August 12, 2025

Exploring Hate Speech Dynamics: The Emotional, Linguistic, and Thematic Impact on Social Media Users

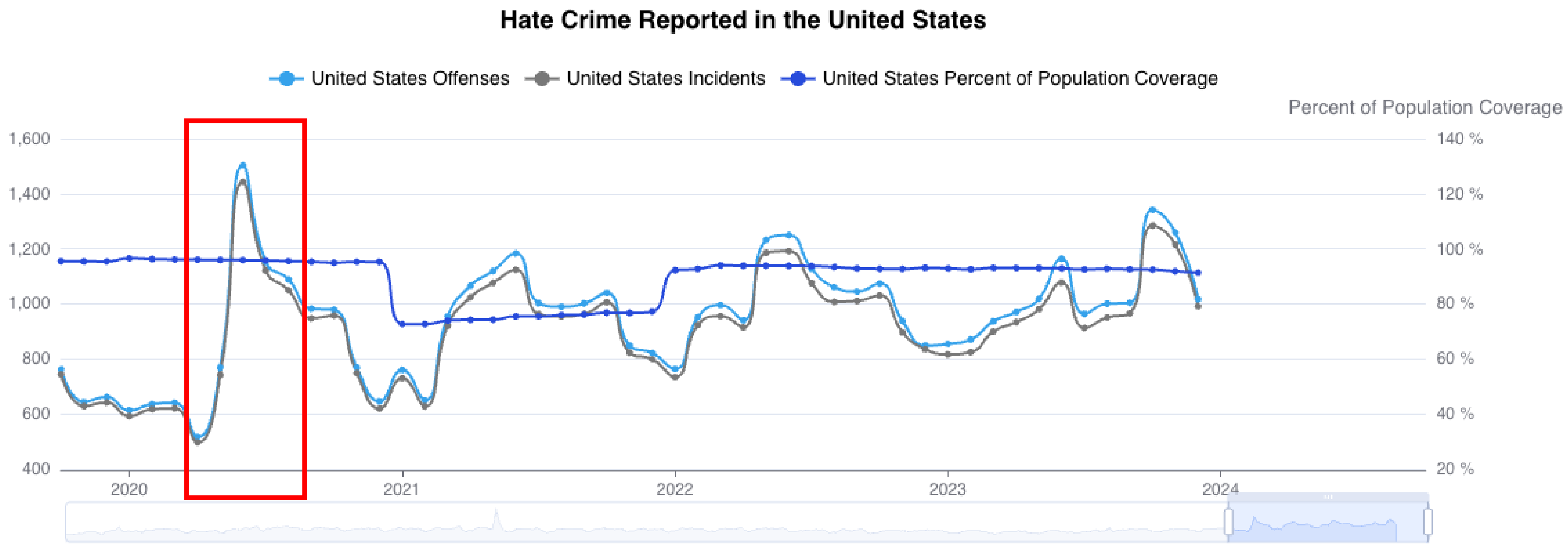
Amira Ghenai, Zeinab Noorian, Hadiseh Moradisani, Pariya Abadeh, Caroline Erentzen, Fattane Zarrinkalam

Information Processing & Management 62.3 (2025). https://doi.org/10.1016/j.ipm.2025.104079

Toronto Metropolitan University

TED ROGERS SCHOOL OF MANAGEMENT

AACSB ACCREDITED



source: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime

Asian racism a year after Atlanta spa shootings

Michelle Chen

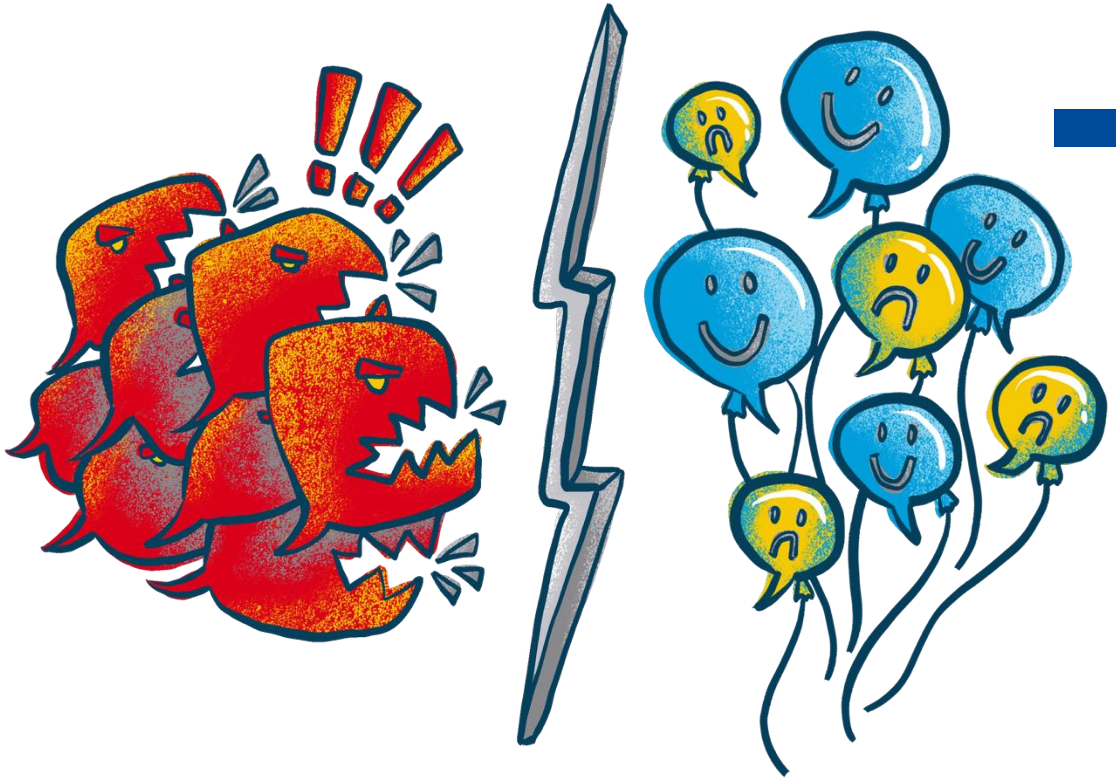
Wed 16 Mar 2022 12.21 GMT

Share

Composite: AFP, Shutterstock, Getty Images, Reuters

Context

- Social media platforms (e.g., Twitter) spread both **positive and harmful content**, including **hate speech**.
- **Hate speech**, especially during crises like COVID-19, surged, particularly targeting **East Asians**.
- Platforms **amplify hate speech** through **echo chambers**, increasing **societal harm** and risk of **offline violence**.



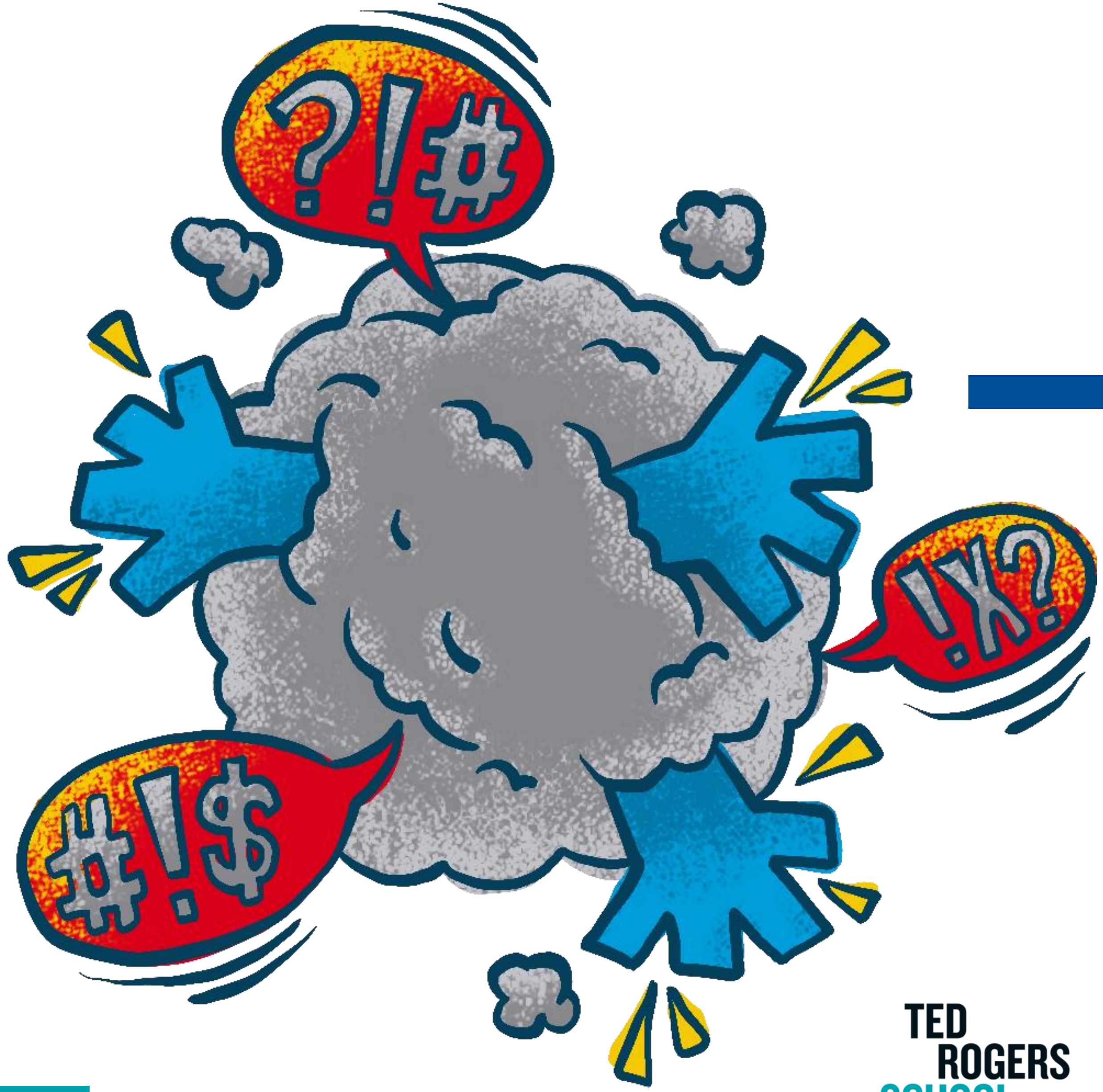
Context

- **Research Gaps**: Existing models focus on keyword detection without examining the network structure, or its progression over time

