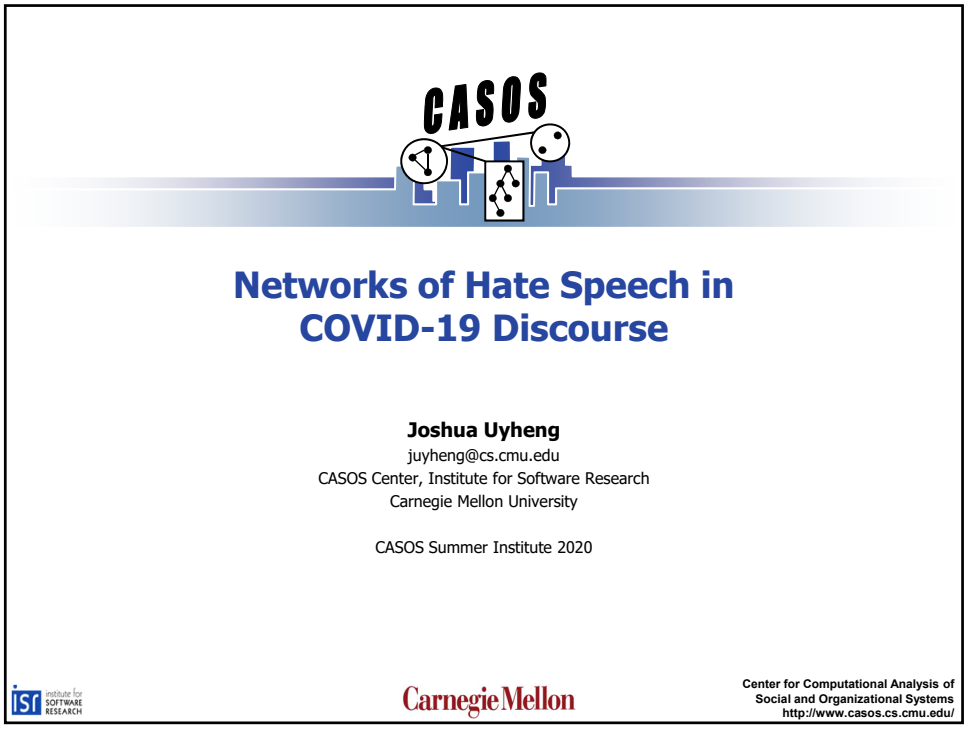
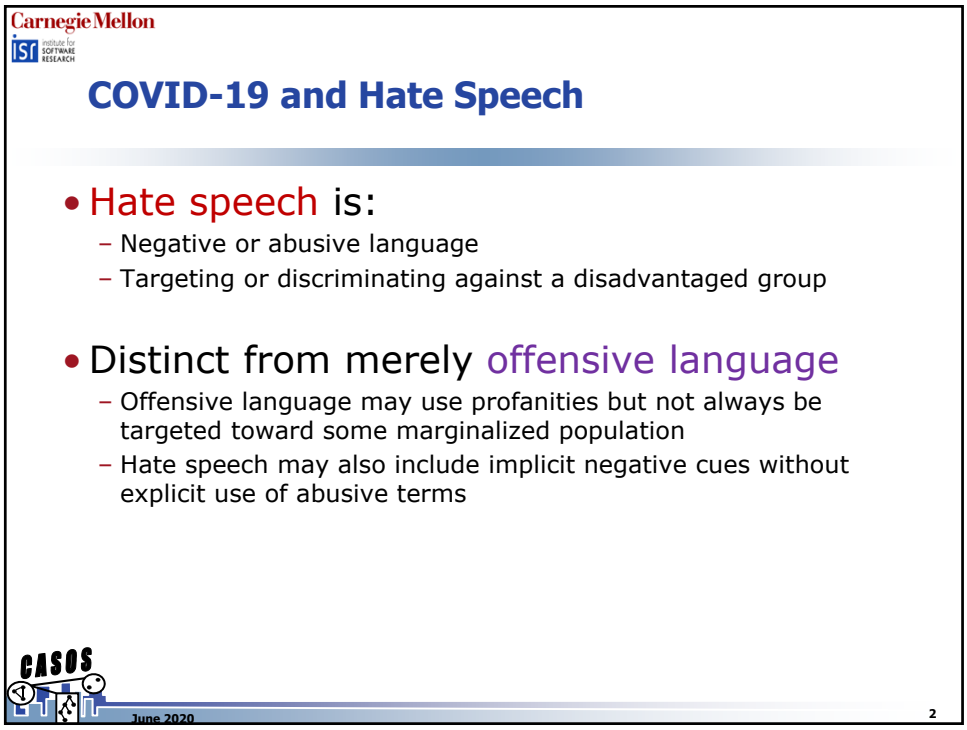
<Your Name>


