageid){
  url = "https://en.wikipedia.org/w/api.php"

  config = list("action" = "parse",
                "format" = "json",
                "pageid" = pageid,
                "utf8" = 1,
                "prop" = paste("text",
                               "langlinks",
                               "categories",
                               "links",
                               "externallinks",
                               "revid",
                               "displaytitle",
                               "properties",
                               "subtitle",
                               sep="|"))

  res = GET(url = url, query = config)
  res_cont = fromJSON(rawToChar(res$content))
  res_cont = res_cont$parse
}
```

HTML / Wiki-text parser

- Select internal wikipedia links
- Cut out references and side boxes



Network Visualization & Data Processing (NLP, Time Series)

- Social Network Visualization with open-source software Gephi and R/iGraph
- NLP and Times Series analysis with R and/or Python

Gephi
makes graphs handy

Collection Setup for Wikidata

1	**Setup of initial input by search**	Wikipedia article for SPLC hate groups
2	**Collecting first-level nodes w/ article content**	151 organizations (out of ~800)
3	**Parsing links out of article content**	HTML parser for included links
4	**Collecting second-level nodes w/o content**	914 identified persons with 1,209 links
5	**Collecting article edits via revision IDs**	3.4 million observations
6	**Collecting page views for past five years**	2.8 million observations

3. Analysis Methods

3.1 Social Network Analysis (SNA)

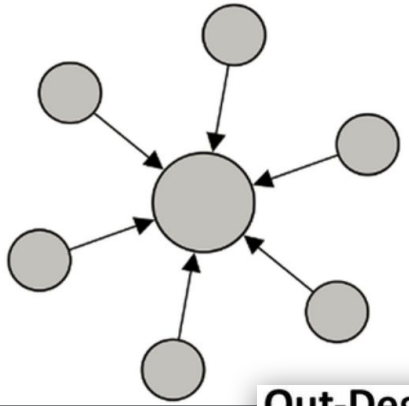
Lead Questions

Who are the central persons to connect U.S. hate groups?

Which groups are central? Which are not?

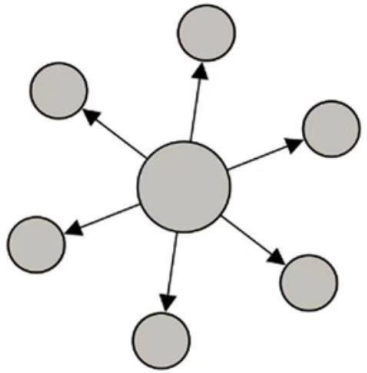
SNA- Directed Graph

In-Degree
'Effect'
Variables with high in-degree are impacted by multiple other variables. An in-degree of 0 means a variable is not influenced by others in the system.



First level data('Organization' in our case):
Only has out-degrees

Out-Degree
'Cause'
High out-degree variables have an ability to change many others in the system. Variables with an out-degree of 0 do not directly influence others.



Second level data('Person' in our case):
Only has in-degrees

SNA - Centrality

- Centrality Measures:
  - Degree Centrality:

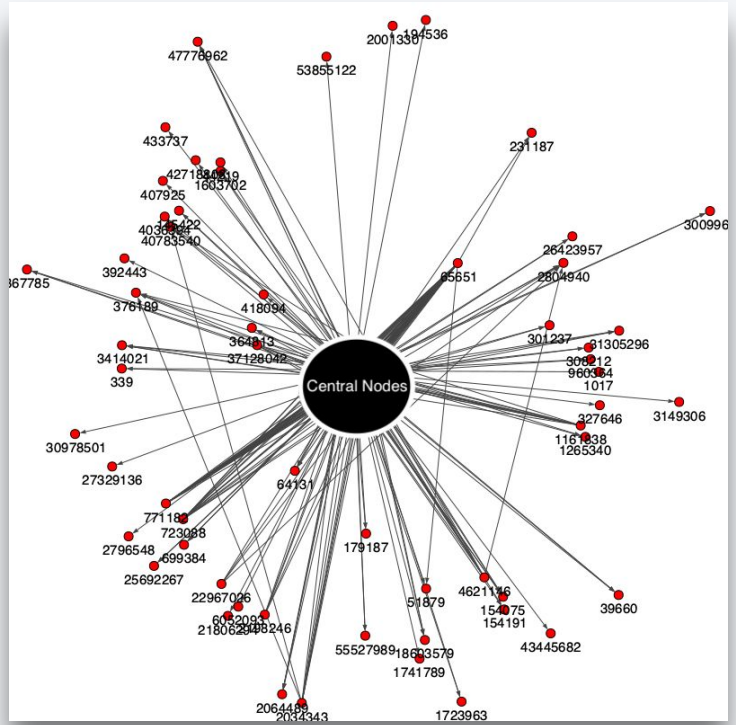
$$C_D(x) = \frac{deg(x)}{N-1}$$

  - Closeness Centrality:

$$C(x) = \left[\sum_{y \neq x}^{N} \frac{d(x,y)}{(N-1)}\right]^{-1}$$

  - Betweenness Centrality:

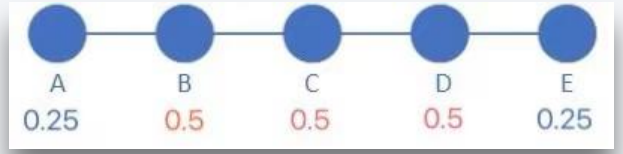
$$B(v) = \sum_{x \neq y} \frac{\sigma(x, y \mid v)/\sigma(x, y)}{(N-1)(N-2)\ /\ 2}$$



SNA- The Math behind it

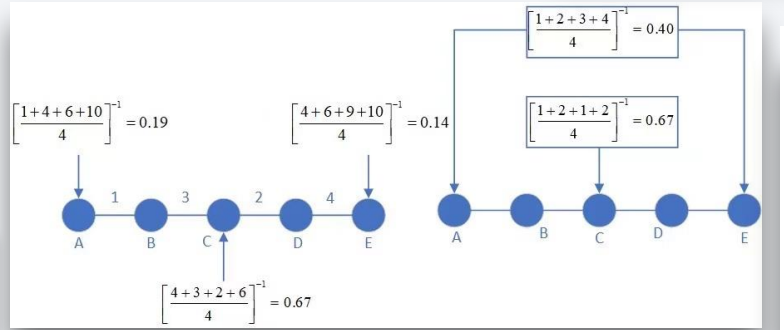
Degree Centrality:

$$C_D(x) = \frac{deg(x)}{N-1}$$



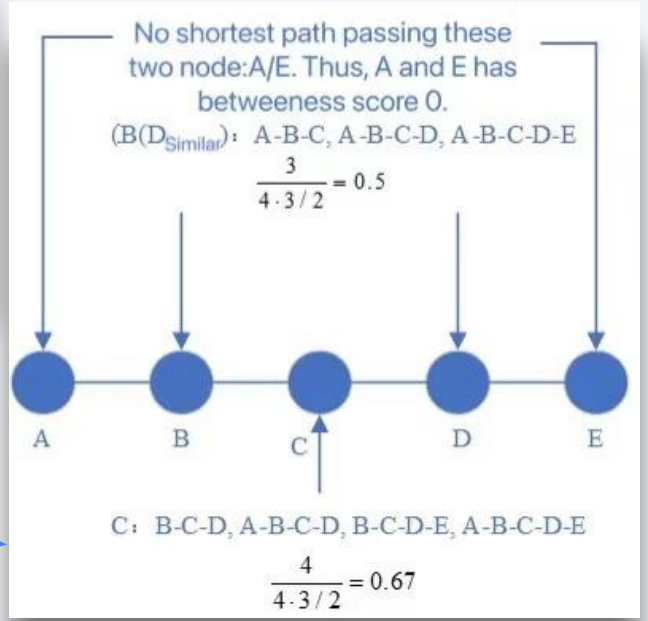
Closeness Centrality:

$$C(x) = \left[\sum_{y \neq x}^{N} \frac{d(x,y)}{(N-1)}\right]^{-1}$$



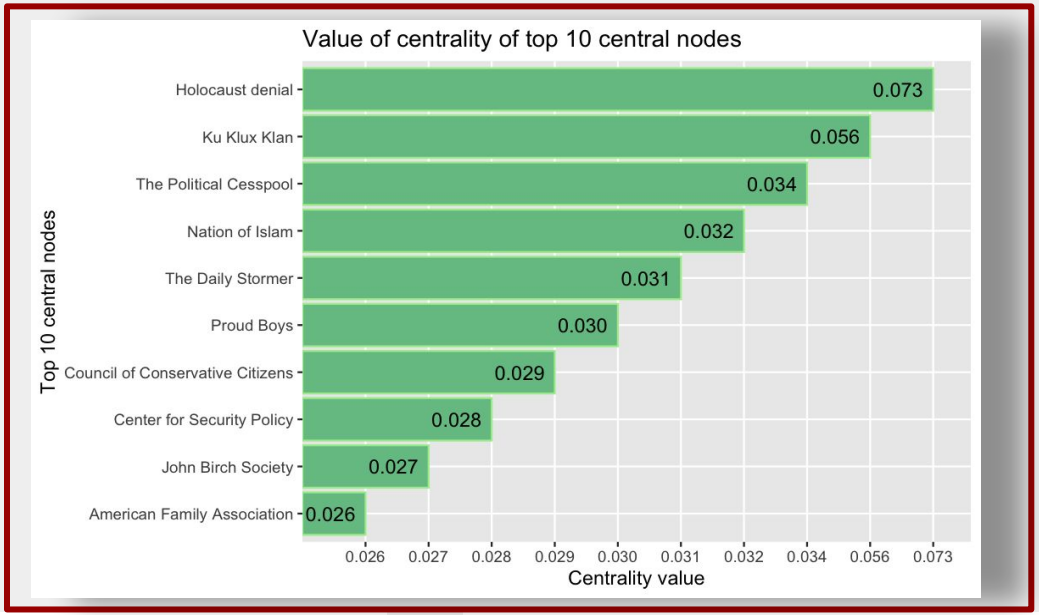
Betweenness Centrality:

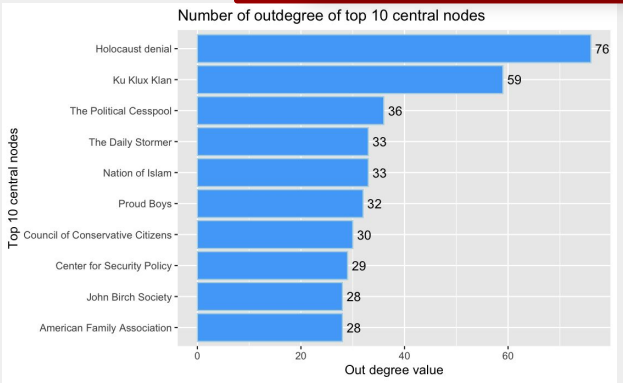
$$B(v) = \sum_{x \neq y} \frac{\sigma(x, y \mid v)/\sigma(x, y)}{(N-1)(N-2)\,/\,2}$$



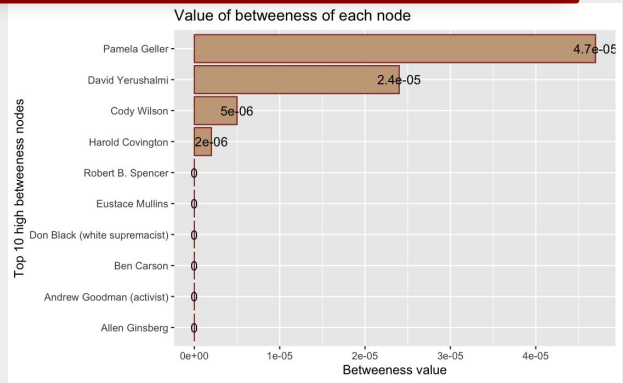
SNA_main.py
class Jgraph(object):
def __init__(self, matrix)
'''INPUT a matrix(Data
self.matrix = matrix
self.__directed = Fals
self.shape = self.matr
self.name = self.matri
self.__check()
self.file_path = os.pa
self.shortest_file_pat
self.g = igraph.Graph(
self.visual_style = {"
self.__graph()
self.__get_centrality(
def __check(self):
for i in range(self.sh
if self.matrix.ilo
warnings.warn(
Jgraph › __init__()
SNA_main
degree_centrality:
{'65651': 0.07272727272727272, '1
closeness_centrality_without_weigh
{'1108056': 1.0, '12885889': 1.0,
closeness_centrality:
{'12885889': 1.0, '13006875': 1.0
betweenness_centrality_without_wei
{'28456749': 9.532713706942382e-0
betweenness_centrality:
{'28456749': 9.532713706942382e-0
Process finished with exit code 0
0.5, 'vertex_label_dist': 2}
5573891': 0.031578947368421054, '476021
: 1.0, '25039291': 1.0, '2574929': 1.0,
: 1.0, '29547715': 1.0, '31672403': 1.0,
88239933e-06, '1001251': 0.0, '10079827'
488239933e-06, '1001251': 0.0, '10079827'
Run it on Python
MMMMMM..