- **Study Objective**:

    - Investigate the linguistic/thematic patterns among hate speech users

    - Provides insights for proactive hate speech mitigation



TED ROGERS SCHOOL OF MANAGEMENT

Context

RQ1: ***What** is the effect of hate speech on the linguistic and cognitive characteristics of social media users who post hateful content compared to those who do not?*

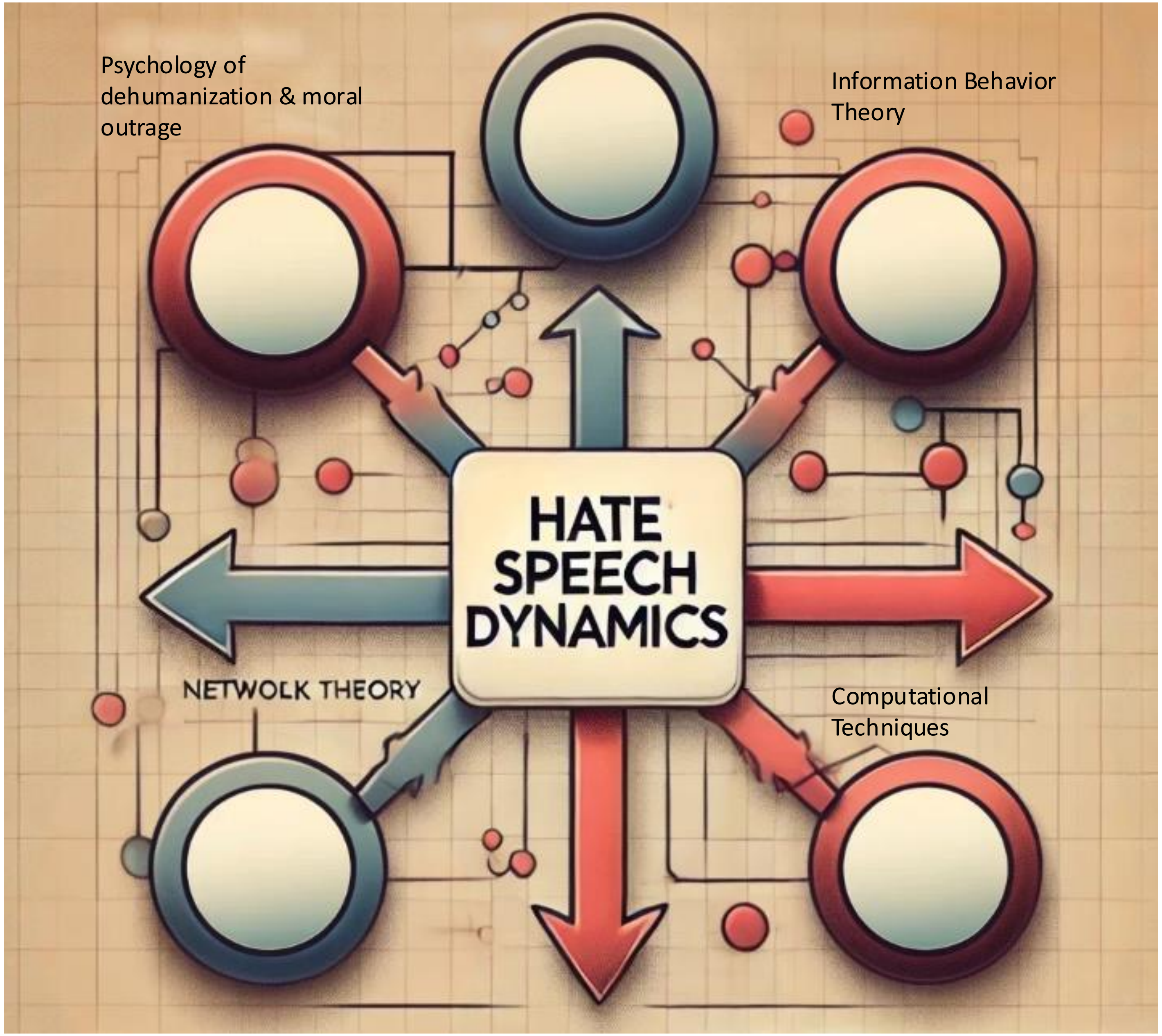
RQ2: **To what extent** *do the thematic patterns and specificity of hate speech narratives on social media differ from those of non-hate speech content?*

TED ROGERS SCHOOL OF MANAGEMENT

Outline

- Theoretical Foundation
- Methodology
  - Data Collection
  - Propensity Score Analysis & Network Analysis
- Results
- Conclusion & Future Work

Theoretical Foundation



Theoretical Foundation

RQ1: Linguistic and Cognitive Markers in Hate Speech

H1a
- Hate speech users show **higher** levels of negative emotions (anger, anxiety, sadness)

[Alorainy et al. (2018), ElSherief et al. (2018), Giner-Sorolla & Russell (2019), Haybron (2002), Mathew et al. (2018), Matsumoto et al. (2016), Sell et al. (2009)]

H1b
- Hate speech users use language related to **power, risk, and death**

[Elsherief et al. (2018), Goff et al. (2008), Markowitz & Slovic (2020), Paasch-Colberg et al. (2021)]

H1c
- Hate speech users employ more **third-person pronouns**, indicating detachment

[Elsherief et al. (2018), Faulkner & Bliuc (2018), Zannettou et al. (2020), Perdue et al. (1990), Shih et al. (2013), Matos & Miller (2023)]

H1d
- Hate speech involves more **profanity**

[Carter (1944), Leader et al. (2009), Bartlett et al. (2014), Bilewicz & Soral (2020), Jeshion (2013), Thurlow (2001), Anderson & Lepore (2013), Vallée (2014)]

H1e
- Hate speech is linked with **moral outrage** language

[Brady et al. (2021), Crockett (2017), Salerno & Peter-Hagene (2013), Grubbs et al. (2019), Young & Young (2020), Faulkner & Bliuc (2018), Solovev & Pröllochs (2023)]

TED ROGERS SCHOOL OF MANAGEMENT

Theoretical Foundation

RQ2: Thematic Coherence and Complexity in Hate Speech Narratives

H2a
- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]

H2b
- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

H2c
- Hate speech narratives display **lower** topic specificity

[Suedfeld & Tetlock (1977), Jakob et al. (2023), Faulkner & Bliuc (2018), Gregory & Piff (2021), Dhont & Hodson (2014), Hodson & Busseri (2012)]

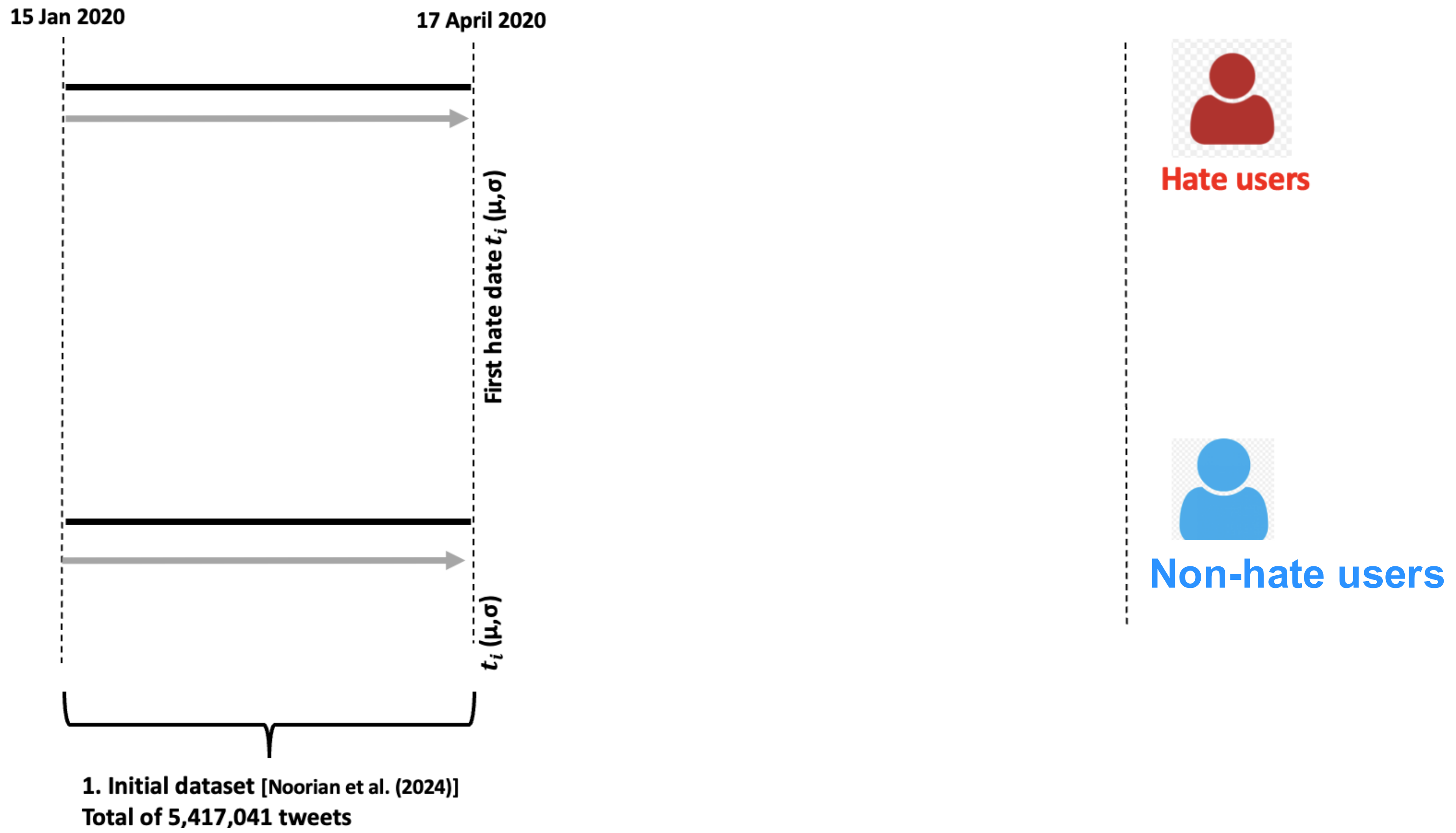
Methodology – Data Collection

Hate users: Twitter users posting at least 3 hateful tweets about anti-Asians during COVID-19