CASOS
Networks of Hate Speech in COVID-19 Discourse
Joshua Uyheng
juyheng@cs.cmu.edu
CASOS Center, Institute for Software Research
Carnegie Mellon University
CASOS Summer Institute 2020
Carnegie Mellon
Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/


Carnegie Mellon
COVID-19 and Hate Speech
• Hate speech is:
– Negative or abusive language
– Targeting or discriminating against a disadvantaged group
• Distinct from merely offensive language
– Offensive language may use profanities but not always be targeted toward some marginalized population
– Hate speech may also include implicit negative cues without explicit use of abusive terms
CASOS
June 2020
2

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Definition/s of hate speech

- Hate speech is:
  - Negative or abusive language
  - Targeting or discriminating against a disadvantaged group
- Distinct from merely offensive language
  - Offensive language may use profanities but not always be targeted toward some marginalized population
  - Hate speech may also include implicit negative cues without explicit use of abusive terms

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Hate speech as a *social phenomenon*

- Language does not exist in a vacuum
  - It is perpetuated *by* groups
  - It is committed *against* groups
- Over time, it is important to see how hate speech shapes social interaction
  - Formation of communities
  - Accrual of individual influence

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Value of a dynamic network perspective

- Network science helps us:
  - Understand large-scale and complex patterns of relationships
  - See a social phenomenon at multiple scales
- Dynamic network methods are:
  - Interoperable with machine learning and other cutting-edge computational tools
  - Enable intuitive visualizations

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Objectives of this case study

- In the context of the COVID-19 pandemic:
  - How can we empirically examine hate speech in its socially networked setting?
  - How can we characterize individuals and groups which do and do not engage in hate speech?

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

A QUICK DETOUR

WHAT IS HATE SPEECH?

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Can we use a data-driven method to figure out what hate speech "is"?

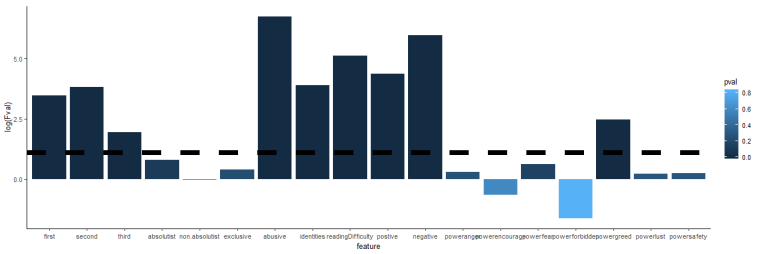
- 24K tweets labeled as hate speech, offensive language, or neither
  - 1430 hate speech (5.77%)
  - 191909 offensive language (77.43%)
  - 4163 neither (16.80%)
- Measured linguistic cues using Netmapper
  - Ran ANOVA tests to see statistically significant differences

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM*.

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Abusives are most significant; absolutist, exclusive, and power words non-significant.

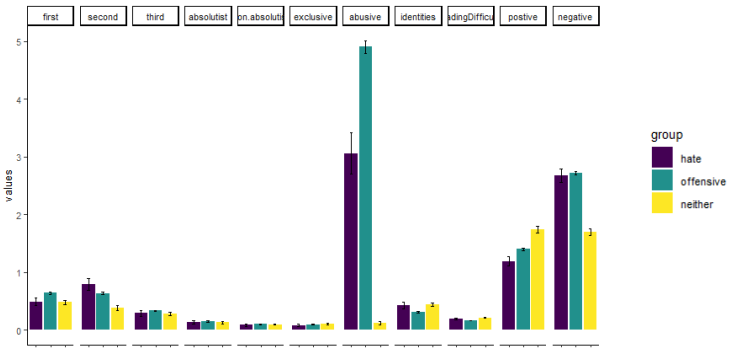


The above plot depicts F values of one-way ANOVA (log scale).
Bars are colored by p-value, with darker shades corresponding to lower p-values.
A dashed line represents the critical F value (log scale) at an alpha = .05.