SMTHG
SERIOUS IS
HAPPENING!
mpare("5.2", PHP_VERSION,
5.2 or greater is required!!!");
n_loaded("pc
SysInfo requi
erly.");
php in order
inc.php';
e APP_ROOT.'/
figuration
e APP_ROOT.
ed('PSI_CONFIG_FILE') || !defined('PSI_DEBUG'))
new Template("/templates/html/error_config.html
tpl->fetch();

Top 10 High Centrality Nodes

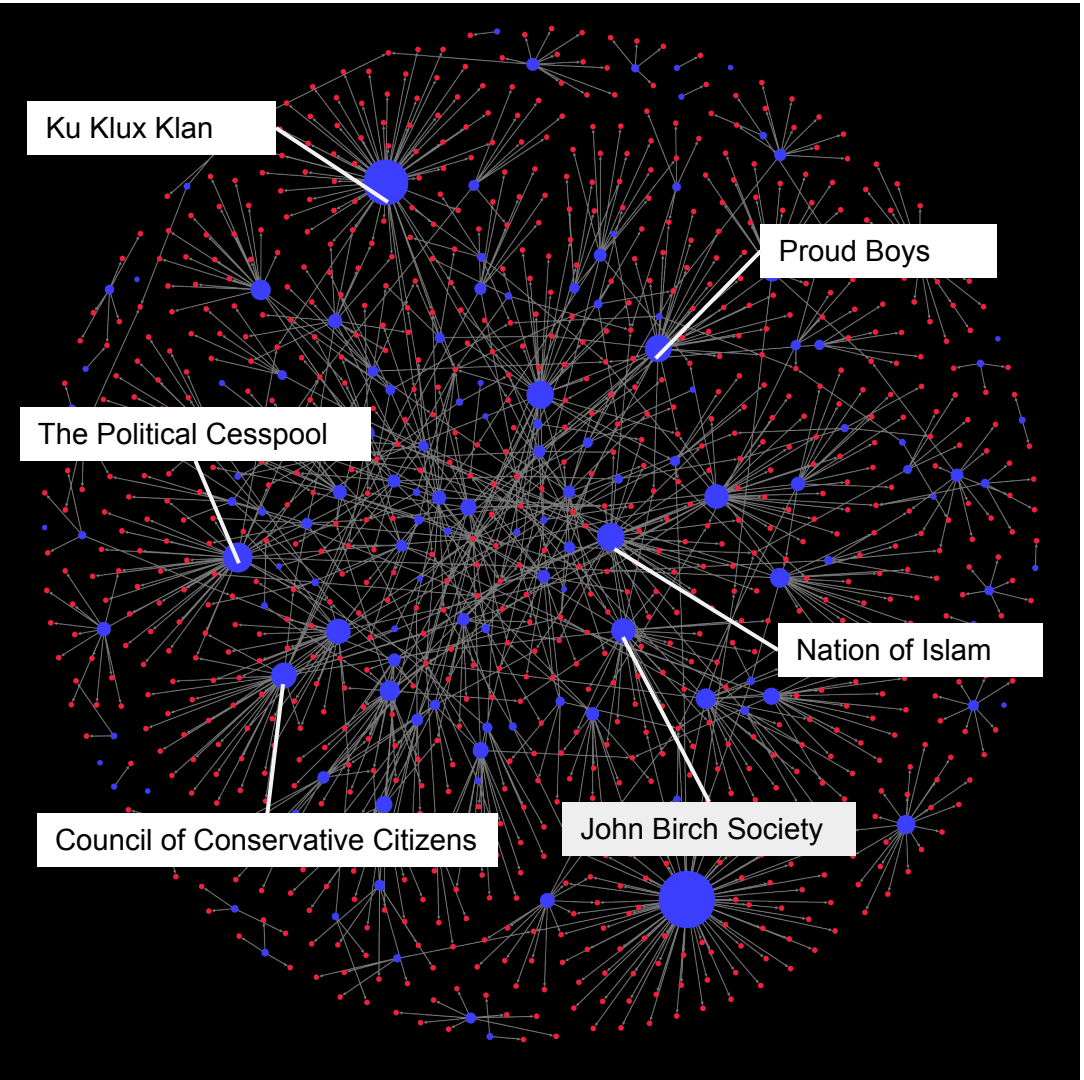




Top 10 High Out-degree Nodes

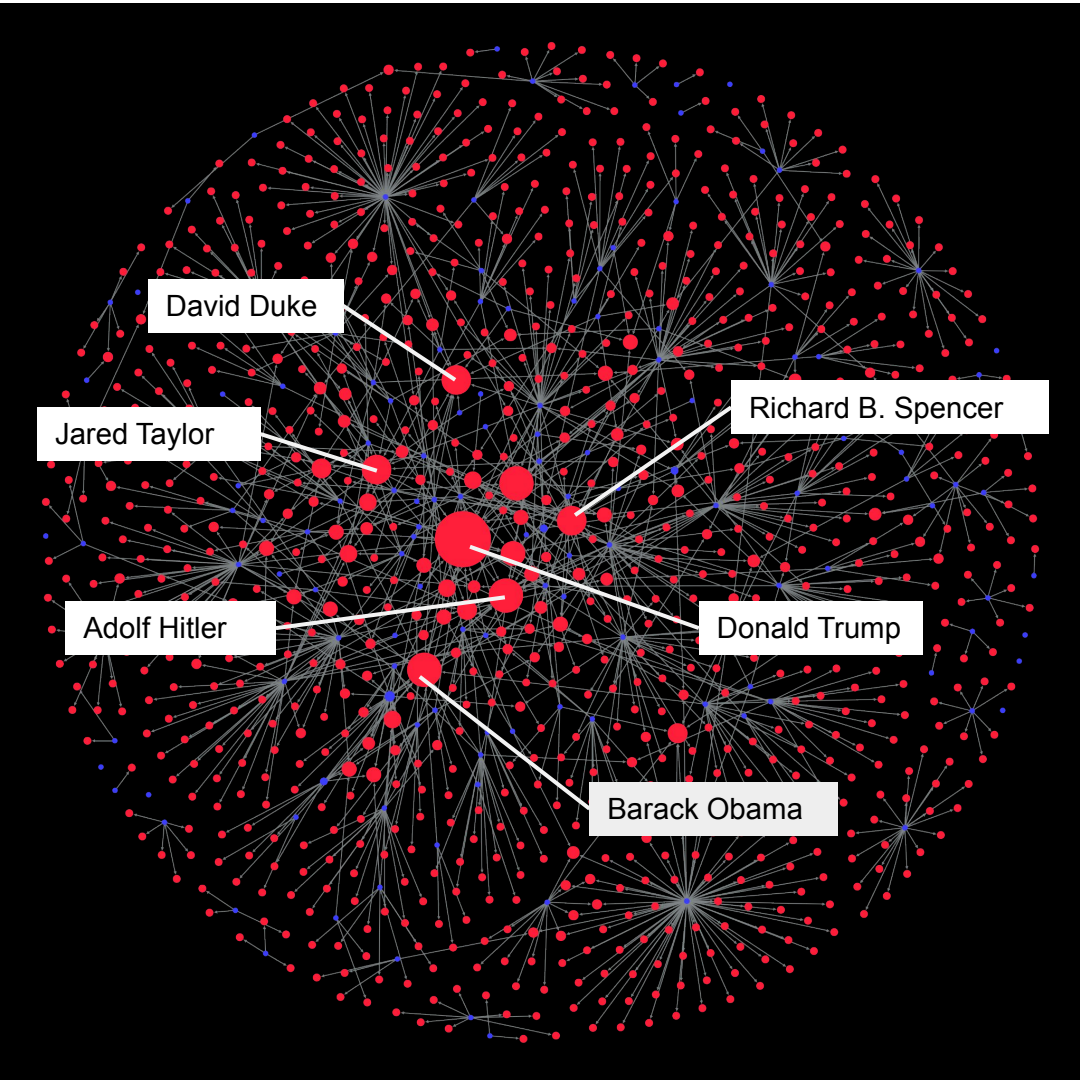


Top 10 High Betweenness Nodes



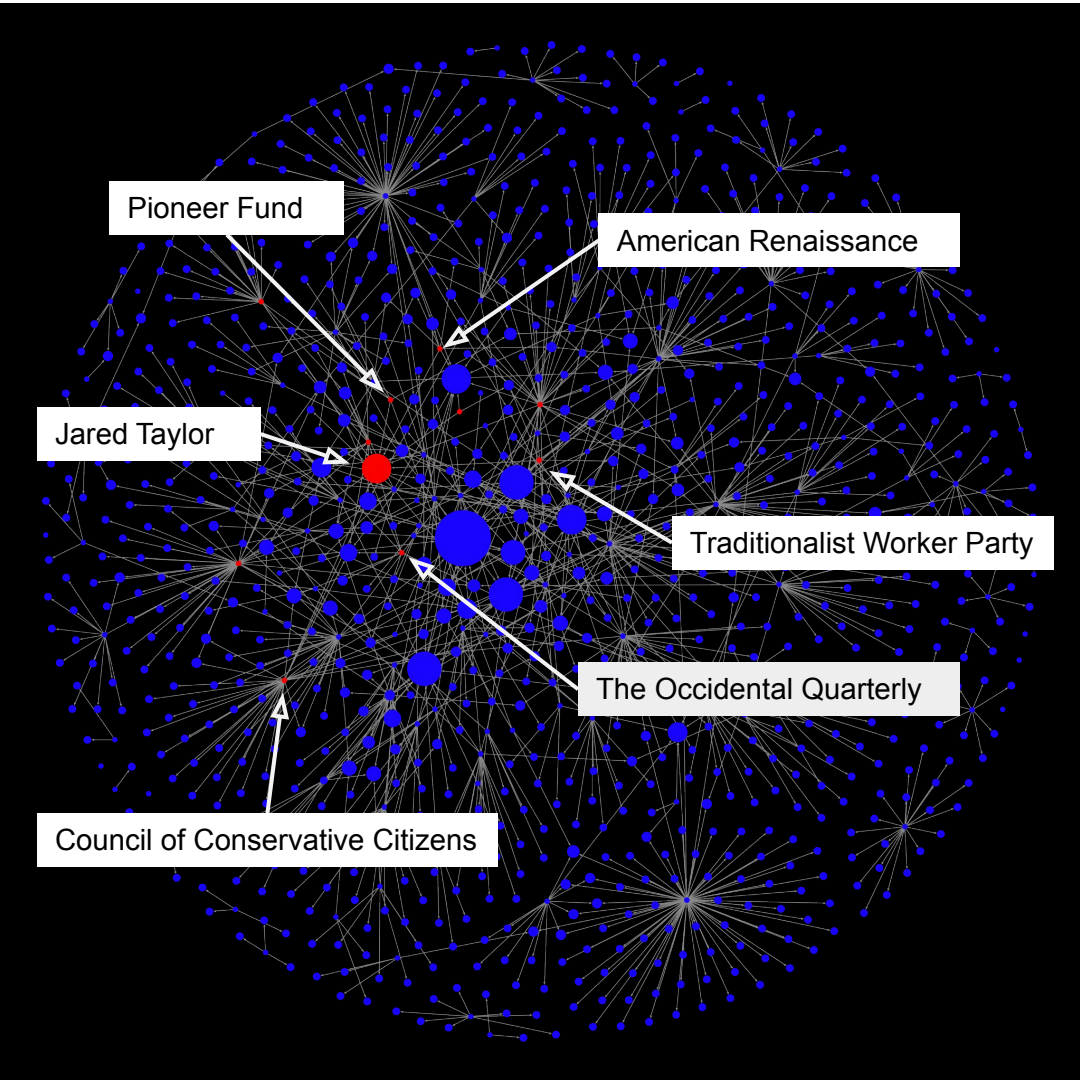
Network by Out-Degree

- Node size by out-degree
  (Organizations referencing Persons)
- Color coding:
  - Blue: Organizations
  - Red: Persons
- More central and indirectly connected organizations in the center



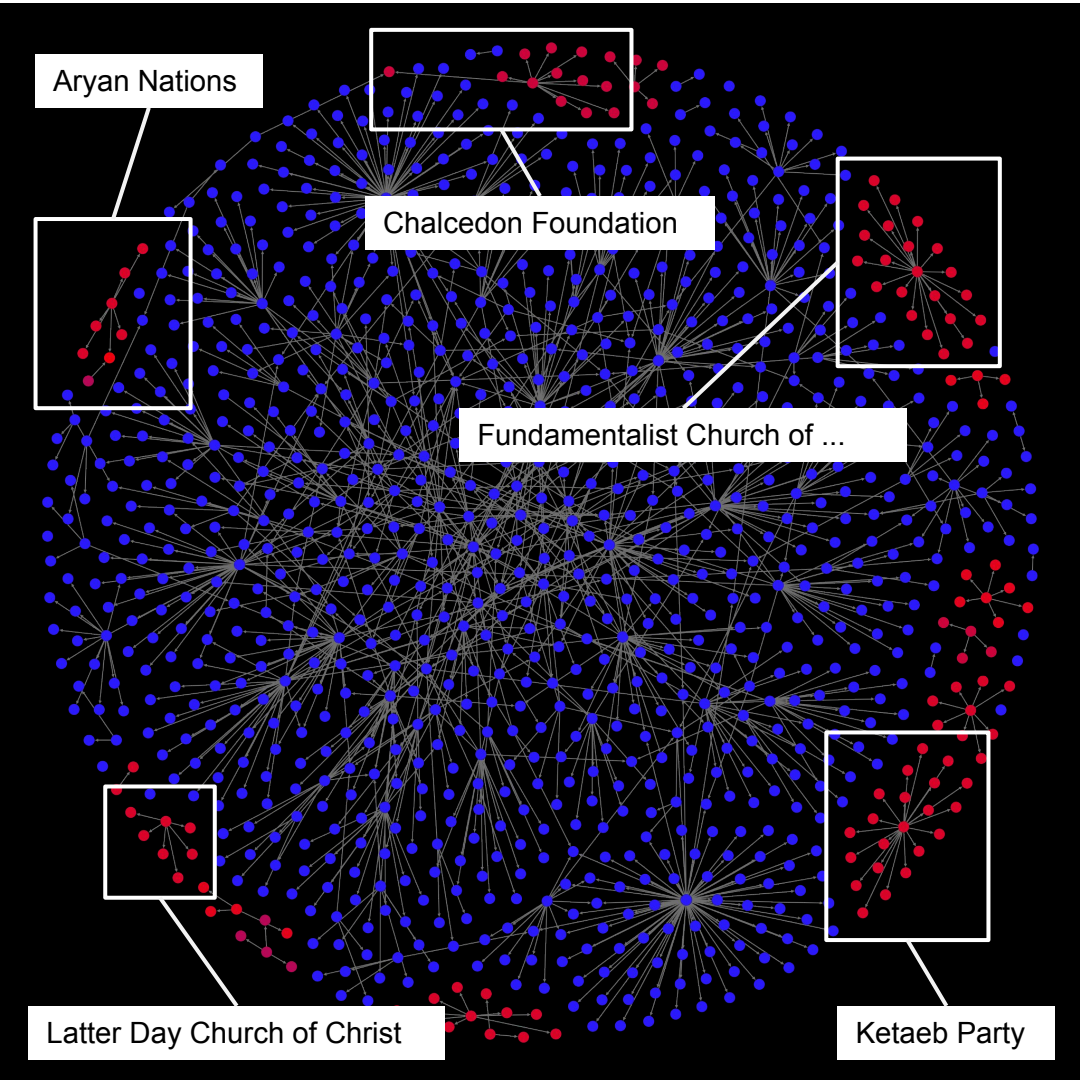
Network by In-Degree

- Node size by in-degree (incoming edges from organizations)
- Color coding:
  - Blue: Organizations
  - Red: Persons
- Central persons in hate groups discourse



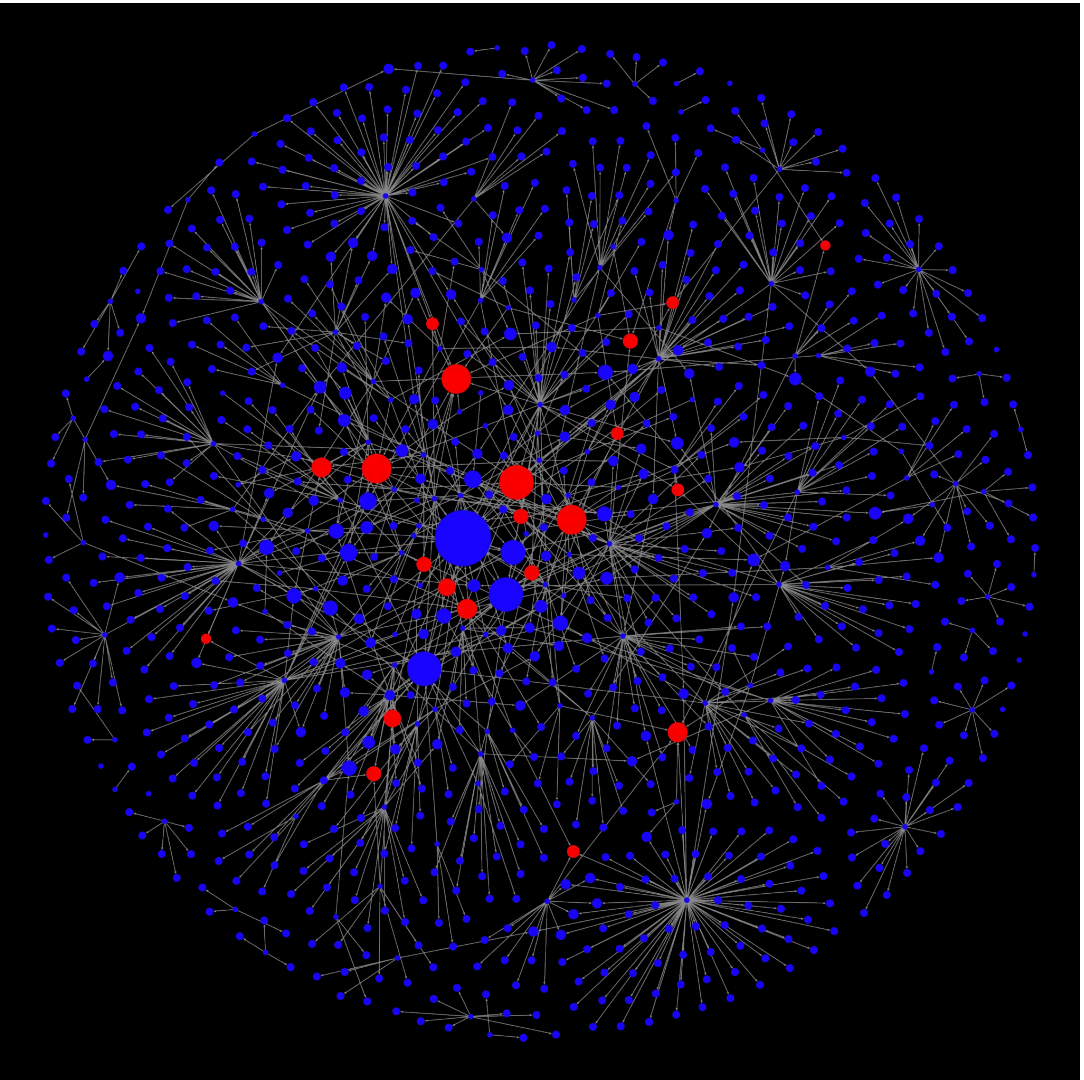
Focus: Jared Taylor

- *1951, Yale / SciencePo
- American white supremacist
  - White genocide conspiracy theory
  - proponent of scientific racism
  - opposes anti-miscegenation laws
- “Far-Right Thinker”
- Ten organizations referencing Taylor



Focus: Outsider Groups

- **Chalcedon Foundation**: Anti-gay hate group
- **Aryan Nations**: American anti-semitic, neo-Nazi, white supremacist terrorist organization
- **Latter Day Church of Christ / Fundamentalist Church of Jesus Christ of Latter-Day Saint**: Mormon fundamentalist group
- **Ketaeb Party**: Libanese Maronite party (misclassification)



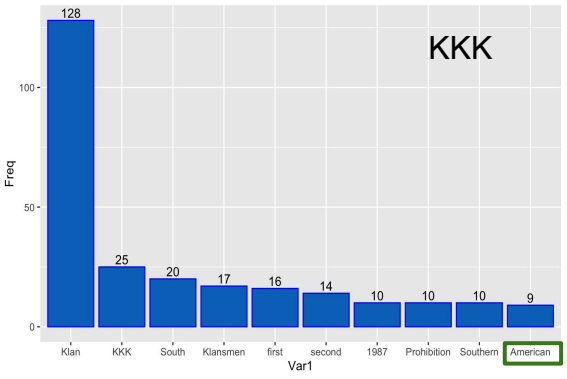
Focus: Neonazis &
White Supremacists

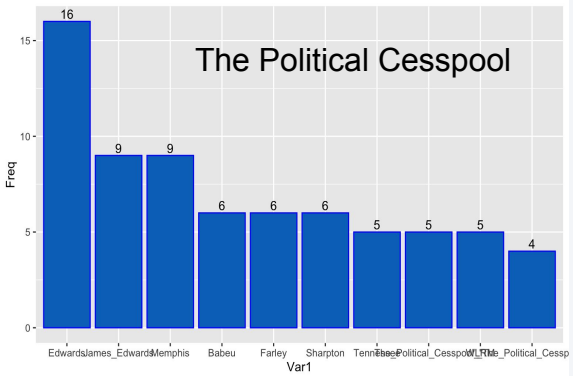
3.2 Natural Language Processing (NLP)

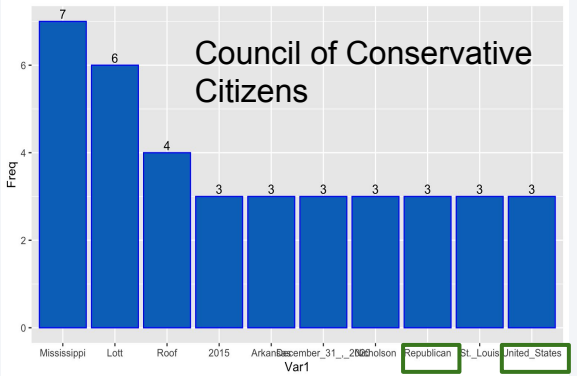
Out-degree Network Analysis

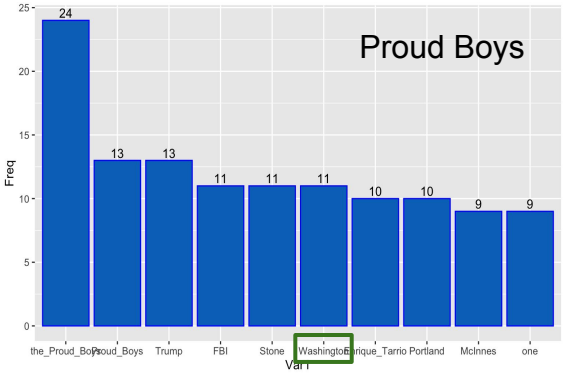
- First level nodes, out-degree measures the number of edges that the vertex of interest points toward.
- Focus on out-degree higher or equal to 28
- Among 1054 label, 6 labels have out-degree higher or equal to 28, and they are:
    - **John Birch Society**, American radical right advocacy group, out degree = 28