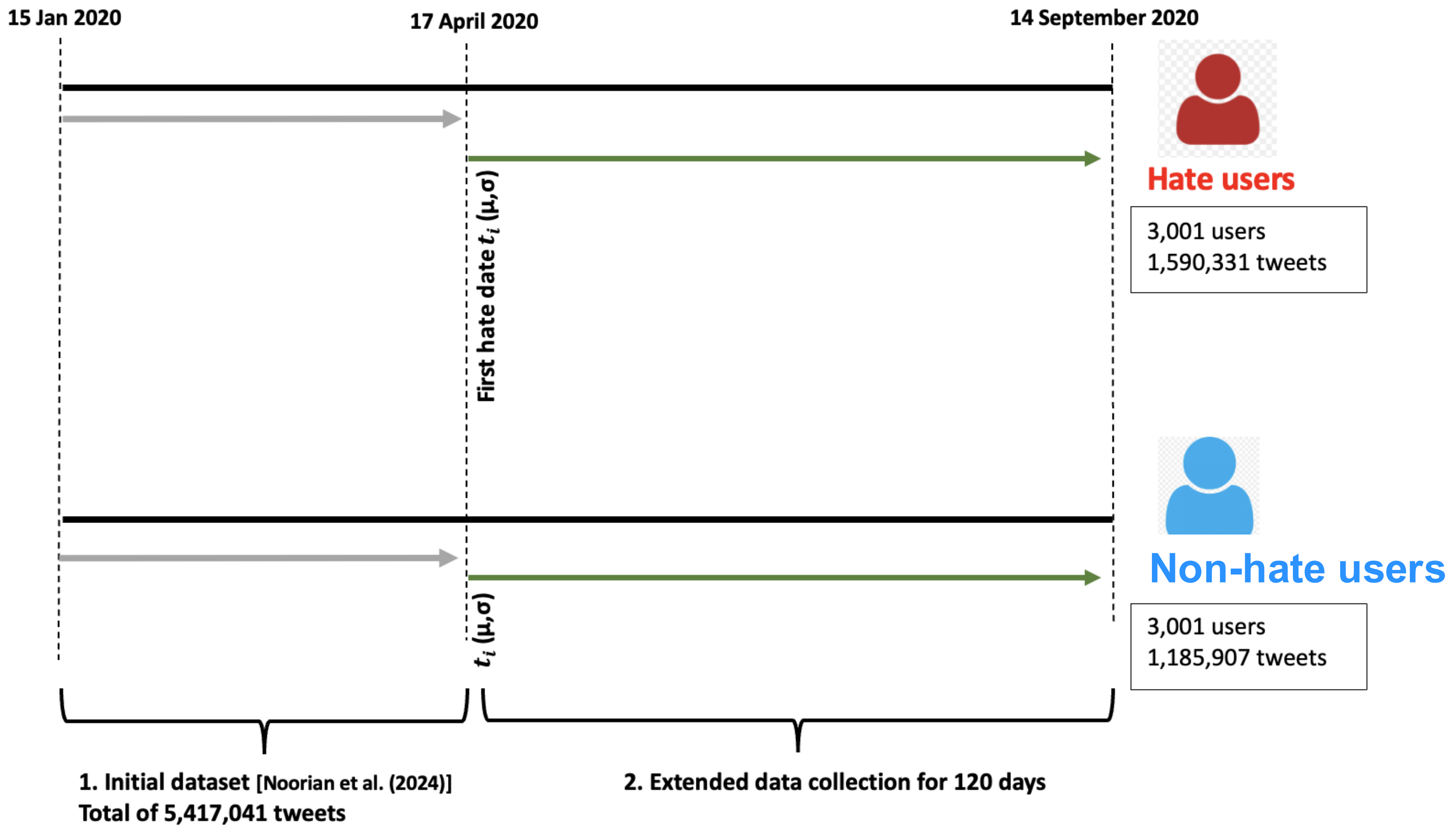
Non-hate users: Twitter users posting at least 3 tweets containing counter-hate/neutral content about anti-Asians during COVID-19

TED ROGERS SCHOOL OF MANAGEMENT

Methodology – Data Collection



Methodology – Data Collection



Methodology

RQ1: ***What** is the effect of hate speech on the linguistic and cognitive characteristics of social media users who post hateful content compared to those who do not?*

What? (outcome)
emotions, linguistic, cognitive factors → LIWC categories & ML classifier

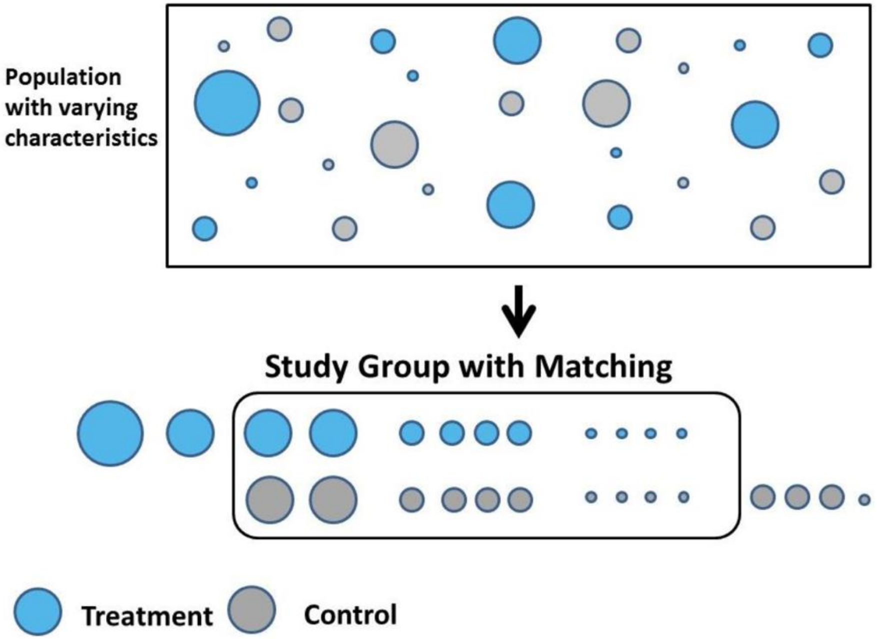
How? (methodology)
Propensity score Analysis

Stat. Significance?
t-test (Cohen's d Cohen)

TED ROGERS SCHOOL OF MANAGEMENT

Methodology – Propensity Score Analysis

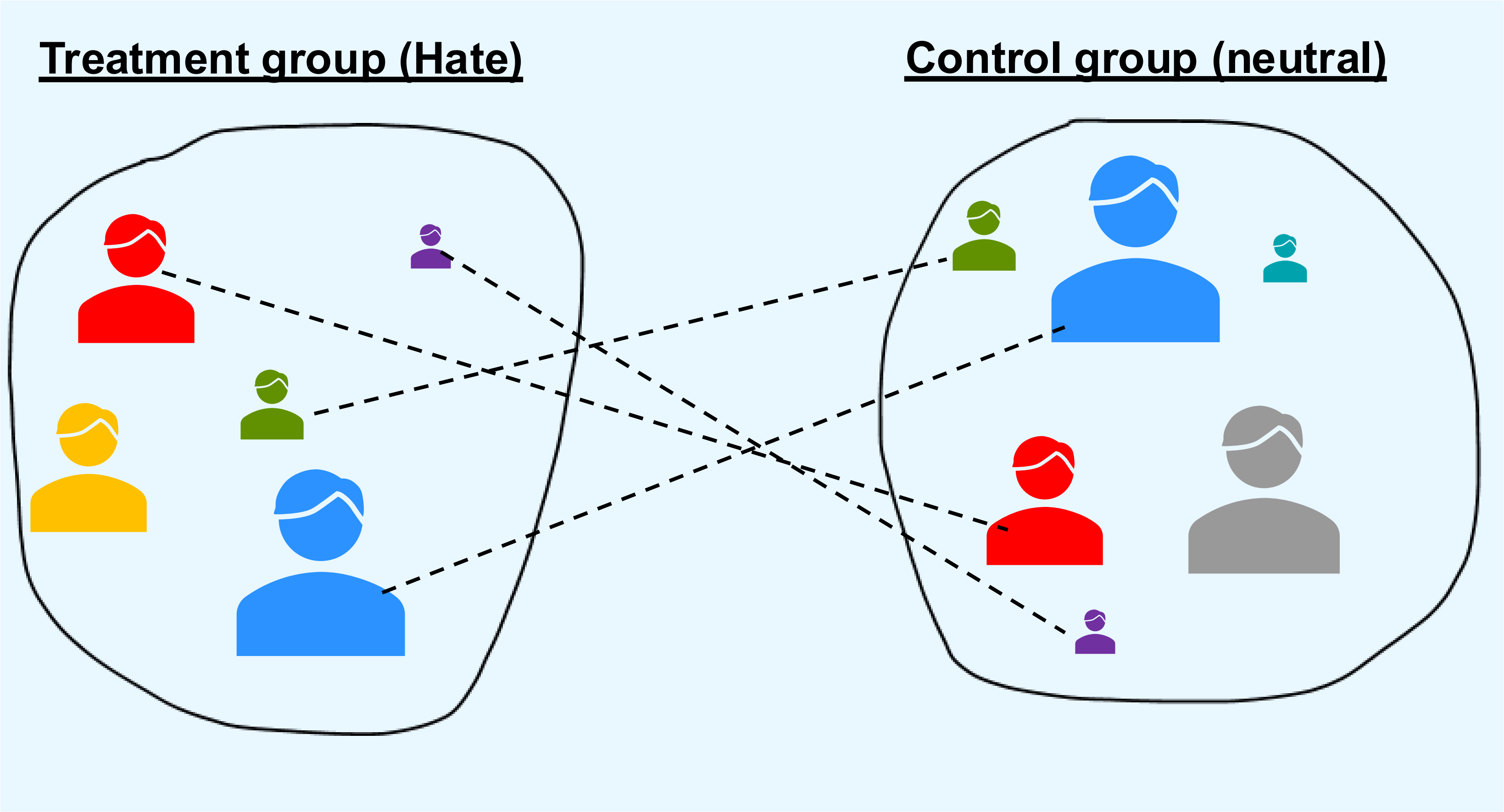
- **Concept**: Estimate what each user's behavior would look like with and without exposure to hate speech
- **Challenge**: Can't observe both outcomes for the same individual
- **Solution**: Match users with similar behaviors and characteristics