CASOS

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Hate speech uses negative and abusive terms, second-person language, and identities.



The above plot depicts the mean values of different linguistic indicators across categories.
Error bars correspond to 95% confidence intervals.

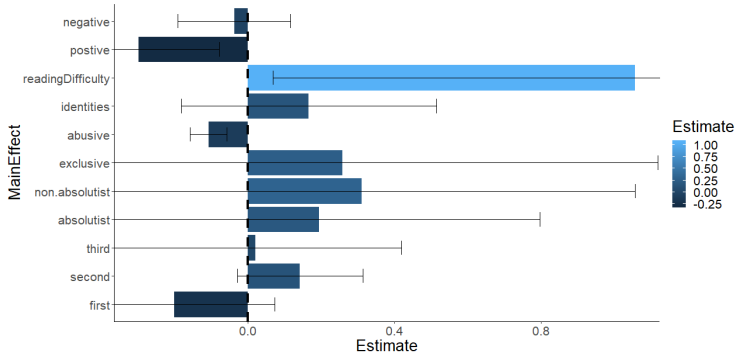
CASOS

CASOS

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Significant main effects detected only for: positive terms, abusive terms, and complexity.

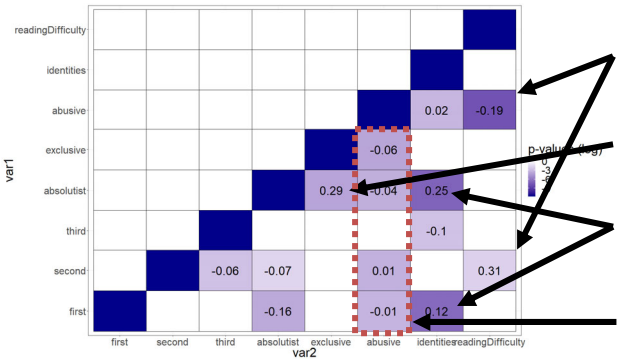


The above plot depicts coefficient values of 'main effects' (i.e., no interactions) in logistic regression classifying hate speech against regular and offensive language.
Error bars correspond to 95% confidence intervals.

CASOS

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

But many interaction effects distinguish hate from regular and offensive speech.



Hate speech is *complex* and uses *more second-person* language but *less abusive terms*.

Hate speech combines *absolutist* and *exclusive* language.

Hate speech combines *identities* with *absolutist* and *first-person* language.

Interestingly, for hate speech, abusive terms interact only a little with other features, likely because we are classifying against offensive language.

The above plot depicts the estimated interaction effects in logistic regression classifying hate speech from regular and offensive language.

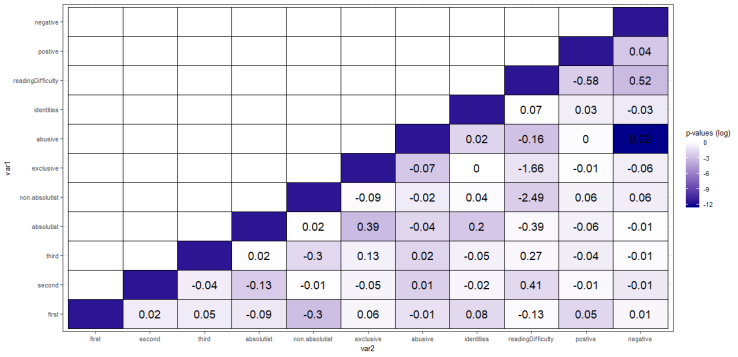
CASOS

CASOS

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

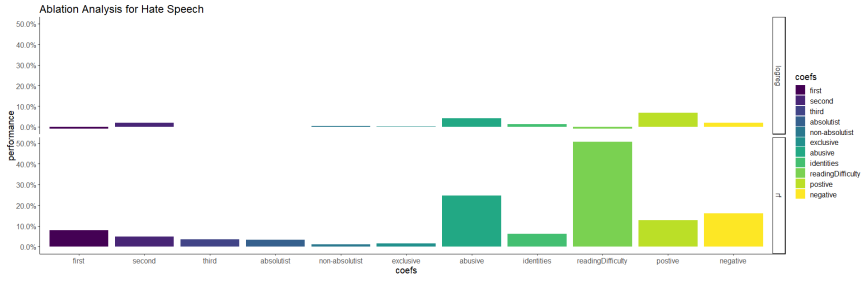
But many interaction effects distinguish hate from regular and offensive speech.