    - **Council of Conservative Citizens**, American white supremacist political group, out degree = 30
    - **Proud Boys**, far right and neo-fascist male only organization, out degree = 32
    - **Nation of Islam**, African-America political and religious movement, out degree = 33
    - **The Political Cesspool**, far right radio show in the U.S., out degree = 36
    - **Ku Klux Klan**, American white supremacist group, out degree = 59
- Connections & overlappings between them

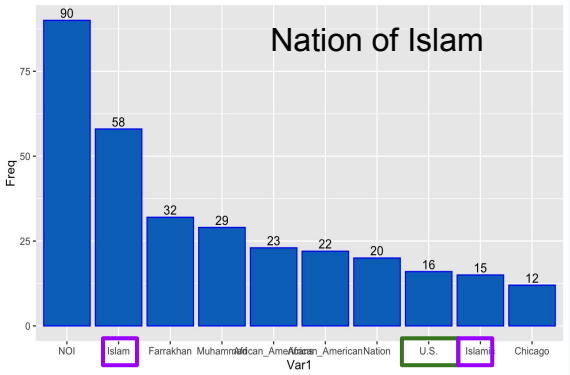
Out-degree Network Analysis

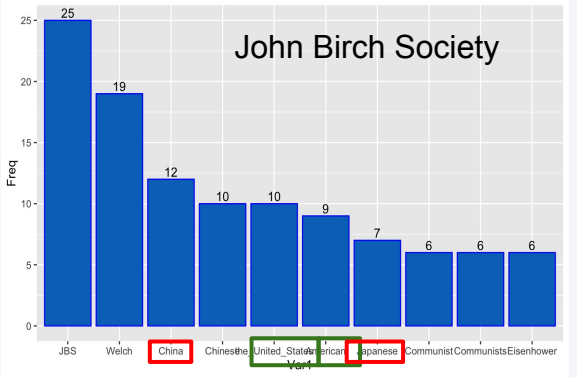












: United States

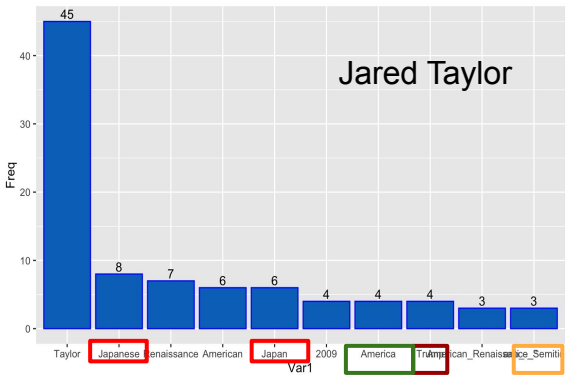
: Islam

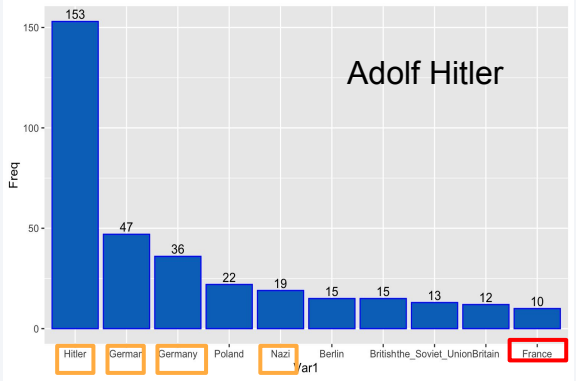
: Asian Countries / China, Japan

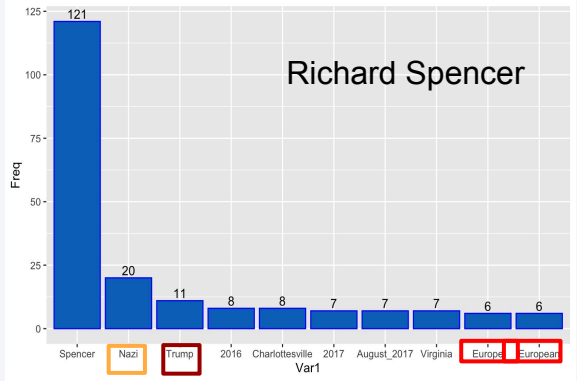
In-degree Network Analysis

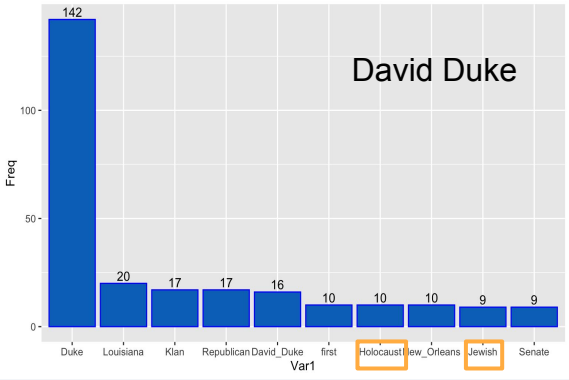
- Second level nodes, in-degree measures the number of edges that point toward the vertex of interest
- Focus on in-degree higher or equal to 10
- Among 1054 label, 5 labels have in-degree higher or equal to 10, and they are:
    - **Jared Taylor**, American white supremacist author, in degree = 10
    - **Richard Spencer**, American white supremacist, in degree = 10
    - **David Duke**, American white supremacist, convicted felon, KKK grand wizard, in degree = 10
    - **Adolf Hitler**, Leader of Germany from 1934 to 1945, in degree = 12
    - **Barack Obama**, 44th president of the United States, in degree = 12
    - **Donald Trump**, 45th president of the United States, in degree = 2
- Connections & overlappings between them

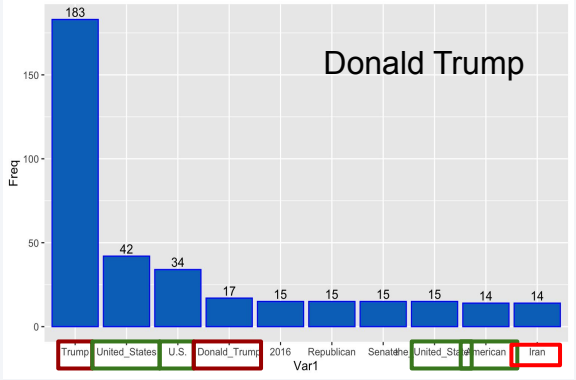
In-degree Network Analysis

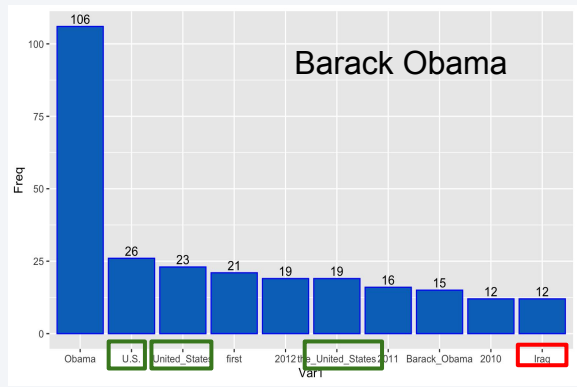












: United States

: Trump

: Germany, Nazi, Jewish

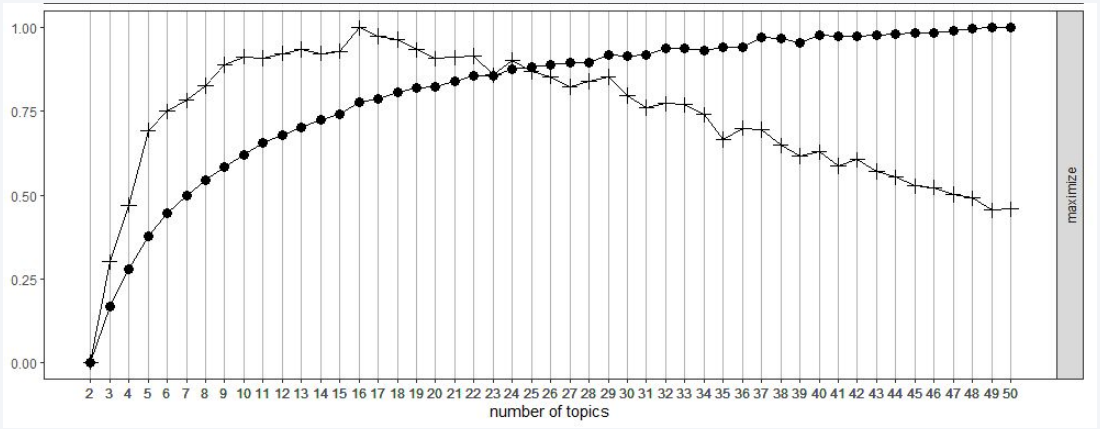
: Asian & European & Middle East Countries

Out-degree Network Analysis **vs.** In-degree Network Analysis

1987, 2015	2009, 2010, 2012, 2016, 2017
• Islam, Islamic • China, Japan	• German, Germany, Holocaust, Jewish, Nazi, Hitler, Anti-Semitic • Japan, Japanese • France, Europe, European • Iraq, Iran
United States, Trump	