Methodology – Propensity Score Analysis

- **Approach**: Mimics a Randomized Controlled Trial (RCT) using propensity score matching
- **Treatment**: posting hate content in SM
- **Goal**: Compare “treatment” users (hate speech users) with “control” (non-hate users)
- **Outcome**: Measures differences in linguistic and thematic features between groups <u>before</u> posting hate

Methodology – Propensity Score Analysis



Methodology – Propensity Score Analysis

- **Propensity Score:** probability of a user being assigned to a specific group (i.e., posting hate speech).
- Calculated using logistic regression, to predict if an observation belongs to the treatment or control group
- Predictions are based on key covariates:
  - Linguistic (LIWC Features), User Activity, Network Features
- Stratified Matching: one-to-many (10 strata)

TED ROGERS SCHOOL OF MANAGEMENT

Methodology

RQ2: **To what extent** *do the thematic patterns and specificity of hate speech narratives on social media differ from those of non-hate speech content?*

What? (outcome)

interconnectedness,
global cohesion,
specificity

How? (methodology)
Topic Analysis &
Network Analysis
cosine similarity
Gini coefficient

Stat. Significant?
linear mixed-effects models

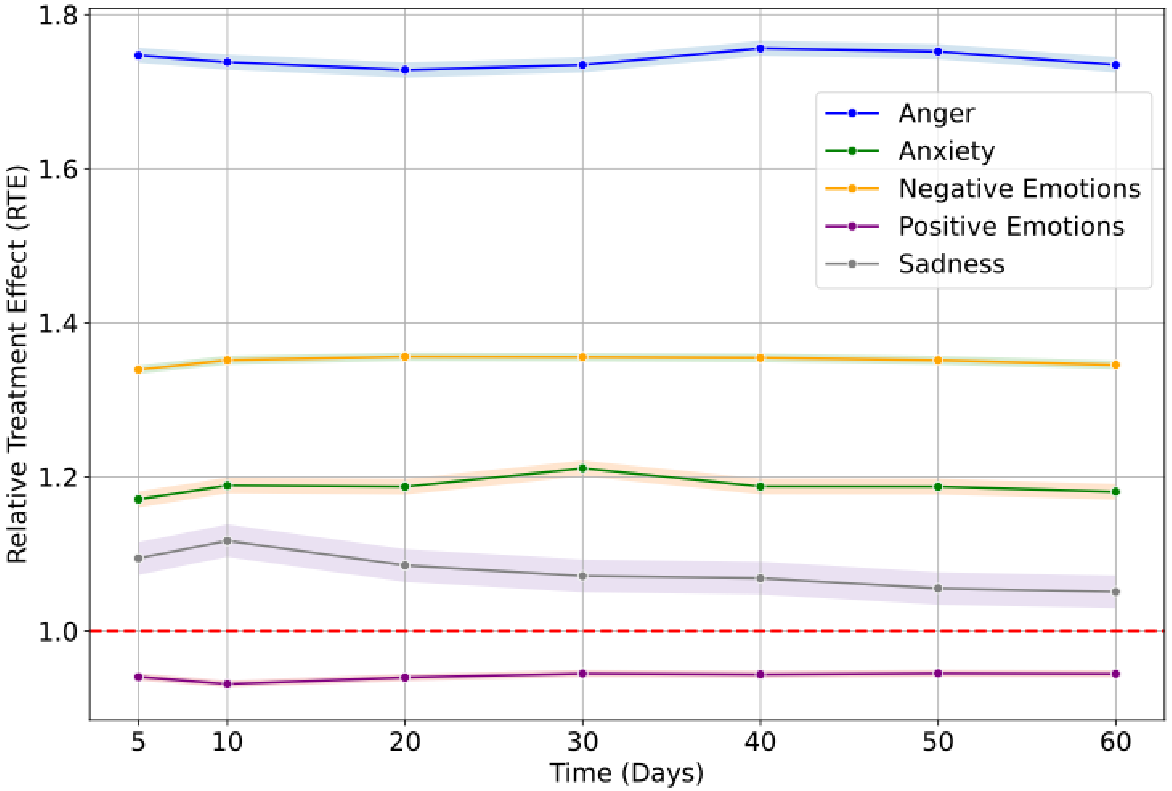
TED ROGERS SCHOOL OF MANAGEMENT

Results

H1a

- Hate speech users show **higher** levels of negative emotions (anger, anxiety, sadness)

[Aloraíny et al. (2018), ElSherief et al. (2018), Giner-Sorolla & Russell (2019), Haybron (2002), Mathew et al. (2018), Matsumoto et al. (2016), Sell et al. (2009)]



Outcome: LIWC categories

$$\mathrm{RTE}_s = \frac{\mathrm{Outcome}_{\mathrm{Treatment},s}}{\mathrm{Outcome}_{\mathrm{Control},s}}$$

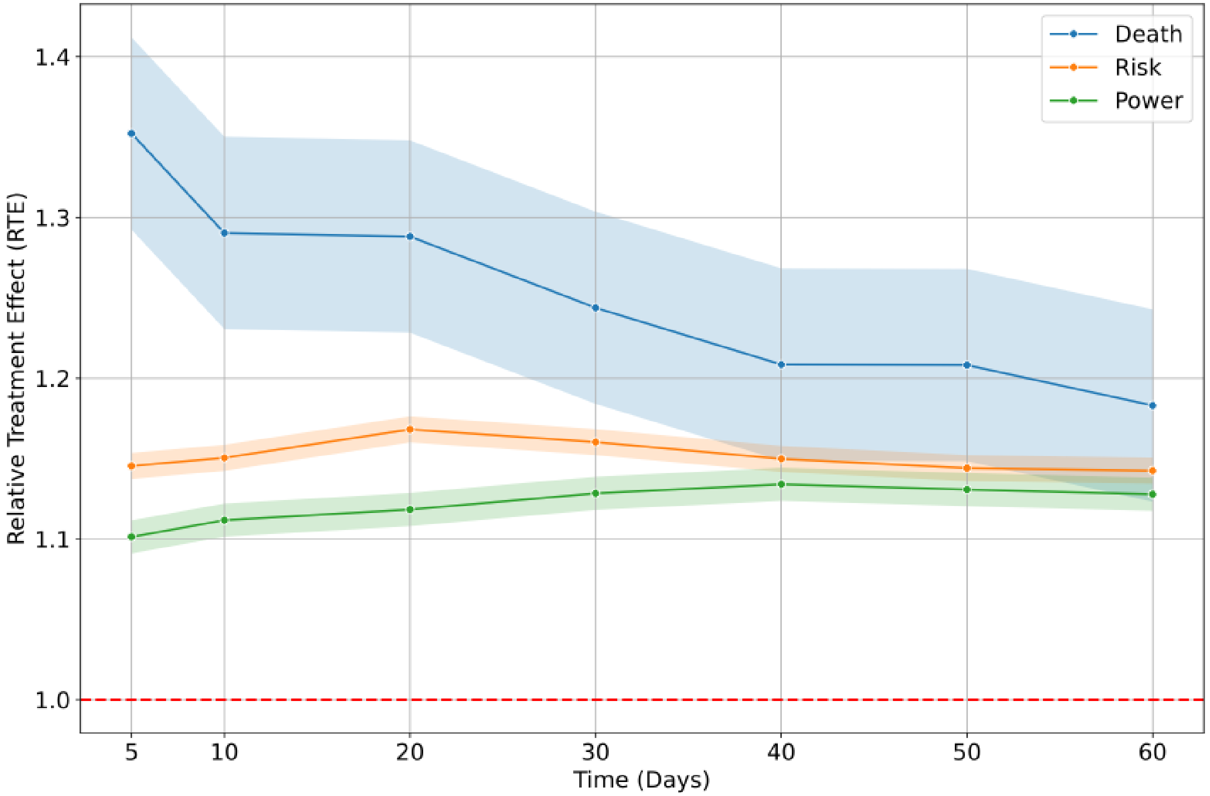
RTE > 1.0 indicates an **increase** in the outcome for the treatment compared to the control

Results

H1b

- Hate speech users use language related to **power, risk, and death**

[Elsherief et al. (2018), Goff et al. (2008), Markowitz & Slovic (2020), Paasch-Colberg et al. (2021)]



Outcome: LIWC categories

$$\mathrm{RTE}_s = \frac{\mathrm{Outcome}_{\mathrm{Treatment},s}}{\mathrm{Outcome}_{\mathrm{Control},s}}$$