The above plot depicts the estimated interaction effects in logistic regression classifying hate speech from regular and offensive language.

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Ablation analysis further suggests most crucial identifiers of hate speech are complexity, abusives, and positive/negative terms.



To perform ablation analysis, we trained classifiers to perform hate speech classification while removing one predictor at a time.
Values presented are percent difference in F1 score compared to model trained on full data. Higher values suggest greater importance for the variable.
The two models used for these experiments were a logistic regression classifier and a 100-tree random forest.

<Your Name>

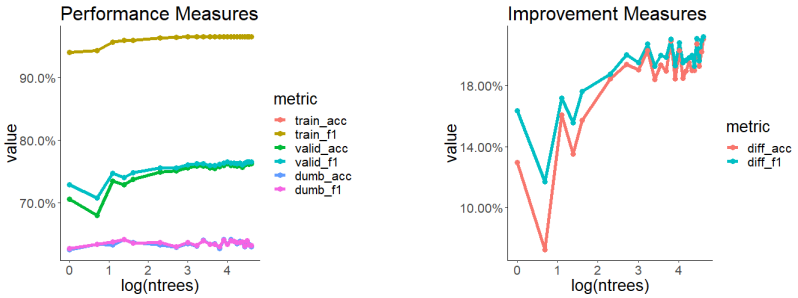
Carnegie Mellon ISR institute for SOFTWARE RESEARCH

Machine Learning Classifier

- Training Procedure
  - Oversampling during training to have equal proportions across categories
  - 70-20-10 train-validate-test split
- Evaluation
  - Measure accuracy, F1 ('weighted') scores
  - Compare against random baseline
  - Choose classifier with best validation performance
  - Final evaluation on test set

Carnegie Mellon ISR institute for SOFTWARE RESEARCH

Random forest with 50 trees gives best validation performance with decent improvement over baseline.



Test accuracy is 76.40% ||| Test F1 score is 76.74%
Accuracy improvement is 22.51% ||| F1 improvement is 21.85%

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

RESULTS

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Data (Preliminary – to be expanded)

- Twitter data
  - Collected using REST API
  - Terms: #COVID19US
    - At some point official hashtag used for pandemic discourse specific to the United States
  - Dates: March 5 – 25 (21 days)
    > data available already up to May still processing

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Exploratory questions

- How much hate speech and offensive language do we detect in online discussion of the #COVID19US hashtag?
- How much bot activity do we detect in online discussion of the #COVID19US hashtag?
- Are the two quantities related?

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Method

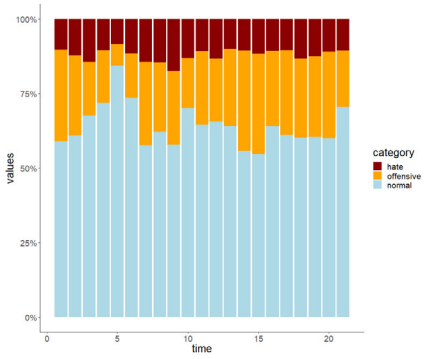
- Hate speech detection
  - Features: Linguistic cues associated with psychological states (see Pennebaker)
  - Model: Random forest with 40 estimators
    - Trained on open dataset of hate speech, offensive language, normal language
    - Achieved ~97% training accuracy and F1; ~75% testing accuracy and F1
- Network analysis with ORA
  - Visualization of agent x agent networks
  - Visualization of lexical networks for hate speech

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Relative levels of hate appear to fluctuate over time.

- #COVID19US discourse is dominated by language that is neither offensive nor hate speech
- However, noticeable proportions of the latter persist
  - Between 8-17% hate speech
  - Between 7-30% offensive



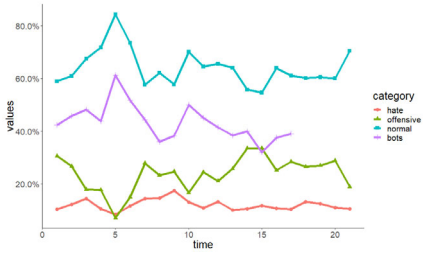
CASOS

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Are bots driving hate speech and offensive language? Results suggest they do not.