1. In in-degree network, subjects are more widely distributed in all aspects.
2. United States, Trump are nodes in both network analysis.

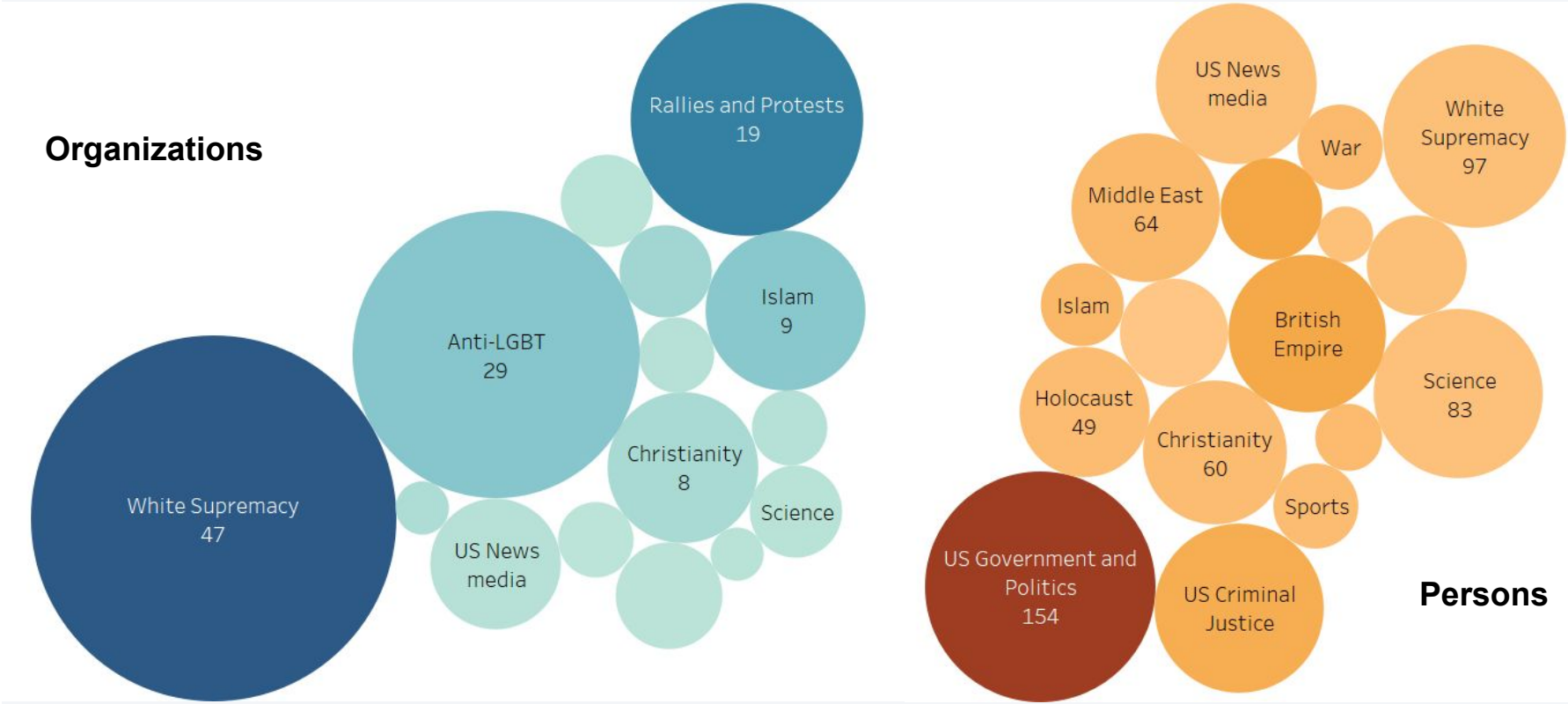
LDA



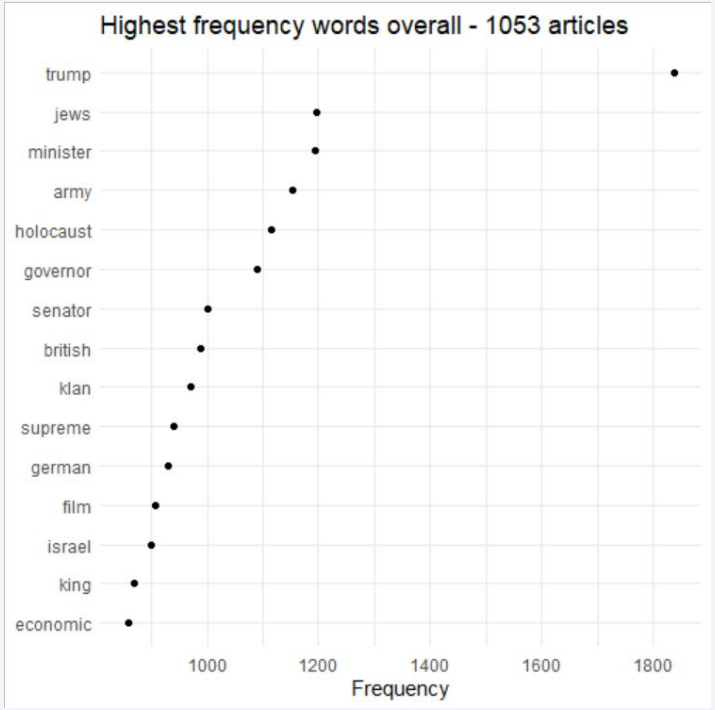
Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
Socialism	US Judicia	Royals	Rallies/Pro	KKK/Jim C	Science	US Politics	US Crimin	War	LGBT	Holocaust	Iraq/Iran	Hillary	Islam	White Sup
soviet	v	harry	rally	klan	theory	voted	guilty	air	gay	holocaust	iraq	trump	islam	immigrati
russian	amendme	san	boys	king	academic	congressic	jury	command	homosexu	german	iran	trump's	muslim	nationalist
russia	circuit	novel	proud	alabama	ideas	republicar	fbi	buckley	homosexu	germany	bin	fbi	muhamm	supremac
communis	lawsuit	anne	video	klux	philosoph	votes	sentenced	battle	baptist	nazi	nations	clinton	muslims	neo-nazi
mao	cases	francisco	far-right	ku	scientific	nominatio	shooting	navy	god	hitler	saddam	russian	malcolm	splc
de	appeals	scientolog	online	tennessee	galileo	legislation	zimmerma	japanese	lgbt	denial	iranian	intelligenc	god	alliance
revolution	kentucky	royal	charlottes	african	argues	bush	arrest	lincoln	sex	historian	un	obama	islamic	aryan
africa	constitutic	art	twitter	carolina	works	nominee	sentence	troops	homosexu	auschwitz	islamic	hillary	elijah	virginia
republic	judicial	duke	site	blacks	psycholog	obama	officers	china	afa	historical	ban	russia	india	group's
china	alabama	stories	alt-right	johnson	century	amendme	car	field	falwell	camps	nuclear	FALSE	followers	nations
property	ruling	jack	protest	georgia	intelligenc	democrats	prosecutic	fort	pastor	nazis	afghanista	wikileaks	noi	nazi
libya	ruled	fiction	violent	kkk	modern	senators	testified	virginia	same-sex	hitler's	intelligenc	allegation	indian	wikipedia
socialist	counsel	novels	portland	atlanta	writings	defeated	florida	aircraft	orientatio	historians	iraqi	emails	african	socialist
revolution	judges	writer	protesters	african-an	taught	reform	shot	reserve	bible	europe	countries	twitter	ibn	immigrant
moscow	liberty	works	banned	confedera	earth	percent	witness	command	lesbian	gas	pakistan	firm	teachings	expanding