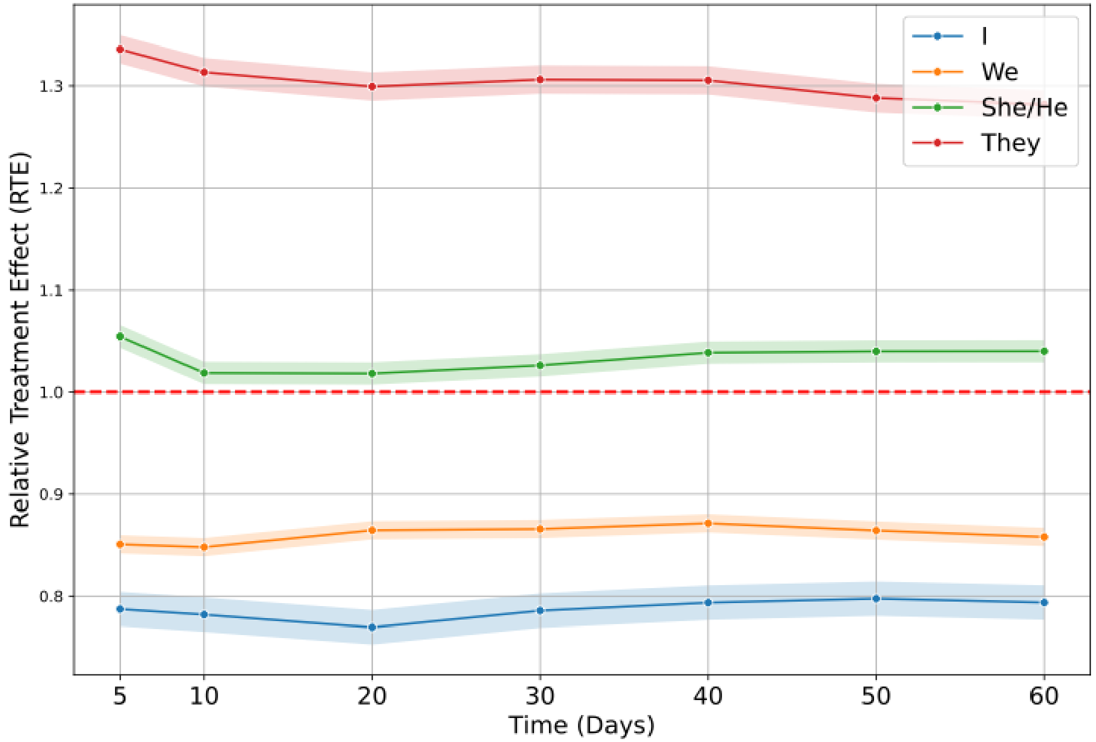
RTE > 1.0 indicates an **increase** in the outcome for the treatment compared to the control

Results

H1c

- Hate speech users employ more **third-person pronouns**, indicating detachment

[Elsherief et al. (2018), Faulkner & Bliuc (2018), Zannettou et al. (2020), Perdue et al. (1990), Shih et al. (2013), Matos & Miller (2023)]



Outcome: LIWC categories

$$\mathrm{RTE}_s = \frac{\mathrm{Outcome}_{\mathrm{Treatment},s}}{\mathrm{Outcome}_{\mathrm{Control},s}}$$

RTE > 1.0 indicates an **increase** in the outcome for the treatment compared to the control

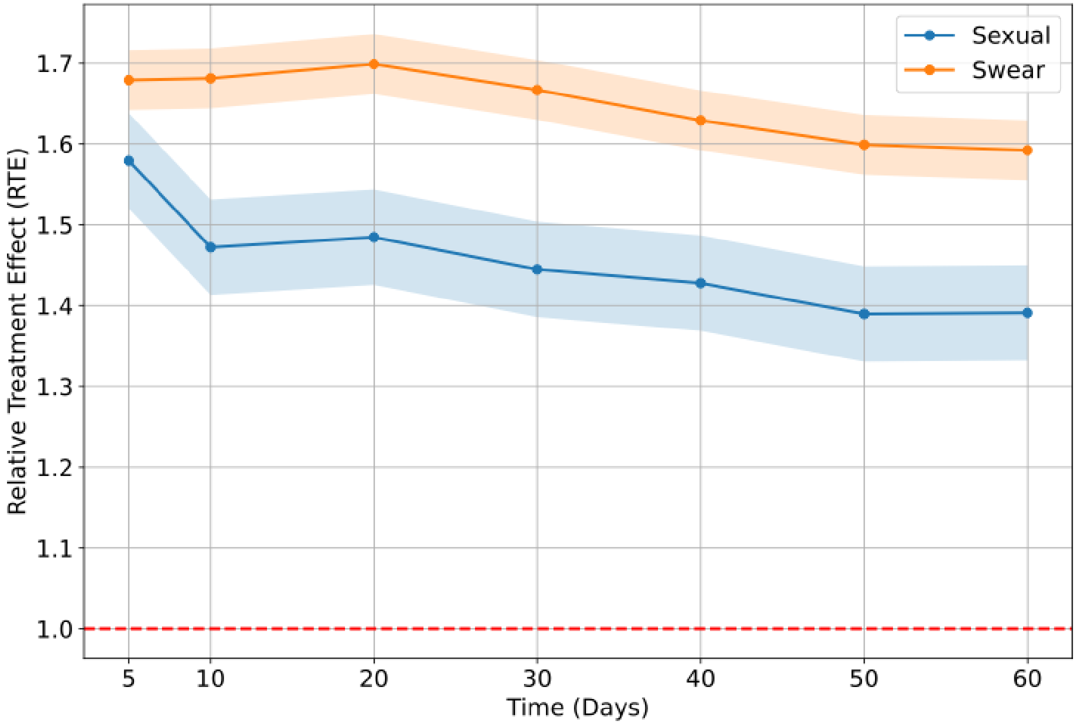
TED ROGERS SCHOOL OF MANAGEMENT

Results

H1d

- Hate speech involves more **profanity**

[Carter (1944), Leader et al. (2009), Bartlett et al. (2014), Bilewicz & Soral (2020), Jeshion (2013), Thurlow (2001), Anderson & Lepore (2013), Vallée (2014)]



Outcome: LIWC categories

$$\mathrm{RTE}_s = \frac{\mathrm{Outcome}_{\mathrm{Treatment},s}}{\mathrm{Outcome}_{\mathrm{Control},s}}$$

RTE > 1.0 indicates an **increase** in the outcome for the treatment compared to the control

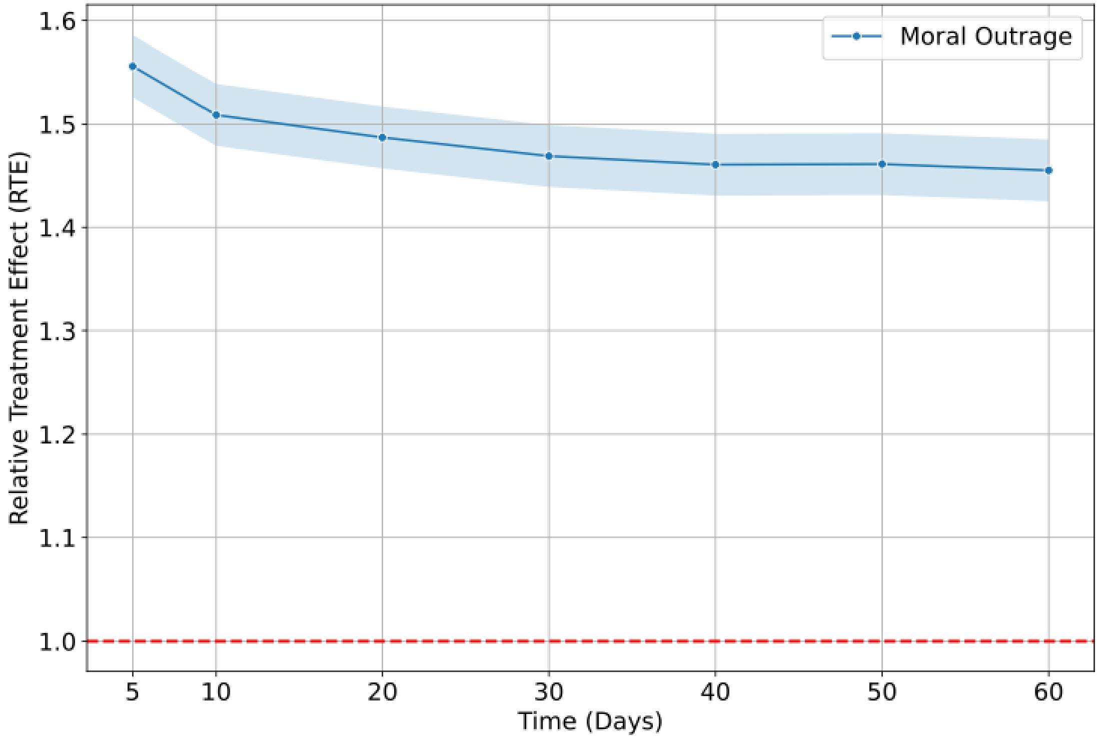
TED ROGERS SCHOOL OF MANAGEMENT

Results

H1e

- Hate speech is linked with **moral outrage** language

[Brady et al. (2021), Crockett (2017), Salerno & Peter-Hagene (2013), Grubbs et al. (2019), Young & Young (2020), Faulkner & Bliuc (2018), Solovev & Pröllochs (2023)]



Outcome: moral outrage classifier [Brady et al. (2021)]

$$\mathrm{RTE}_s = \frac{\mathrm{Outcome}_{\mathrm{Treatment},s}}{\mathrm{Outcome}_{\mathrm{Control},s}}$$

RTE > 1.0 indicates an **increase** in the outcome for the treatment compared to the control

Results

- H2 report results for stratum 5
  - 1,095 users: 614 hate speech users and 481 control
  - 631,504 total tweets
- Repeated experiments for the 4 largest stratum
- Consistent findings