- Bot activity over time is negatively correlated to both offensive language and hate speech
- Bot activity instead positively correlated with normal speech

	hate	offensive	normal	bots
hate	1.0000000	0.1269184	-0.4097136	-0.3371000
offensive	0.1269184	1.0000000	-0.9568375	-0.8094684
normal	-0.4097136	-0.9568375	1.0000000	0.8431966
bots	-0.3371000	-0.8094684	0.8431966	1.0000000



CASOS

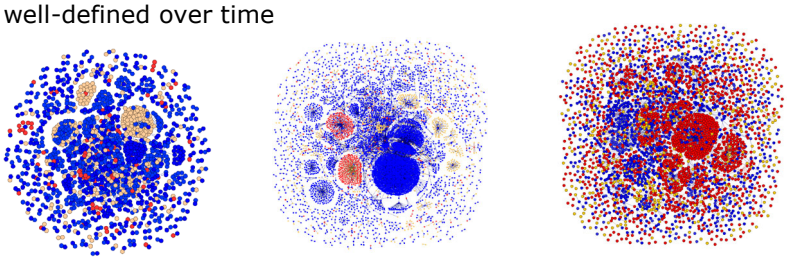
CASOS

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

What is striking, however, is the apparent formation of *hate communities*.

- Networks of users deploying hate speech appear to grow more well-defined over time



March 5 | March 14 | March 25

Figures depict agent x agent networks (replies + retweets + mentions).
Agents colored based on use of hate speech (red), offensive language (orange), and neither (blue).

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

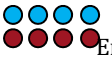
Quantifying community formation: Hate entropy as a measure of randomness

- Entropy measures level of disorder or randomness in a system
- Computation
  - Suppose there are N possible labels for a system of nodes
  - Then for label k in {1, 2, … N}, we define:

$$p_k = \frac{\#nodes\ with\ label\ k}{\#total\ nodes}$$

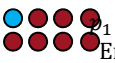
  - Entropy = $-\sum_{k=1}^{N} p_k \log p_k$

- Higher-entropy system: Less homophily



$p_1 = 0.5, p_2 = 0.5$
Entropy = 0.6931472

- Lower-entropy system: More homophily



$p_1 = 0.875, p_2 = 0.125$
Entropy = 0.3767702

- As hate speech grows more clustered, we expect *hate entropy to go down*

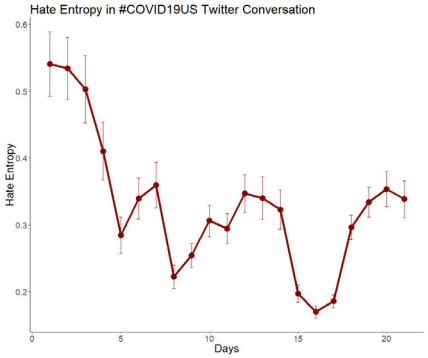
CASOS

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Hate entropy metric shows that distribution of hate speech is *less random*, more clustered.

- Procedure for calculation:
  - Produce Louvain clusters over Agent x Agent network (All Communication)
  - Take only subset of Louvain clusters with size $\geq$ 10
  - Compute entropy of hate class labels per cluster
  - Take mean over time



Interestingly, still not correlated to bot activity – is the hate speech organic?

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

DISCUSSION

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Some Takeaways

- Hate speech is an important yet challenging problem to examine in the context of a global pandemic
- It is important to see hate speech as both a linguistic and socially networked phenomenon
- Interoperable pipelines of network science and machine learning tools can help us approach the problem empirically
- Policies designed to respond to hate speech and other social cyber-security issues must be grounded in multidisciplinary and multi-methodological perspective

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

METHODOLOGY

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Tools

1. Netmapper

- To measure use of abusive terms
- To measure use of identity terms

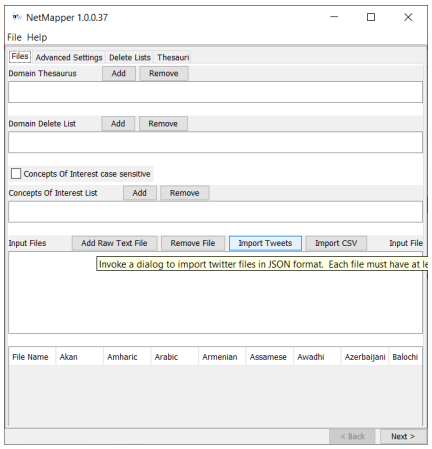
2. ORA

- To visualize social interactions
- To measure important network metrics

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for Netmapper: Loading data

- Load files into Netmapper using the Import Tweets button
- We want the following files:
  - covidhate_20200309.json
  - covidhate_20200314.json
  - covidhate_20200319.json