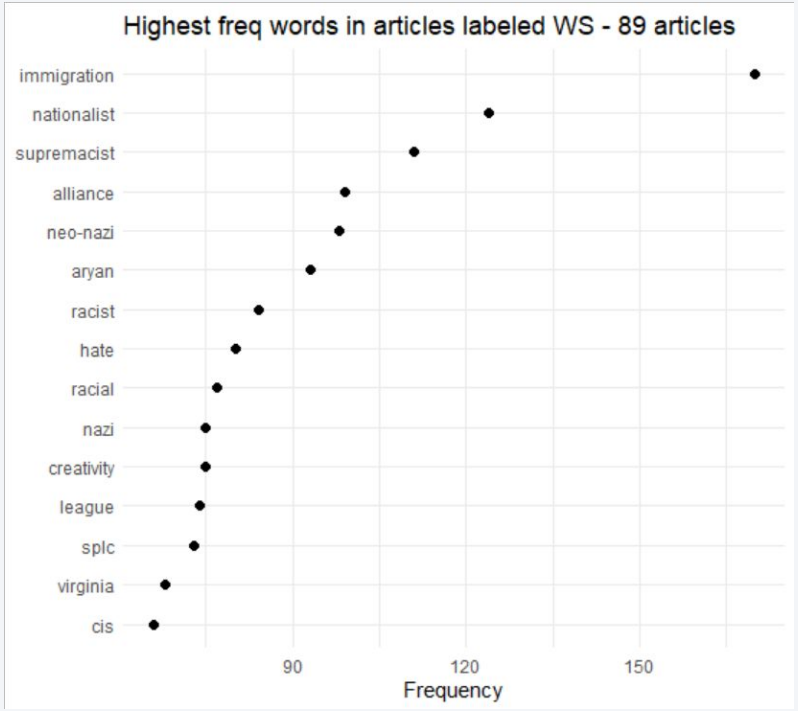
Optimal
K=24
Max Doc Freq = 0.2
Min Term Freq = 0.85
300+ iterations

LDA - by organizations (1st level) and persons (2nd level)

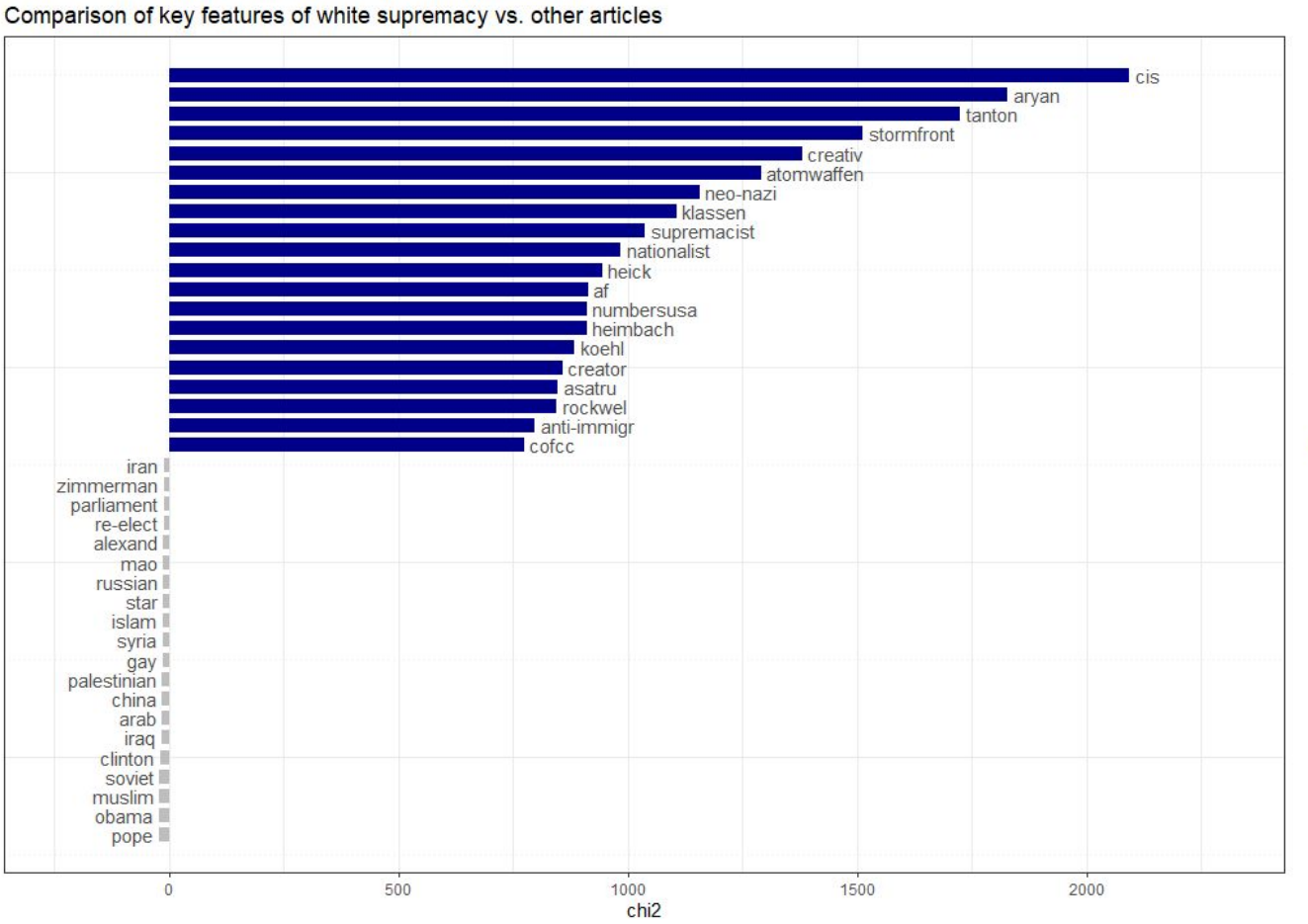


NLP - Highest Frequency Words

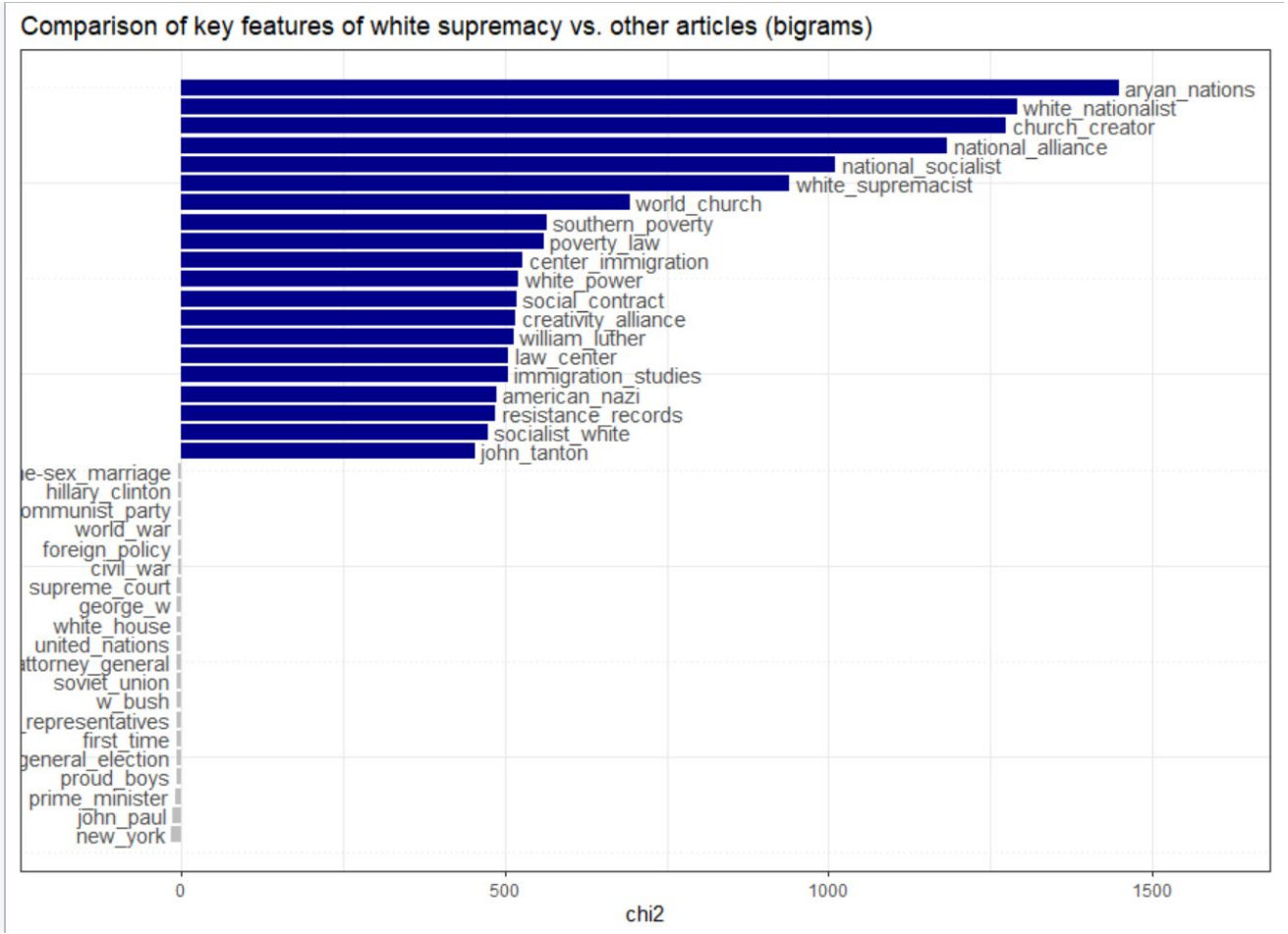




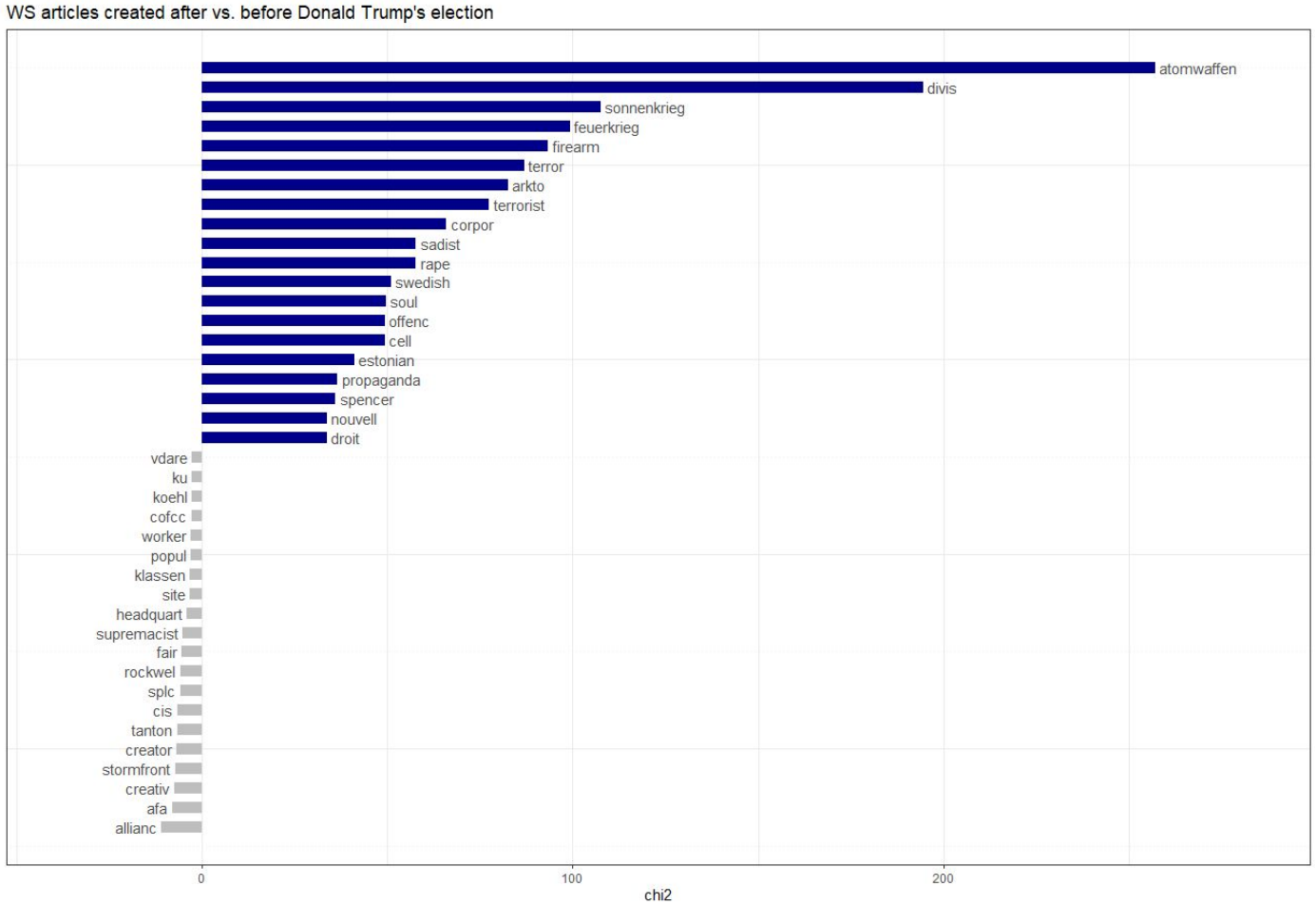
NLP - Keyness

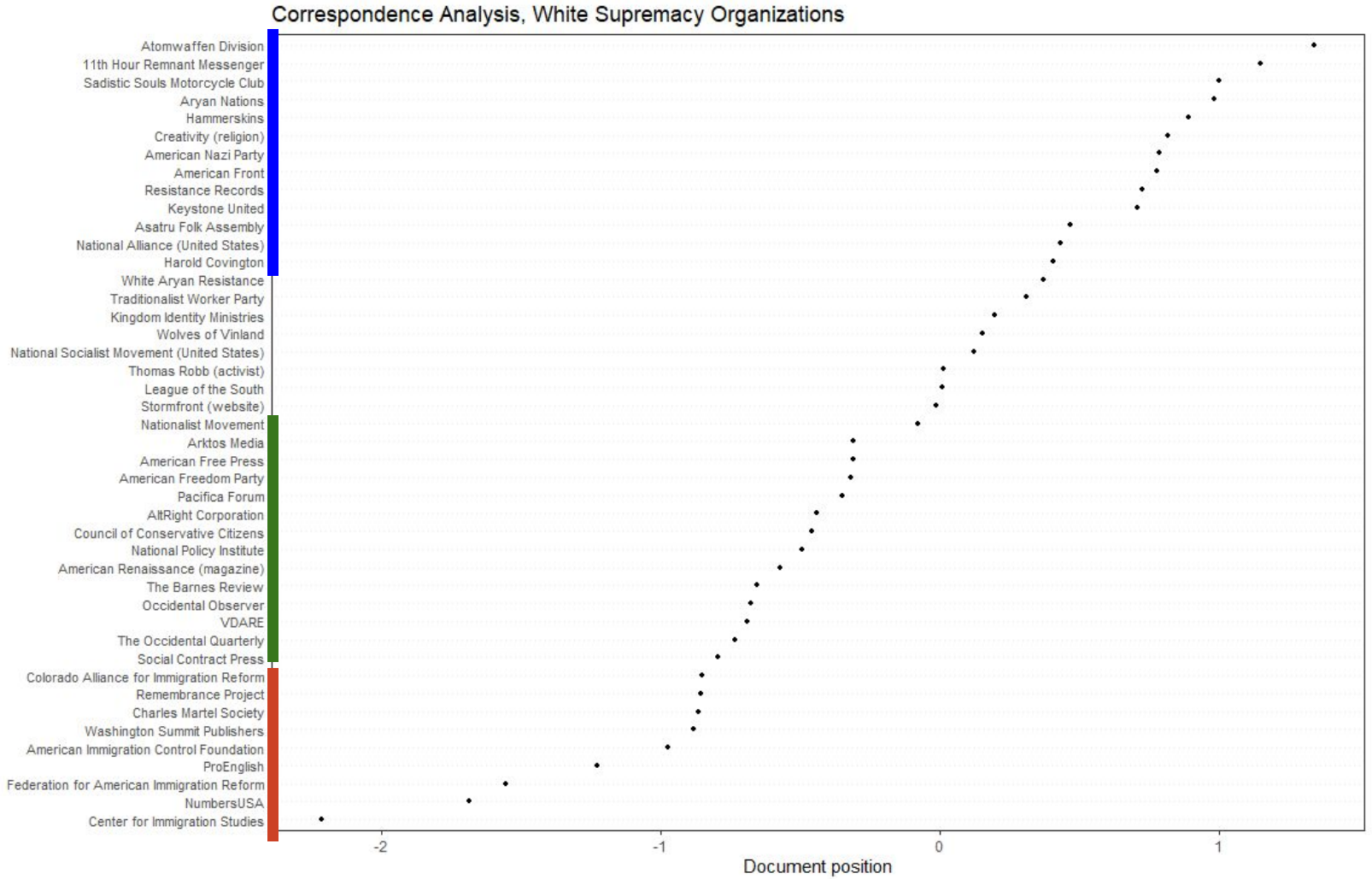

Comparison of key features of white supremacy vs. other articles
cis
aryan
tanton
stormfront
creativ
atomwaffen
neo-nazi
klassen
supremacist
nationalist
heick
af
numbersusa
heimbach
koehl
creator
asatru
rockwel
anti-immigr
cofcc
iran
zimmerman
parliament
re-elect
alexand
mao
russian
star
islam
syria
gay
palestinian
china
arab
iraq
clinton
soviet
muslim
obama
pope
0
500
1000
1500
2000
chi2

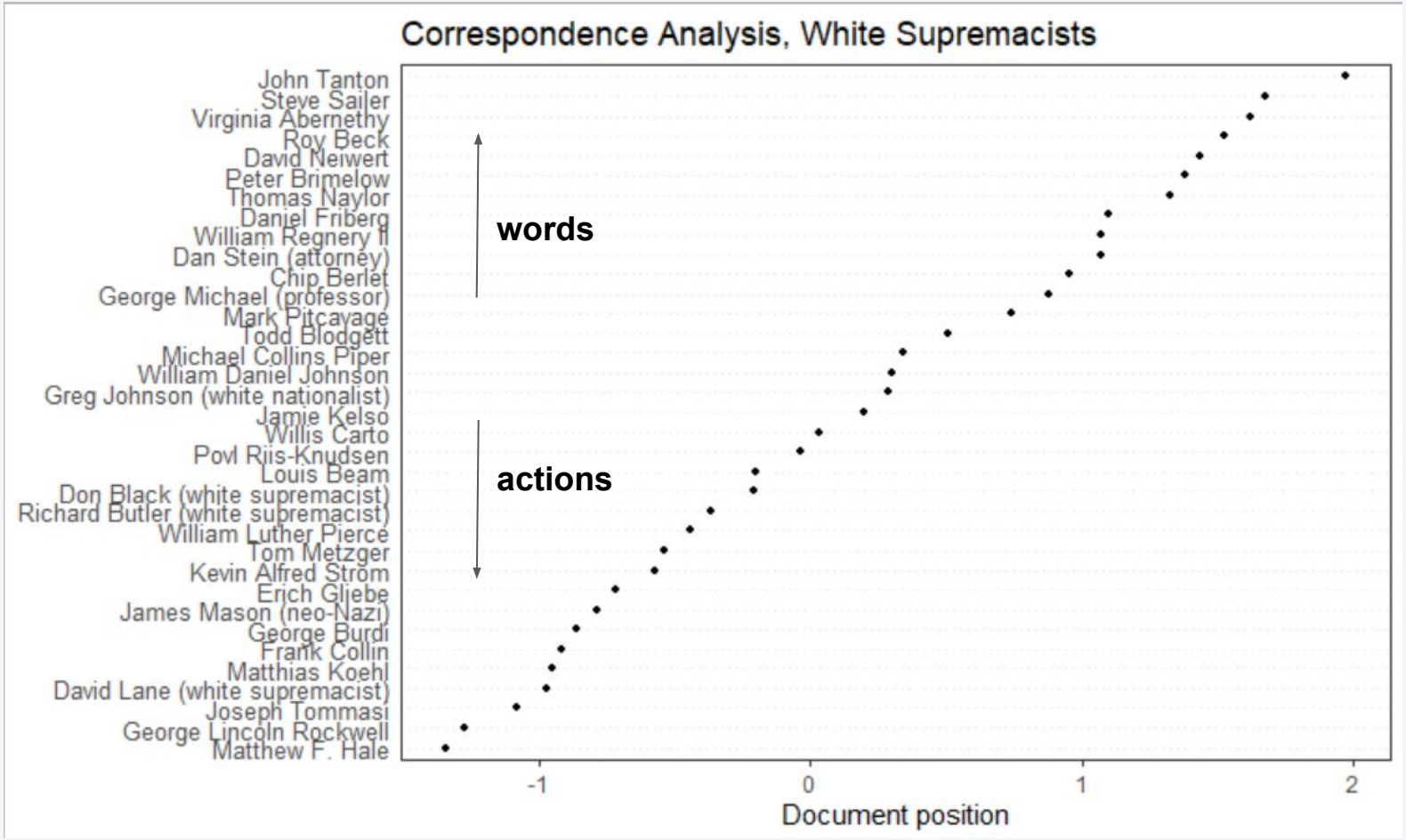
NLP - Keyness


Comparison of key features of white supremacy vs. other articles (bigrams)
aryan_nations
white_nationalist
church_creator
national_alliance
national_socialist
white_supremacist
world_church
southern_poverty
poverty_law
center_immigration
white_power
social_contract
creativity_alliance
william_luther
law_center
immigration_studies
american_nazi
resistance_records
socialist_white
john_tanton
e-sex_marriage
hillary_clinton
ommunist_party
world_war
foreign_policy
civil_war
supreme_court
george_w
white_house
united_nations
ttorney_general
soviet_union
w_bush
representatives
first_time
general_election
proud_boys
prime_minister
john_paul
new_york
0
500
1000
1500
chi2

NLP - Keyness


WS articles created after vs. before Donald Trump's election
atomwaffen
divis
sonnenkrieg
feuerkrieg
firearm
terror
arkto
terrorist
corpor
sadist
rape
swedish
soul
offenc
cell
estonian
propaganda
spencer
nouvell
droit
vdare
ku
koehl
cofcc
worker
popul
klassen
site
headquart
supremacist
fair
rockwel
splc
cis
tanton
creator
stormfront
creativ
afa
allianc
0
100
200
chi2

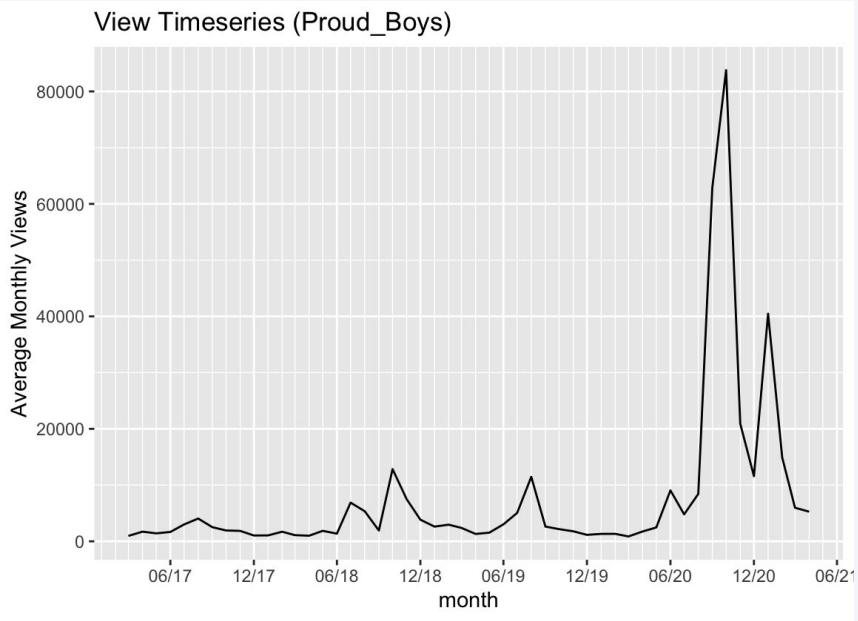

Correspondence Analysis, White Supremacy Organizations
Atomwaffen Division