TED ROGERS SCHOOL OF MANAGEMENT

Results

	Topic 1	Topic 2	Topic 3	...	Topic N
Tweet 1	0.20	0.10	0.50	...	0.05
Tweet 2	0.00	0.70	0.10	...	0.10
Tweet 3	0.15	0.20	0.00	...	0.30
...					
Tweet M	0.05	0.05	0.80	...	0.00

Document-Topic Matrix

	Word 1	Word 2	Word 3	...	Word K
Topic 1	0.05	0.30	0.15	...	0.02
Topic 2	0.10	0.00	0.20	...	0.05
Topic 3	0.25	0.10	0.00	...	0.15
...					
Topic N	0.00	0.05	0.30	...	0.10

Topic-Word Matrix

Results

Non- Hate Topics	Hate-related Topics
RT people COVID amp	RT China people
RT COVID coronavirus amp	Hong Kong protests
Baseball RF good like	Positive comments
Masks face wear ventilators	US politics
Job search resume help	Twitter lockdowns
Food quicker help meals	Bill Gates money
Michigan reopen stay home	UK bloggers
Music radio listen stayhome	Food and cooking
Social distancing mental health	Book promotion
Drawing art enjoy kids	Education
God bless and broadband	Growth and waves
Predictive analytics detect infection	Follow and unfollow
Eid stay home safe	Australian port
Automatically followed checked unfollowed	CEO experiences
Weight loss method fast	American hero
Tutoring supplemental reviews help	Welded doors
Court suspends constitution federal	Joger incident
Studied eastern philosophy hind	Temperature changes
US America Texas Alabama	Unemployment rate
Misidentified remains settlers swords	Redirects and links

BERTopic

Results

Controversial topics

Non- Hate Topics	Hate-related Topics
RT people COVID amp	**RT China people**
RT COVID coronavirus amp	**Hong Kong protests**
Baseball RF good like	Positive comments
Masks face wear ventilators	**US politics**
Job search resume help	**Twitter lockdowns**
Food quicker help meals	**Bill Gates money**
Michigan reopen stay home	UK bloggers
Music radio listen stayhome	Food and cooking
Social distancing mental health	Book promotion
Drawing art enjoy kids	Education
God bless and broadband	Growth and waves
Predictive analytics detect infection	Follow and unfollow
Eid stay home safe	Australian port
Automatically followed checked unfollowed	CEO experiences
Weight loss method fast	American hero
Tutoring supplemental reviews help	**Welded doors**
Court suspends constitution federal	Joger incident
Studied eastern philosophy hind	Temperature changes
US America Texas Alabama	Unemployment rate
Misidentified remains settlers swords	Redirects and links

BERTopic

TED ROGERS SCHOOL OF MANAGEMENT

Results

Neutral/positive topics

Non- Hate Topics	Hate-related Topics
RT people COVID amp	RT China people
RT COVID coronavirus amp	Hong Kong protests
Baseball RF good like	**Positive comments**
Masks face wear ventilators	US politics
Job search resume help	Twitter lockdowns
Food quicker help meals	Bill Gates money
Michigan reopen stay home	UK bloggers
Music radio listen stayhome	Food and cooking
Social distancing mental health	Book promotion
Drawing art enjoy kids	Education
God bless and broadband	Growth and waves
Predictive analytics detect infection	Follow and unfollow
Eid stay home safe	Australian port
Automatically followed checked unfollowed	CEO experiences
Weight loss method fast	American hero
Tutoring supplemental reviews help	Welded doors
Court suspends constitution federal	Joger incident
Studied eastern philosophy hind	Temperature changes
US America Texas Alabama	Unemployment rate
Misidentified remains settlers swords	Redirects and links

BERTopic

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]

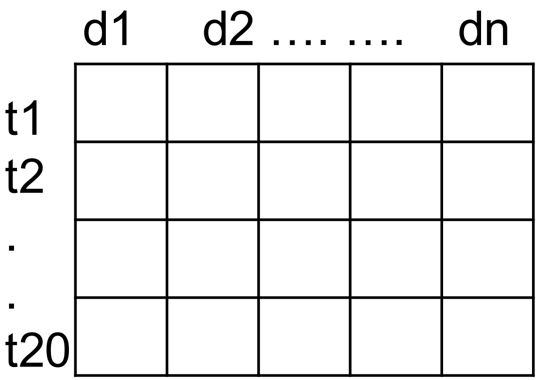
TED ROGERS SCHOOL OF MANAGEMENT

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]



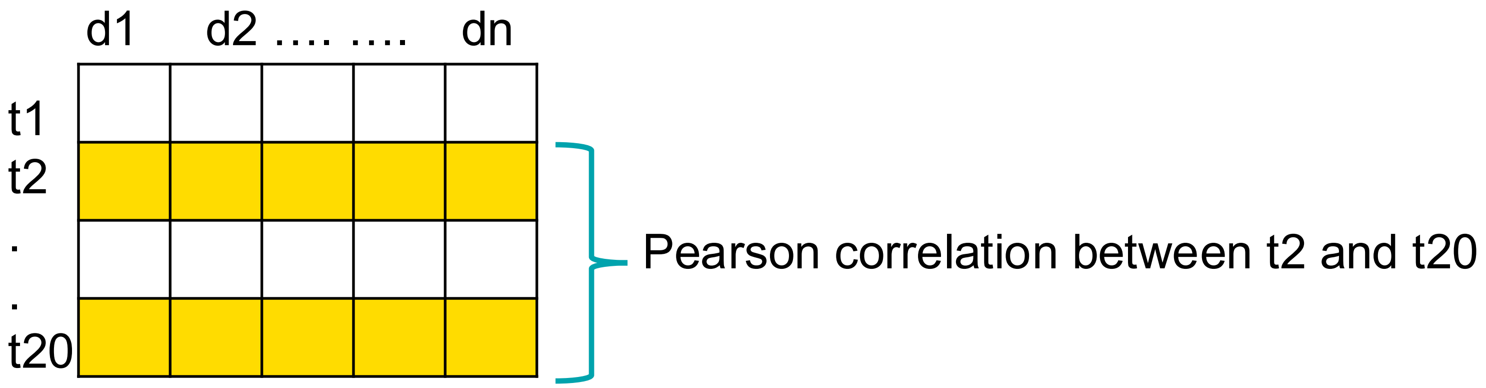
Document-Topic Matrix

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]



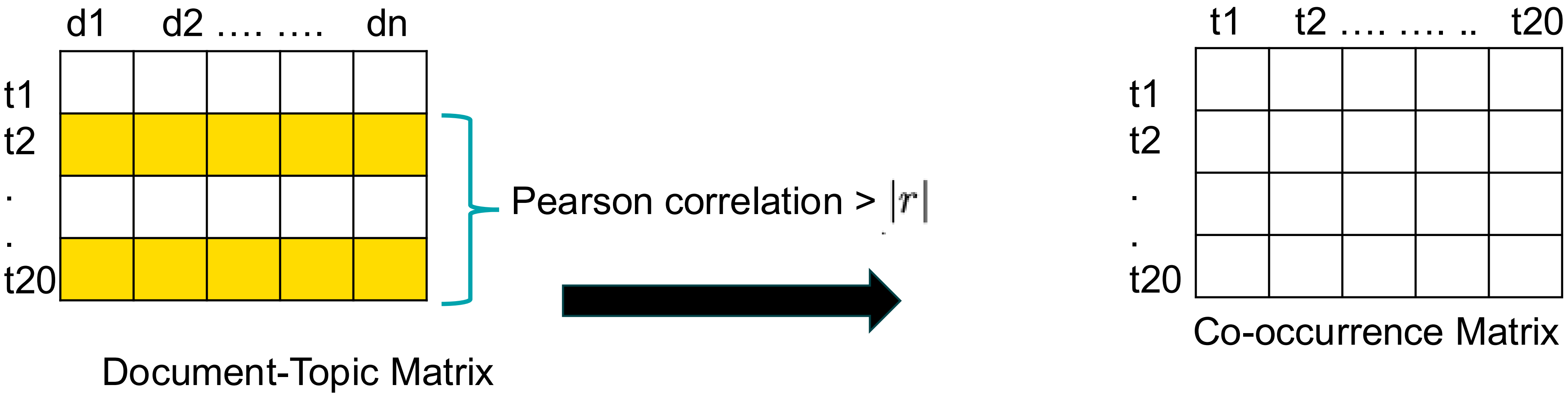
Document-Topic Matrix

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]