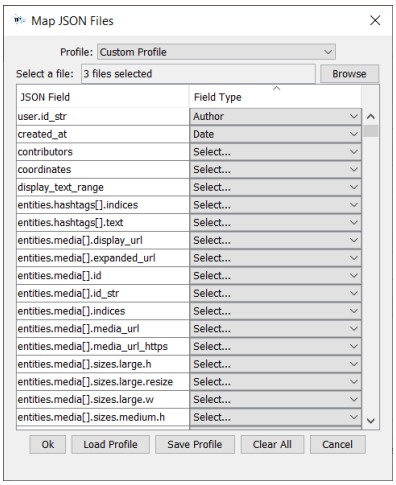
CASOS

<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for Netmapper: Analysis

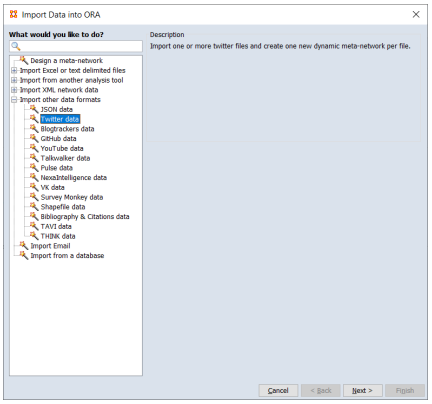
- Make sure relevant Netmapper fields match their corresponding JSON fields
  - Author: user.id_str
  - Date: created_at
  - Tweet ID: id_str
  - Text: full_text
- Run and save Netmapper files
  - Make sure we are getting "usage measures"



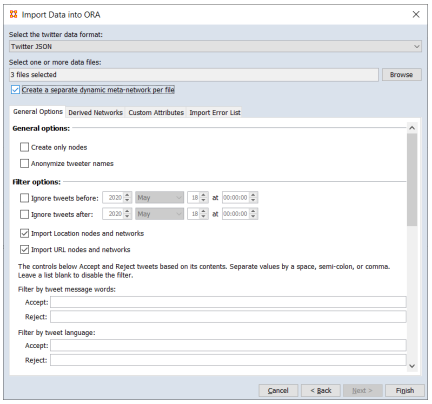
Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Loading data

Import Twitter data



Create a separate dynamic meta-network per file

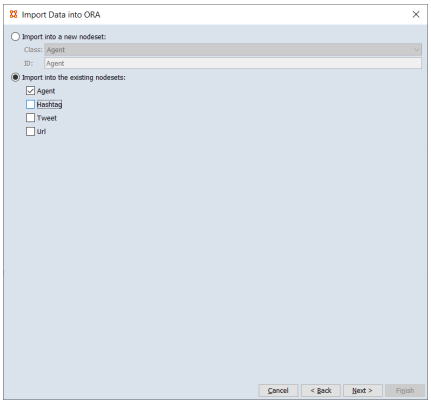


<Your Name>

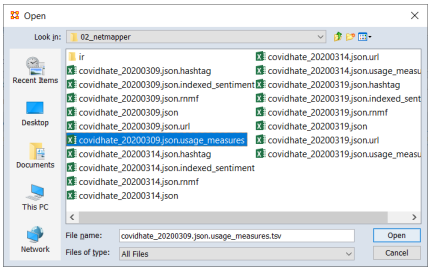
Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Loading attributes