11th Hour Remnant Messenger
Sadistic Souls Motorcycle Club
Aryan Nations
Hammerskins
Creativity (religion)
American Nazi Party
American Front
Resistance Records
Keystone United
Asatru Folk Assembly
National Alliance (United States)
Harold Covington
White Aryan Resistance
Traditionalist Worker Party
Kingdom Identity Ministries
Wolves of Vinland
National Socialist Movement (United States)
Thomas Robb (activist)
League of the South
Stormfront (website)
Nationalist Movement
Arktos Media
American Free Press
American Freedom Party
Pacifica Forum
AltRight Corporation
Council of Conservative Citizens
National Policy Institute
American Renaissance (magazine)
The Barnes Review
Occidental Observer
VDARE
The Occidental Quarterly
Social Contract Press
Colorado Alliance for Immigration Reform
Remembrance Project
Charles Martel Society
Washington Summit Publishers
American Immigration Control Foundation
ProEnglish
Federation for American Immigration Reform
NumbersUSA
Center for Immigration Studies
-2
-1
0
1
Document position


Correspondence Analysis, White Supremacists
John Tanton
Steve Sailer
Virginia Abernethy
Roy Beck
David Neiwert
Peter Brimelow
Thomas Naylor
Daniel Friberg
William Regnery II
Dan Stein (attorney)
Chip Berlet
George Michael (professor)
Mark Pitcavage
Todd Blodgett
Michael Collins Piper
William Daniel Johnson
Greg Johnson (white nationalist)
Jamie Kelso
Willis Carto
Povl Riis-Knudsen
Louis Beam
Don Black (white supremacist)
Richard Butler (white supremacist)
William Luther Pierce
Tom Metzger
Kevin Alfred Strom
Erich Gliebe
James Mason (neo-Nazi)
George Burdi
Frank Collin
Matthias Koehl
David Lane (white supremacist)
Joseph Tommasi
George Lincoln Rockwell
Matthew F. Hale
words
actions
-1
0
1
2
Document position

3.3 Time Series Analysis

Example: The Proud Boys

- Examining monthly view averages over time to ascertain temporal patterns in public attention
- Clear spike between August 2020 and October 2020; secondary spike in January 2021.



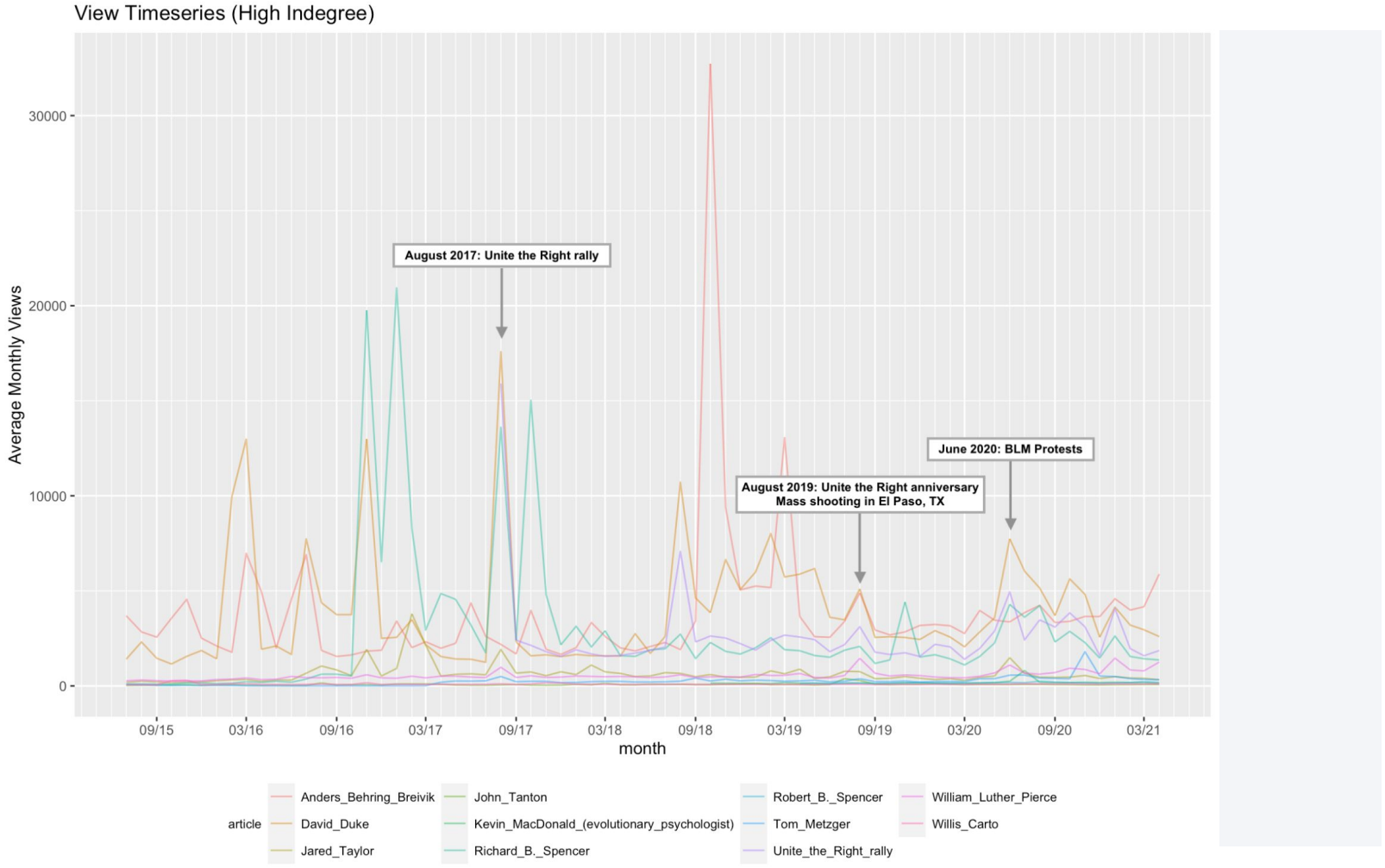
Why might these spikes occur?