Results

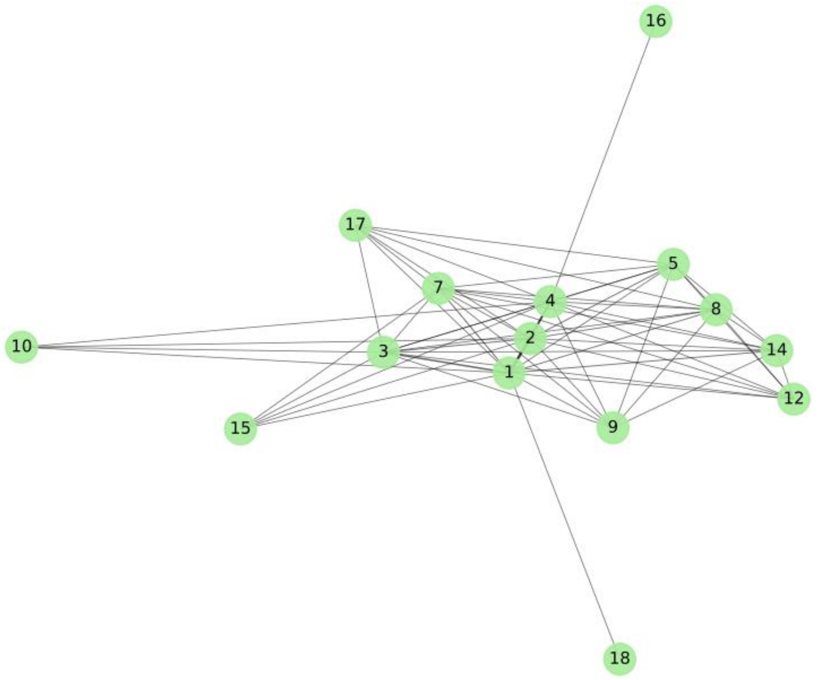
H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]

	t1	t2 …. …. ..			t20
t1					
t2					
.					
.					
t20					

Co-occurrence Matrix



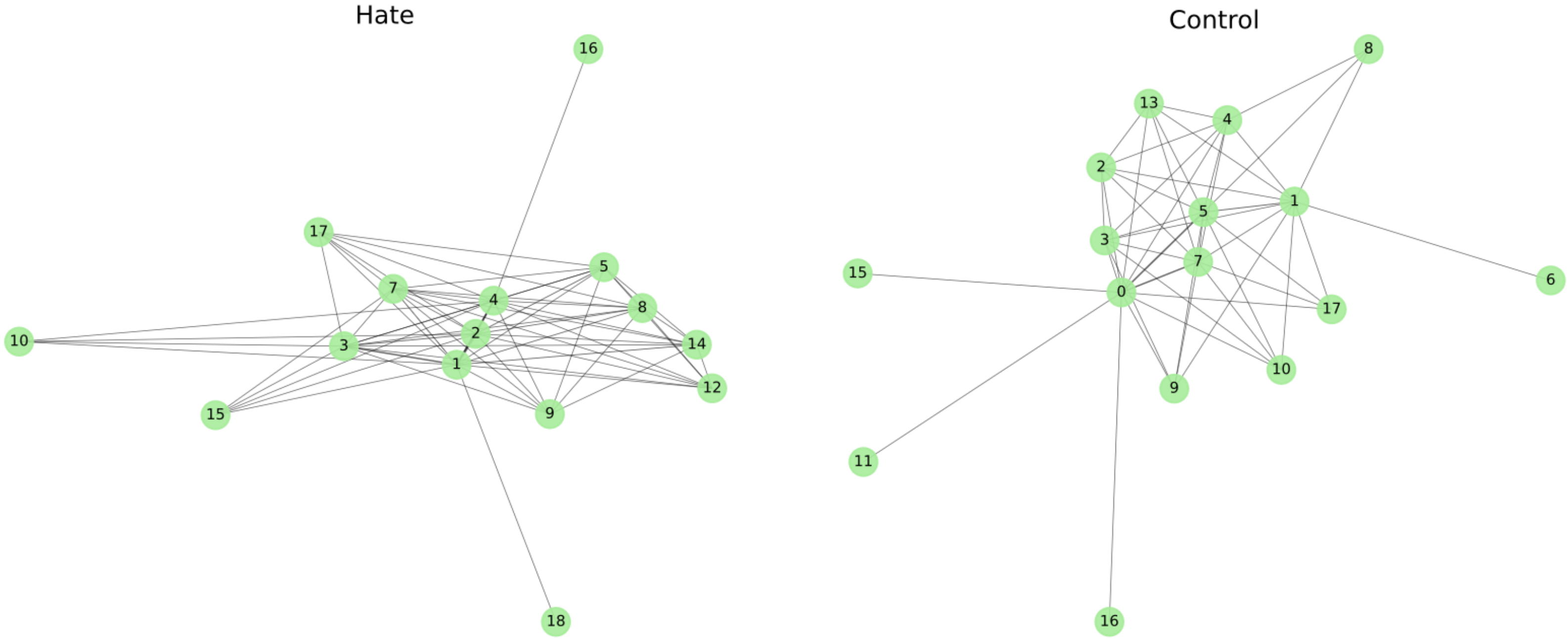
TED ROGERS SCHOOL OF MANAGEMENT

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]

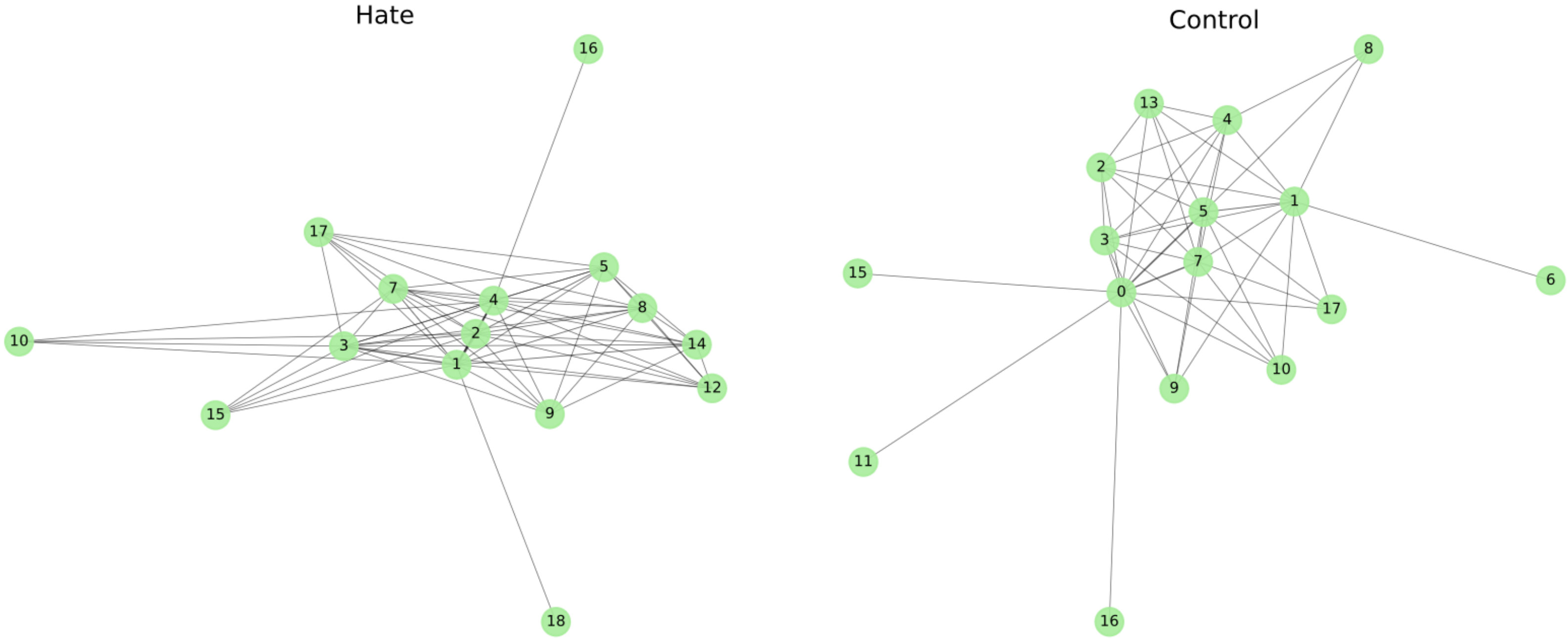


Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]



Entropy: nodes connected in random way

Clustering coefficient: how likely nodes are to be clustered together

Shortest path: average shortest path between nodes

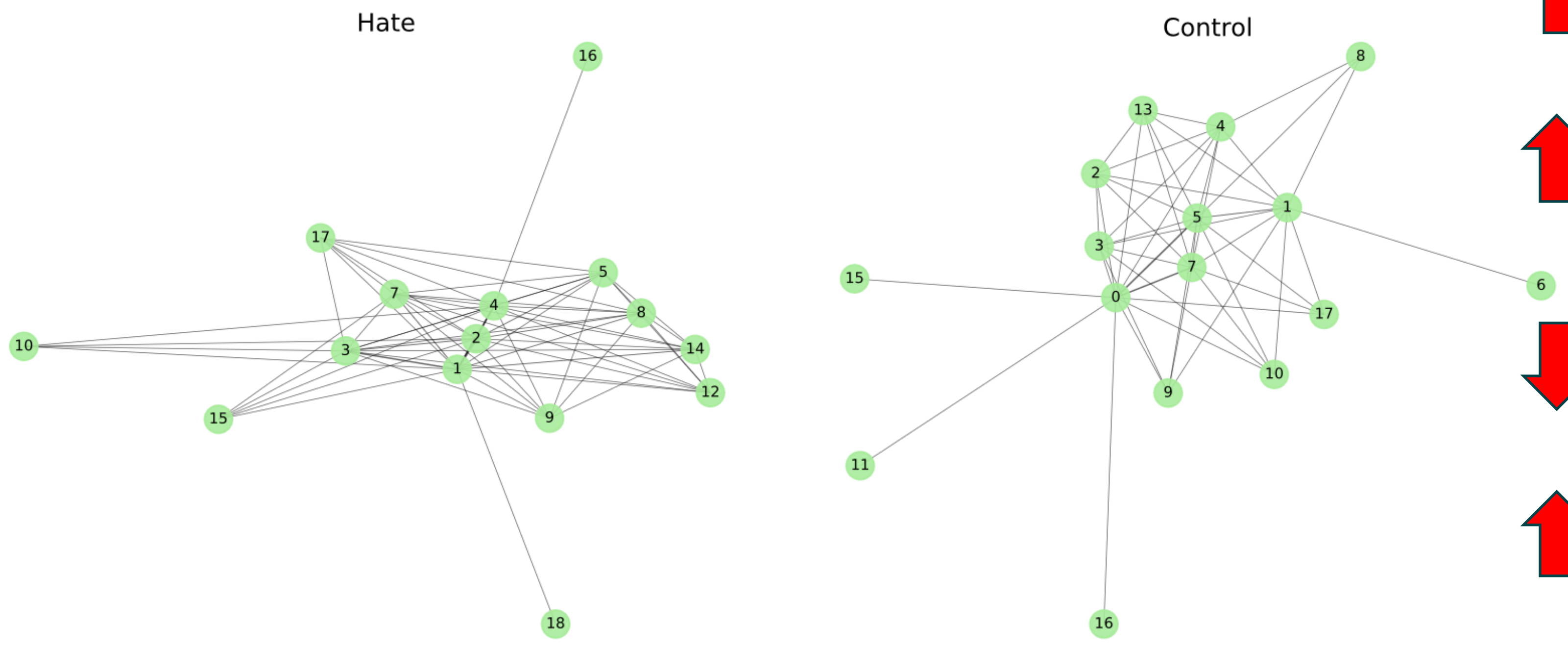
Density: ratio of actual edges to total possible edges