Load attributes only for Agents



Use the appropriate "usage_measures" files



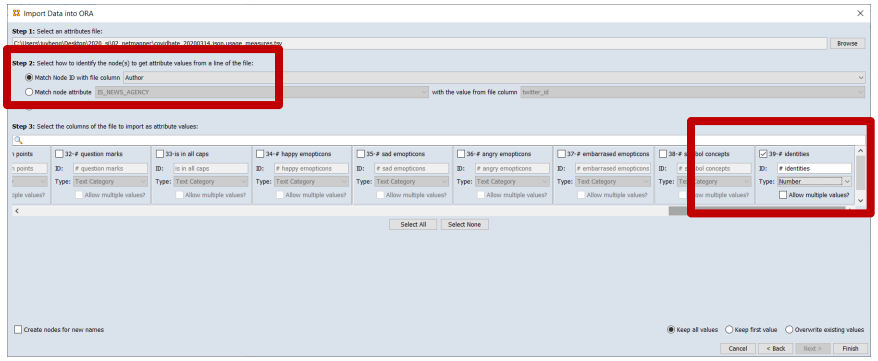
CASOS

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Loading attributes

Match NODE ID with file column Author

Make sure to click only abusives and #identities



CASOS

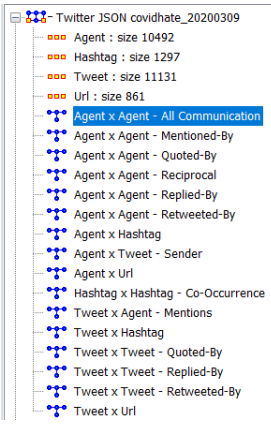
CASOS

<Your Name>

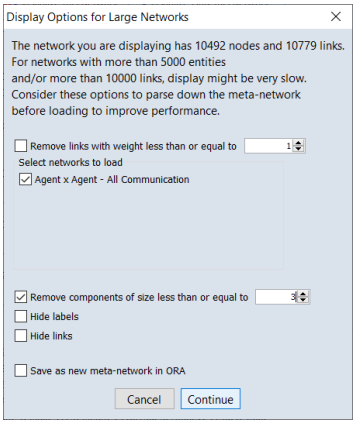
Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Visualize!

Visualize All Communication



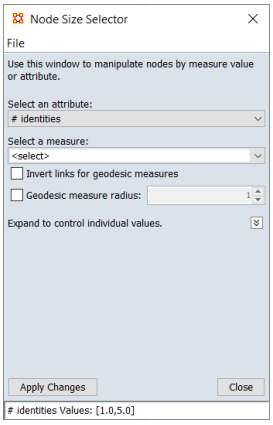
Remove components smaller than 3 nodes



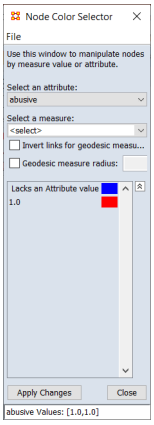
Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Visualize!

Size by identities invoked



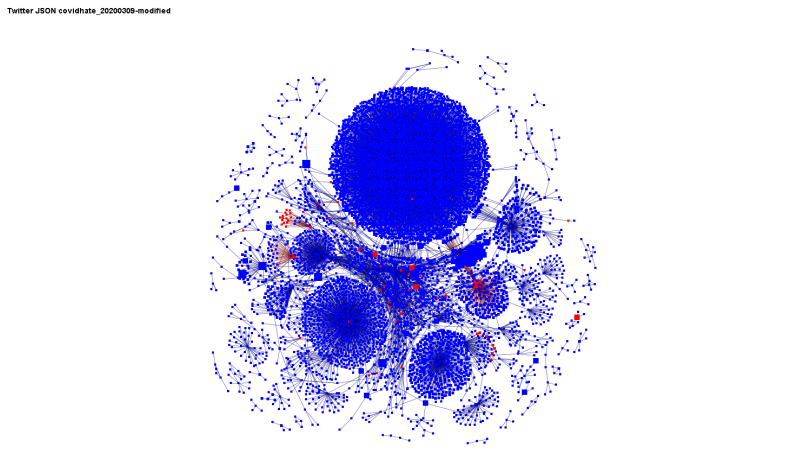
Color by use of abusive terms



<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

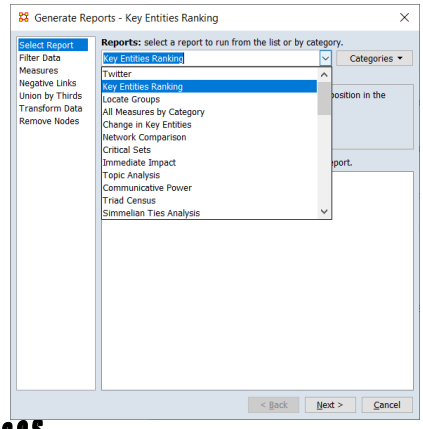
Sample ORA network visualization



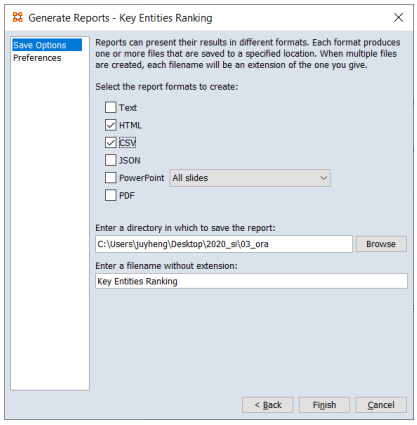
Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Run Reports