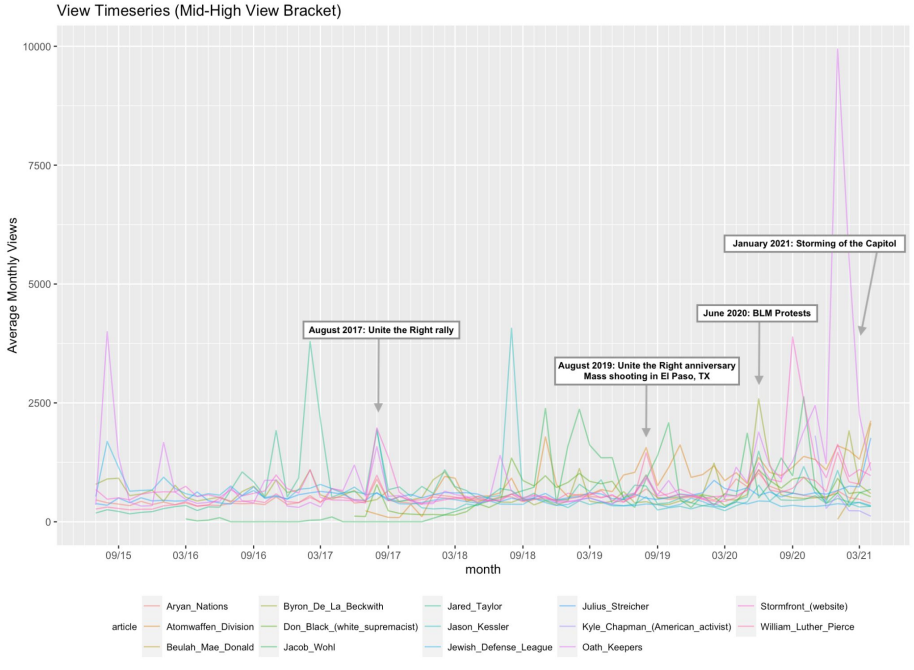
National

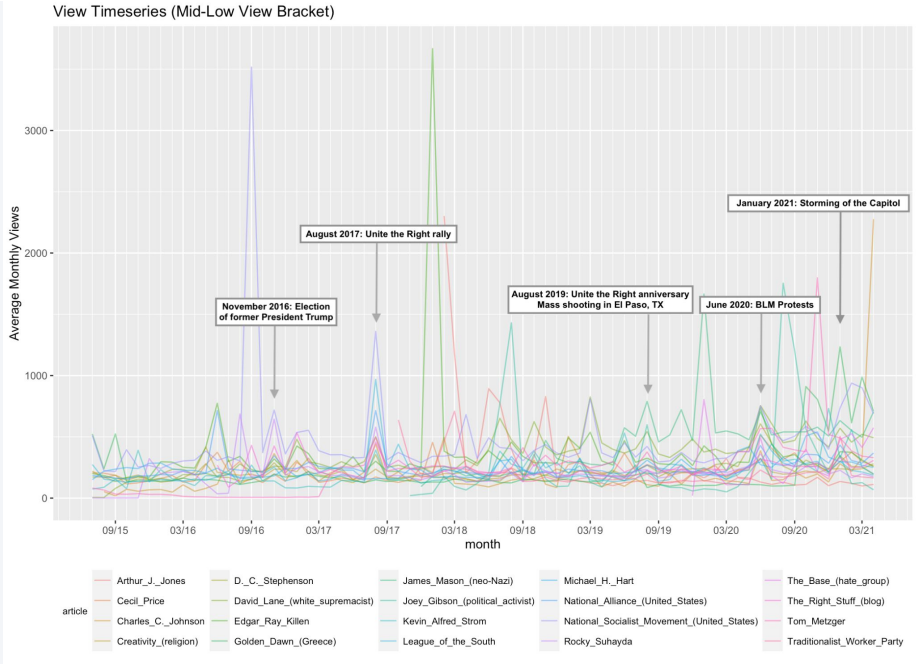
Portland police stand by as Proud Boys and far-right militias flash guns and brawl with antifa counterprotesters

A Day of Protest in Portland as 'Proud Boys' Converge on the City

Jonathan Levinson/OPB, 2020.


View Timeseries (High Indegree)
Average Monthly Views
30000
20000
10000
0
August 2017: Unite the Right rally
August 2019: Unite the Right anniversary
Mass shooting in El Paso, TX
June 2020: BLM Protests
09/15
03/16
09/16
03/17
09/17
03/18
09/18
03/19
09/19
03/20
09/20
03/21
month
article
Anders_Behring_Breivik
David_Duke
Jared_Taylor
John_Tanton
Kevin_MacDonald_(evolutionary_psychologist)
Richard_B._Spencer
Robert_B._Spencer
Tom_Metzger
Unite_the_Right_rally
William_Luther_Pierce
Willis_Carto





- Mid-Low bracket shows a spike in November 2016 around the election.
- All brackets show a dramatic spike in August 2017 around the Unite the Right rally, and again in August 2019.
- All brackets show a dramatic spike in June 2020 around the BLM protests.
- All brackets show a spike or overall increase in views after the January 2021 Storming of the Capitol.

4. Findings

Findings

- Wikipedia data is a helpful resource to better understand public interest and awareness in topics like hate groups
- Wikipedia data benefits from very regular edits of articles but covers only the “surface knowledge” on topics (150 out of 800 hate groups)
- Social network analysis helps to identify common reference between organizations
- The centrality of networks (center vs. periphery) represents interconnectedness of groups in the light of public knowledge (focus: Far-right hate groups, not religious groups)
- “US” and “Trump” are top 2 most frequent words related to Wikipedia network nodes; names of some countries and years are hot subjects around main nodes as well.
- NLP techniques can be used to add “color” identifying groups based on topic and lexical similarity

Findings, cont’d.

- Public attention fluctuates around national news coverage.
- As far-right activities receive news coverage, readers may pursue more information about far-right groups, attacks, and leaders.
- Most of the peaks in public attention around news events were relatively short-lived, with the exception of the Storming of the Capitol in January.

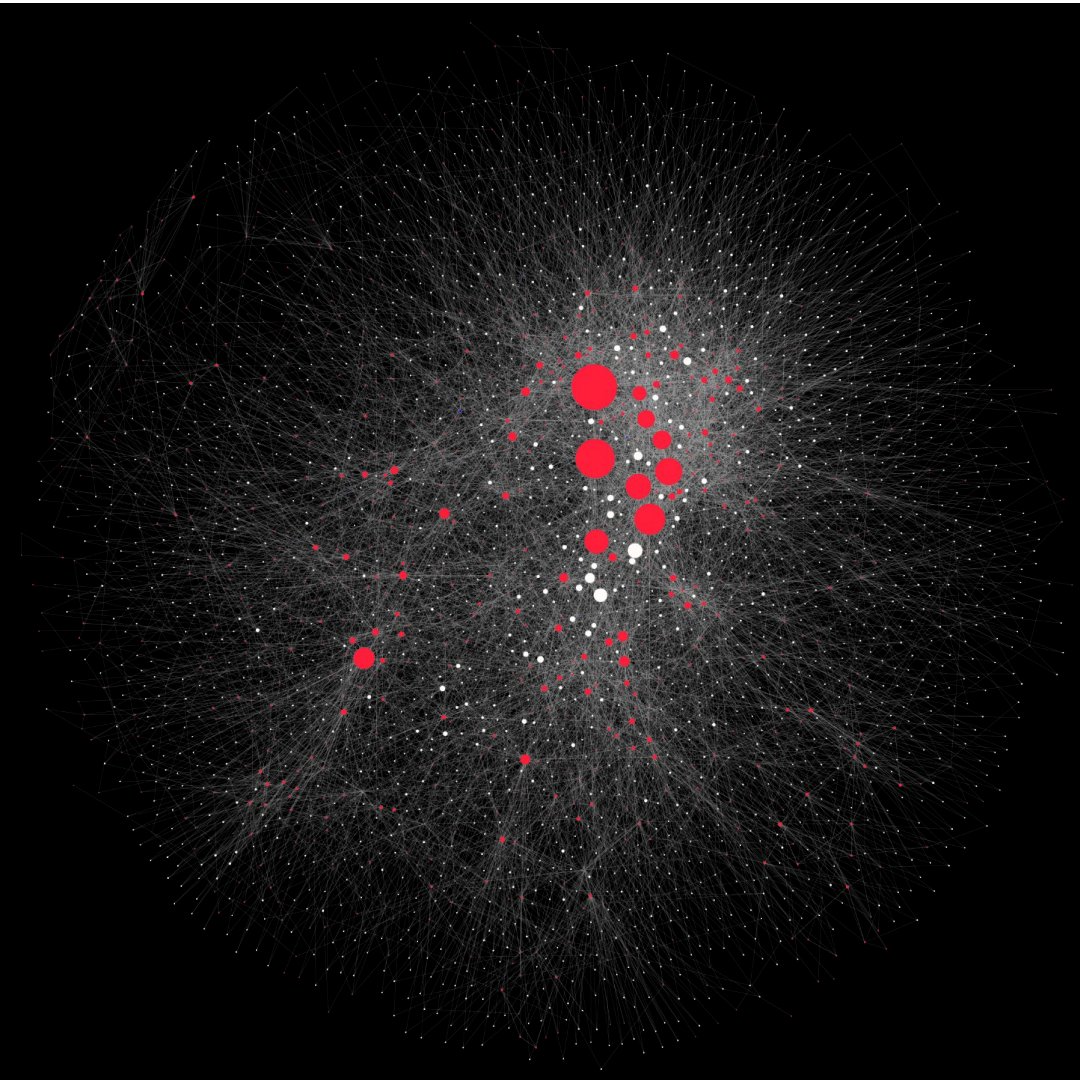
5. Policy Implications & Next Steps

Policy Implications

- Public awareness plays an important role in shaping policy
- Wikipedia can help assess the focus of public attention and steer communication efforts in policy implementation
- Wikipedia analysis can serve as a starting ground to identify common denominators between topics
- SNA wikipedia analysis can help identify central and less central group via centrality cluster (Prioritization strategy in communication, politics or research)

Future Ideas & Next Steps

- Compare different languages / coverage in particular countries:
  - Hate Groups in the US (EN wikipedia) v. Hate Groups in Germany (DE wikipedia)
- Compare different Hate Group types:
  - Anti-Immigrant v. Racist Skinhead v. Anti-LGBTQ
- Include click metrics, possibly via Google searches for keywords
  - Some studies have compared the viability of Google Trends vs. Wikipedia metrics in other areas (eg. finance and economics)
- Examine duration of public attention spikes around news coverage
  - How long does public attention (and impetus to separately research a topic) last after initial news coverage, and what factors might influence that duration?



Expanding the Network to the Third Level

Explore the networks yourself:

- 2nd level: https://bit.ly/3g4HJjP
- 3rd level: https://bit.ly/3uNzVXH

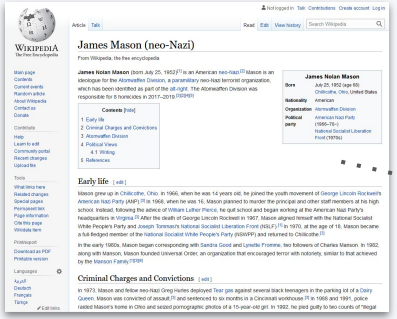
Wikipedia - Demography of Editors - Implication

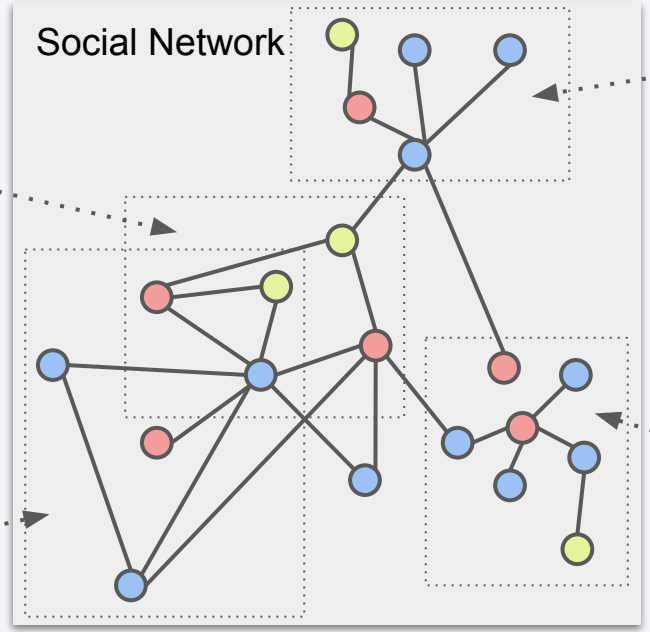
About Public knowledge of far right terrorism, some assumptions:

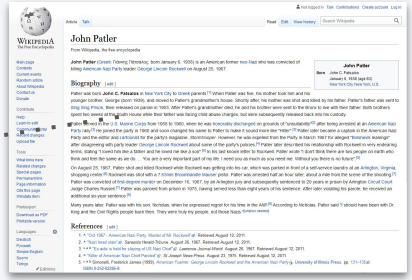
-Wiki tracks public knowledge/understanding of far right terrorism only from those who care and are active in discussing on far right terrorism, and majority of them are non-religious, young people

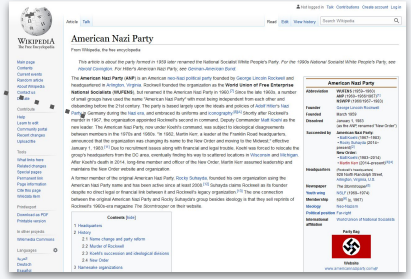
-Which may lead to bias on our analysis of public knowledge of far right terrorism, because wikipedia does not contain much information created/edited by people who don’t care about it or who does not think ring wing groups are associated with terrorists

Plotting a Network from Multiple Articles

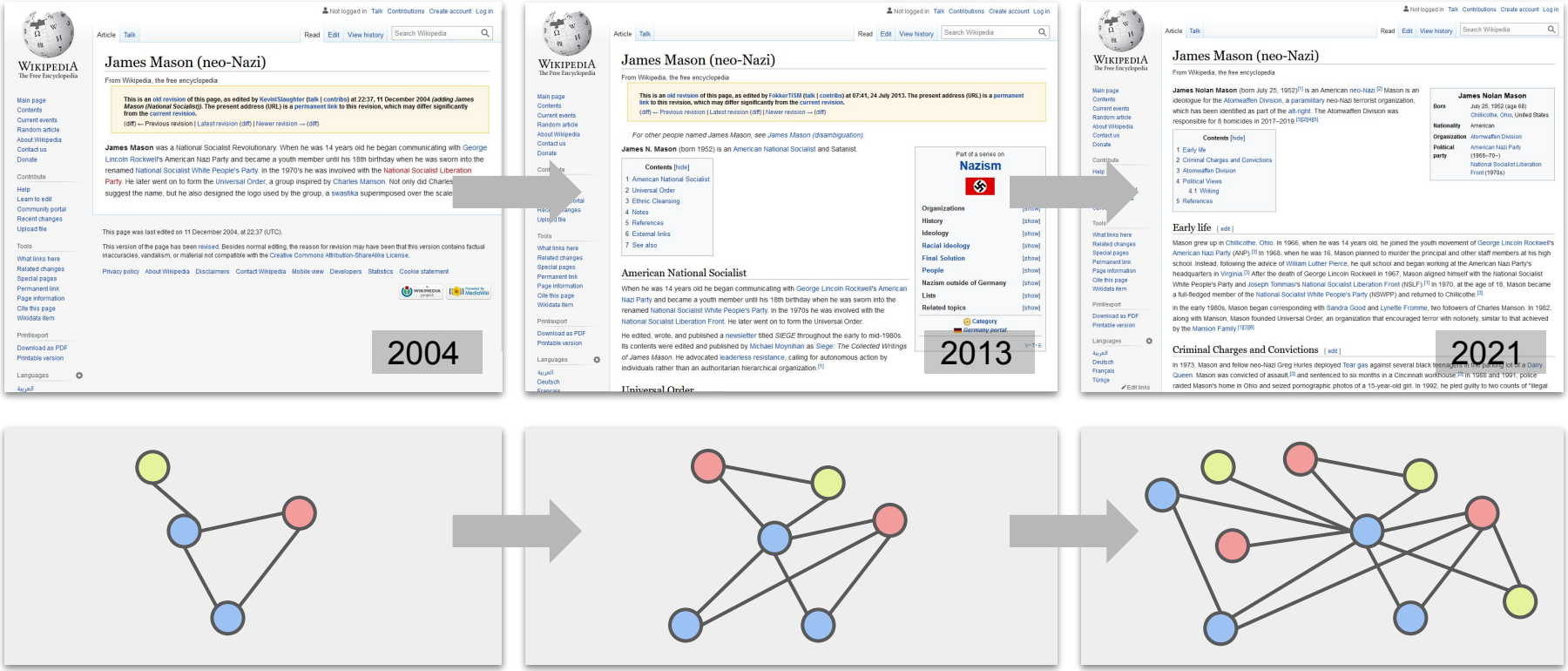




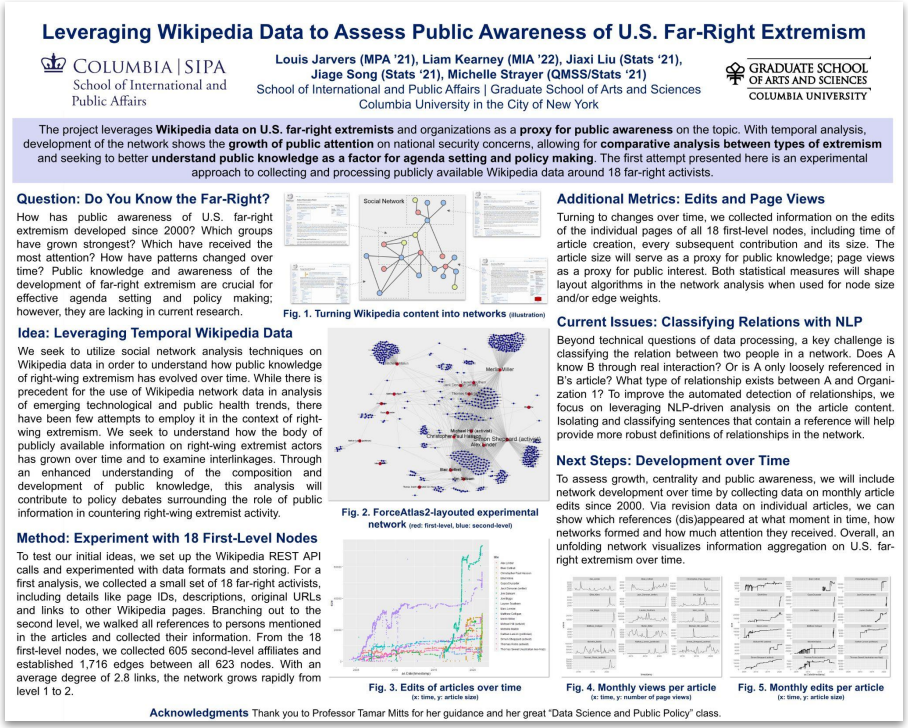




Development of public knowledge over time



Poster Submission



Leveraging Wikipedia Data to Assess Public Awareness of U.S. Far-Right Extremism

COLUMBIA | SIPA School of International and Public Affairs

Louis Jarvers (MPA '21), Liam Kearney (MIA '22), Jiaxi Liu (Stats '21), Jiage Song (Stats '21), Michelle Strayer (QMSS/Stats '21)
School of International and Public Affairs | Graduate School of Arts and Sciences
Columbia University in the City of New York

GRADUATE SCHOOL OF ARTS AND SCIENCES COLUMBIA UNIVERSITY

The project leverages **Wikipedia data on U.S. far-right extremists** and organizations as a **proxy for public awareness** on the topic. With temporal analysis, development of the network shows the **growth of public attention** on national security concerns, allowing for **comparative analysis between types of extremism** and seeking to better **understand public knowledge as a factor for agenda setting and policy making**. The first attempt presented here is an experimental approach to collecting and processing publicly available Wikipedia data around 18 far-right activists.

Question: Do You Know the Far-Right?