Results

H2a

- Hate speech exhibits a **tightly** connected network of related topics

[Papcunova et al. (2023), Salmela & Von Scheve (2017), Wood et al. (2012), Van Prooijen & Van Vugt (2018)]



Entropy: nodes connected in random way

Clustering coefficient: how likely nodes are to be clustered together

Shortest path: average shortest path between nodes

Density: ratio of actual edges to total possible edges

Hate- related topics are more interconnected than non-hate topics

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

Global cohesion: compare tweets together

Local cohesion: within each tweet (H2c)

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

Global cohesion: compare tweets together

Local cohesion: within each tweet (H2c)

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

	Word 1	Word 2	Word 3	...	Word N
Tweet 1	0.10	0.00	0.05	...	0.00
Tweet 2	0.00	0.15	0.00	...	0.08
Tweet 3	0.12	0.00	0.07	...	0.00
...					
Tweet M	0.00	0.09	0.00	...	0.05

TF-IDF Matrix

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

	Word 1	Word 2	Word 3	...	Word N
Tweet 1	0.10	0.00	0.05	...	0.00
Tweet 2	0.00	0.15	0.00	...	0.08
Tweet 3	0.12	0.00	0.07	...	0.00
...					
Tweet M	0.00	0.09	0.00	...	0.05

TF-IDF Matrix

cosine similarity →

tweet1 tweet2….tweetM

tweet1					
ttweet2					
.					
.					
tweet M					

cosine similarity Matrix

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

	Word 1	Word 2	Word 3	...	Word N
Tweet 1	0.10	0.00	0.05	...	0.00
Tweet 2	0.00	0.15	0.00	...	0.08
Tweet 3	0.12	0.00	0.07	...	0.00
...					
Tweet M	0.00	0.09	0.00	...	0.05

TF-IDF Matrix

cosine similarity →

tweet1 tweet2….tweetM

tweet1					
ttweet2					
.					
.					
tweet M					

cosine similarity Matrix

linear mixed-effects model to test for significance:
Cousin similarity ~ tweet_type + word_count + (1 | user_id])

Beta = 0.001, SE < 0.0001, t-value = 39.06, p-value < 0.001, R2m/c = 0.05/0.26

TED ROGERS SCHOOL OF MANAGEMENT

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

Global cohesion: compare tweets together

Local cohesion: within each tweet (H2c)

Hate- related topics show <u>high global</u> coherence than non-hate topics

Results

H2c

- Hate speech narratives display **lower** topic specificity

[Suedfeld & Tetlock (1977), Jakob et al. (2023), Faulkner & Bliuc (2018), Gregory & Piff (2021), Dhont & Hodson (2014), Hodson & Busseri (2012)]

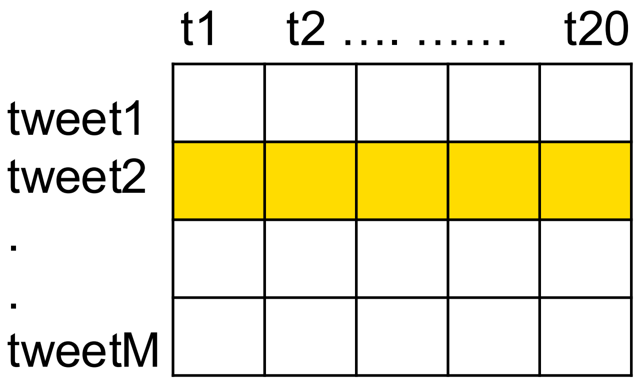
TED ROGERS SCHOOL OF MANAGEMENT

Results

H2c

- Hate speech narratives display **lower** topic specificity

[Suedfeld & Tetlock (1977), Jakob et al. (2023), Faulkner & Bliuc (2018), Gregory & Piff (2021), Dhont & Hodson (2014), Hodson & Busseri (2012)]



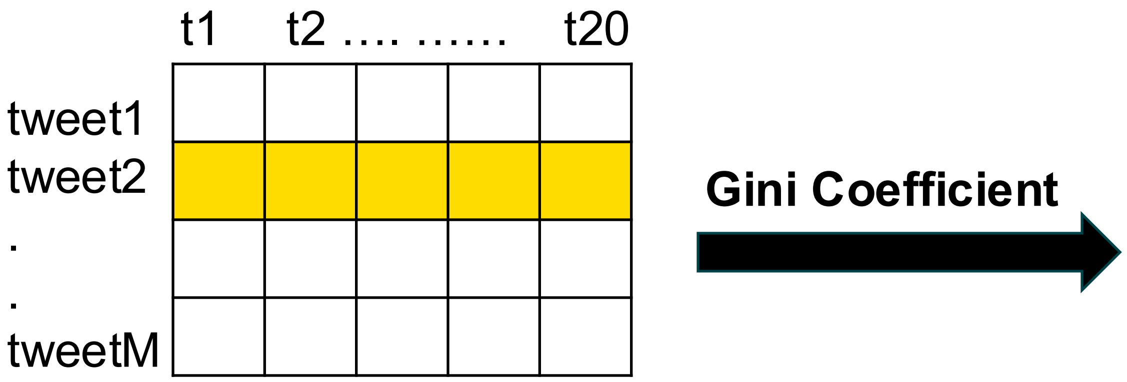
Topic distribution Matrix

Results

H2c

- Hate speech narratives display **lower** topic specificity

[Suedfeld & Tetlock (1977), Jakob et al. (2023), Faulkner & Bliuc (2018), Gregory & Piff (2021), Dhont & Hodson (2014), Hodson & Busseri (2012)]



Topic distribution Matrix

[0.7, 0.1, 0.05,..0.03] >> unequal distribution>> **high** Gini coefficient

[0.1, 0.1, 0.05,..0.13] >> equal distribution>> **low** Gini coefficient

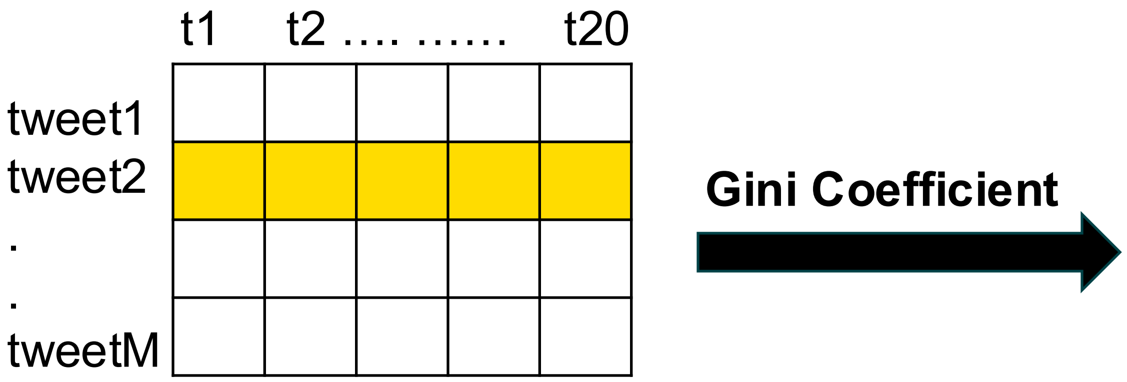
TED ROGERS SCHOOL OF MANAGEMENT

Results

H2c

- Hate speech narratives display **lower** topic specificity

[Suedfeld & Tetlock (1977), Jakob et al. (2023), Faulkner & Bliuc (2018), Gregory & Piff (2021), Dhont & Hodson (2014), Hodson & Busseri (2012)]



Topic distribution Matrix

[0.7, 0.1, 0.05,..0.03] >> unequal distribution>> **high** Gini coefficient

[0.1, 0.1, 0.05,..0.13] >> equal distribution>> **low** Gini coefficient

linear mixed-effects model to test for significance:
Gini coefficient ~ tweet_type + word_count + (1 | user_id])

Beta = -0.004, SE < 0.001, t-value = -12.33, p-value < 0.001. The R2m/c is 0.01/ 0.17

Results

H2b

- Hate speech tweets show **lower coherence**

[Lewandowsky et al. (2018), Miani et al. (2022), Goertzel (1994), Swami et al. (2010), Douglas et al. (2017)]

Global cohesion: compare tweets together

Local cohesion: within each tweet (H2c)

Hate- related topics show <u>low local</u> coherence than non-hate topics

TED ROGERS SCHOOL OF MANAGEMENT

What we learnt

Linguistic differences

Cognitive differences

Narrative cohesion

Implications

Practical
- Content moderation
- Emotional engagement
- Support for targeted users

Theoretical
- Network and cohesion
- Emotional content & diffusion
- Research novelty

TED ROGERS SCHOOL OF MANAGEMENT

Future Work

- Broader Platform Analysis
- Longitudinal Studies
- Cross-Cultural Analysis
- Intervention Strategies

TED ROGERS SCHOOL OF MANAGEMENT

Funding:

- TRSM Research Development Grant
- TRSM Matching Funds
- NSERC DG

Collaborators:

- **Zeinab Noorian**

Assistant Professor, TRSM, Toronto Metropolitan University

- **Hadiseh Moradisani**

PhD student, School of Engineering, University of Guelph

- **Pariya Abadeh**

MSc student, School of Engineering, University of Guelph

- **Caroline Erentzen**

Assistant Professor, Department of Psychology, Toronto Metropolitan University

- **Fattane Zarrinkalam**

Assistant Professor, School of Engineering, University of Guelph