Select Key Entities Ranking



Choose Default Settings and Save HTML and CSV Output



<Your Name>

Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

Instructions for ORA: Run Reports

Who Attribute Analysis is helpful for high-level view

KEY ENTITIES RANKING REPORT

Input data: Twitter JSON covidhate_20200309

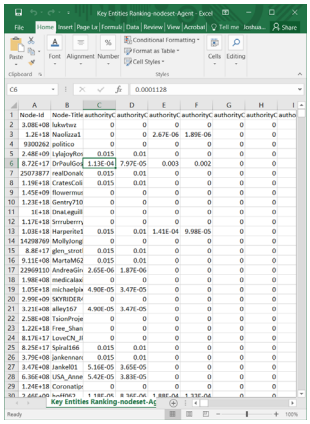
Start time: Mon May 18 12:40:57 2020

Data Description

Table of Contents

Produced by ORA, a joint product of the CASOS center at Carnegie Mellon University and Netanomics

CSVs provide raw metrics for downstream analysis

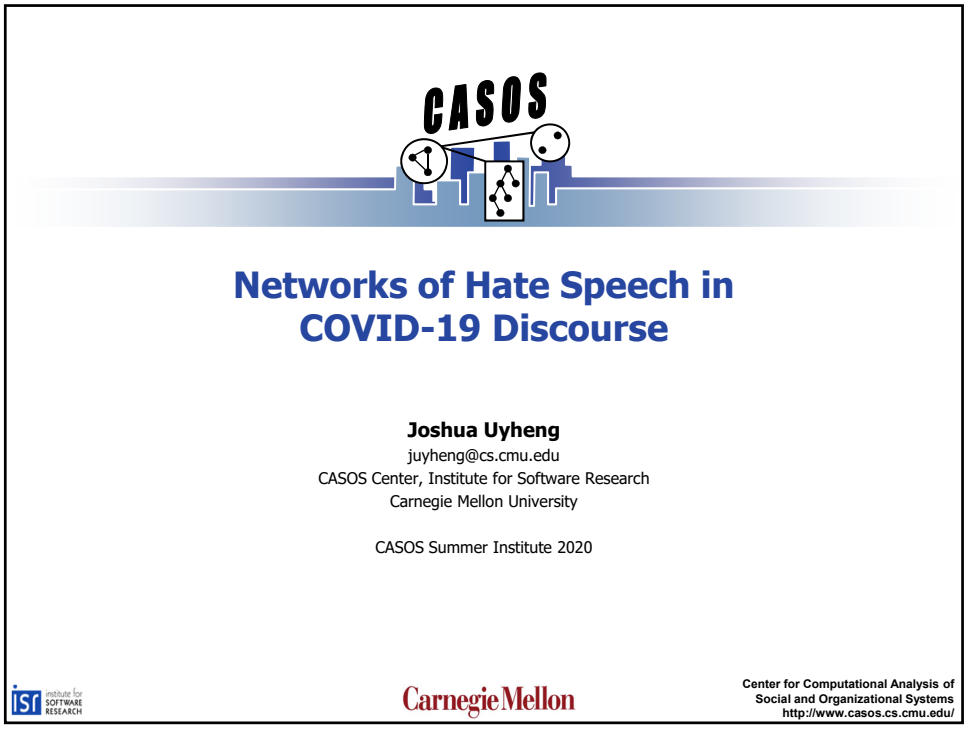


Carnegie Mellon
ISR institute for SOFTWARE RESEARCH

DEMONSTRATION

CASOS

<Your Name>


CASOS
Networks of Hate Speech in COVID-19 Discourse
Joshua Uyheng
juyheng@cs.cmu.edu
CASOS Center, Institute for Software Research
Carnegie Mellon University
CASOS Summer Institute 2020
ISR institute for SOFTWARE RESEARCH
Carnegie Mellon
Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/


CASOS