How has public awareness of U.S. far-right extremism developed since 2000? Which groups have grown strongest? Which have received the most attention? How have patterns changed over time? Public knowledge and awareness of the development of far-right extremism are crucial for effective agenda setting and policy making; however, they are lacking in current research.

Fig. 1. Turning Wikipedia content into networks (illustration)

Idea: Leveraging Temporal Wikipedia Data

We seek to utilize social network analysis techniques on Wikipedia data in order to understand how public knowledge of right-wing extremism has evolved over time. While there is precedent for the use of Wikipedia network data in analysis of emerging technological and public health trends, there have been few attempts to employ it in the context of right-wing extremism. We seek to understand how the body of publicly available information on right-wing extremist actors has grown over time and to examine interlinkages. Through an enhanced understanding of the composition and development of public knowledge, this analysis will contribute to policy debates surrounding the role of public information in countering right-wing extremist activity.

Fig. 2. ForceAtlas2-layouted experimental network (red: first-level, blue: second-level)

Method: Experiment with 18 First-Level Nodes

To test our initial ideas, we set up the Wikipedia REST API calls and experimented with data formats and storing. For a first analysis, we collected a small set of 18 far-right activists, including details like page IDs, descriptions, original URLs and links to other Wikipedia pages. Branching out to the second level, we walked all references to persons mentioned in the articles and collected their information. From the 18 first-level nodes, we collected 605 second-level affiliates and established 1,716 edges between all 623 nodes. With an average degree of 2.8 links, the network grows rapidly from level 1 to 2.

Fig. 3. Edits of articles over time (x: time, y: article size)

Additional Metrics: Edits and Page Views

Turning to changes over time, we collected information on the edits of the individual pages of all 18 first-level nodes, including time of article creation, every subsequent contribution and its size. The article size will serve as a proxy for public knowledge; page views as a proxy for public interest. Both statistical measures will shape layout algorithms in the network analysis when used for node size and/or edge weights.

Current Issues: Classifying Relations with NLP

Beyond technical questions of data processing, a key challenge is classifying the relation between two people in a network. Does A know B through real interaction? Or is A only loosely referenced in B's article? What type of relationship exists between A and Organization 1? To improve the automated detection of relationships, we focus on leveraging NLP-driven analysis on the article content. Isolating and classifying sentences that contain a reference will help provide more robust definitions of relationships in the network.

Next Steps: Development over Time

To assess growth, centrality and public awareness, we will include network development over time by collecting data on monthly article edits since 2000. Via revision data on individual articles, we can show which references (dis)appeared at what moment in time, how networks formed and how much attention they received. Overall, an unfolding network visualizes information aggregation on U.S. far-right extremism over time.

Fig. 4. Monthly views per article (x: time, y: number of page views)

Fig. 5. Monthly edits per article (x: time, y: article size)

Acknowledgments Thank you to Professor Tamar Mitts for her guidance and her great "Data Science and Public Policy" class.

- First level nodes: 18 alt-right activists
- Second level nodes: 605 persons
- Edges: 1,716 edges
- Overview: Page view and edit data
- Submission of project idea to Columbia Data Science Day Spring 2021 on April 21

BACKUP

Tools & Analysis

Tools:

- Python wiki API scraping
- JSON processing for NLP entity extraction and formatting
- Gephi network analysis including temporal development (alternatively: Neo4j)

Analysis (beyond visualization):

- Statistical analysis: Times series analysis
- Network analysis:
  - Connections (indegree, outdegree)
  - Distribution (centrality, density, bridges)
  - Segmentiation
- What do we make out of this?

Feedback from Prof Mitts

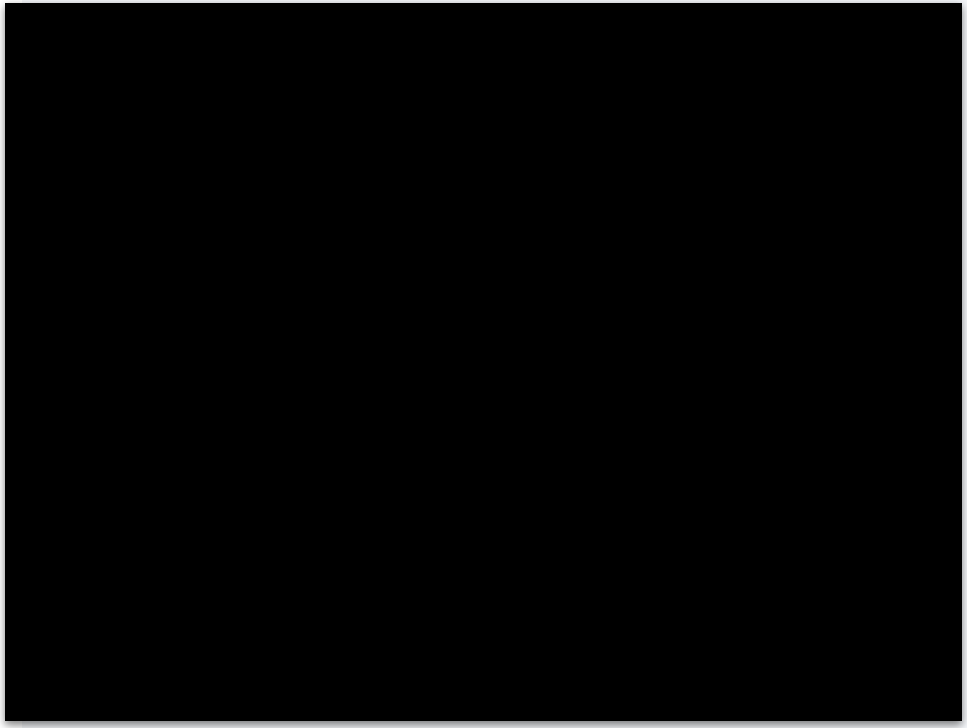
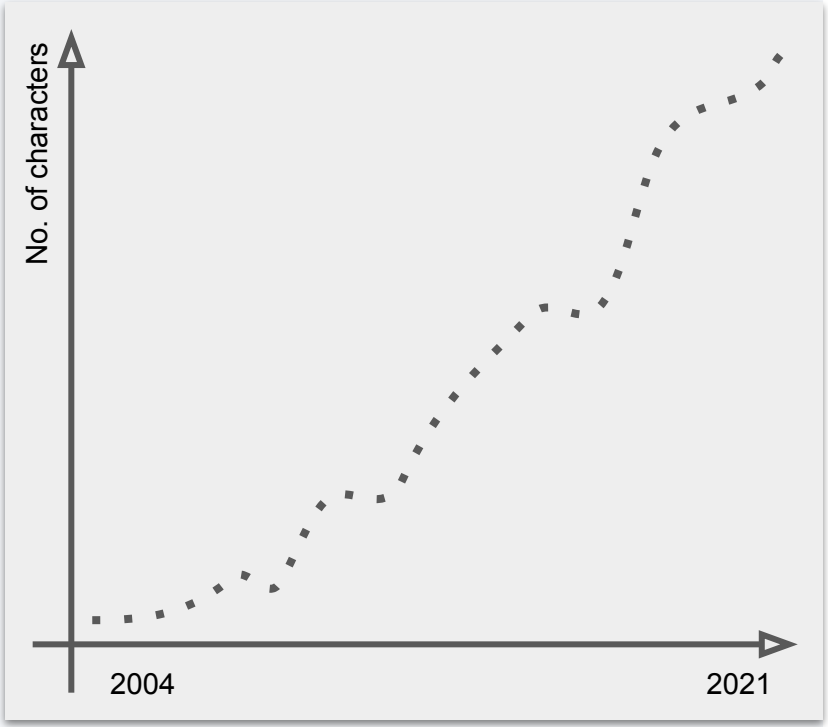
This sounds like a very interesting project. The main question to ask is what Wikipedia data (and linked pages) can allow you to measure. You'll basically be working with text data, which is created/posted by different authors in different time periods (could be used as metadata). The content you'll be analyzing will describe right wing extremism, but this will be from the perspective of public knowledge on these groups. So the main question to think about: are you interested in studying **public knowledge** about right-wing extremism or the **activity** of right-wing extremist groups? If the former, then analyzing with data could be a good approach. If the latter, then Wikipedia data would not get you very far.

In terms of methods, you should think about what you'd like to measure -- is it attention to particular narratives advanced by right-wing extremist groups in the United States (e.g., QAnon, the deep state, rigged elections, etc.)? If not, what other concepts/topics would you like to study?

While attention to far-right extremism is rising in public discourse, this is a relatively recent development. Your project could shed light on how knowledge on far-right extremism has grown over time. Happy to talk more on these ideas if you'd like.

Other data sources / ideas

Wikipedia traffic data?

https://wikiworkshop.org/2019/papers/Wiki_Workshop_2019_paper_9.pdf

Networks expansion over time