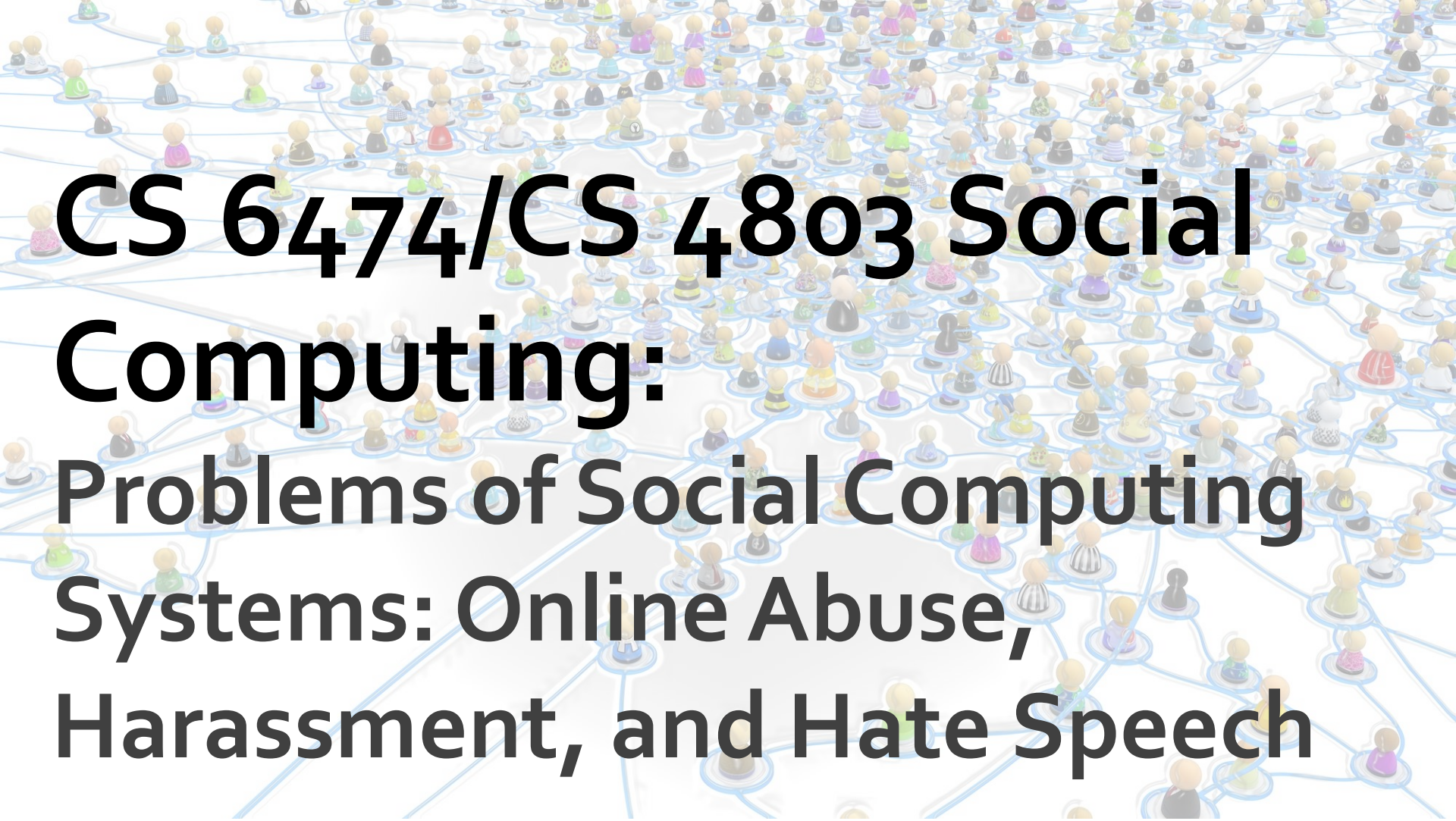


CS 6474/CS 4803 Social Computing:

Problems of Social Computing Systems: Online Abuse, Harassment, and Hate Speech

Munmun De Choudhury

munmund@gatech.edu

Week 9 | March 5, 2025

Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions

News headlines

“How trolls are ruining the internet”

Time (2016)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

News headlines

"How trolls are ruining the internet"

"When will the internet be safe for women?"

The Atlantic (2016)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

News headlines

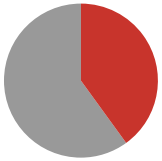
"How trolls are ruining the internet"

"When will the internet be safe for women?"

"Furious trolls are everywhere"

Salon (2014)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil



40% of online users have been harassed

Pew Research (2014)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

More headlines

“Why we’re shutting off our comments”

Popular Science (2013)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

More headlines

"Why we're shutting off our comments"

"We're turning comments off for a while"

The Verge (2013)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

More headlines

"Why we're shutting off our comments"

"We're turning comments off for a while"

"Sick of internet comments? Us, too"

Chicago Sun-Times (2014)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

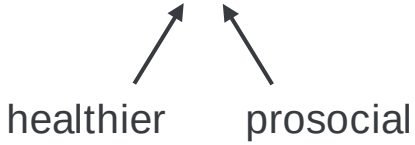
RQ

Why is trolling so prevalent?

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Implication

Understanding trolling lets us design ❤ communities



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

What is trolling?

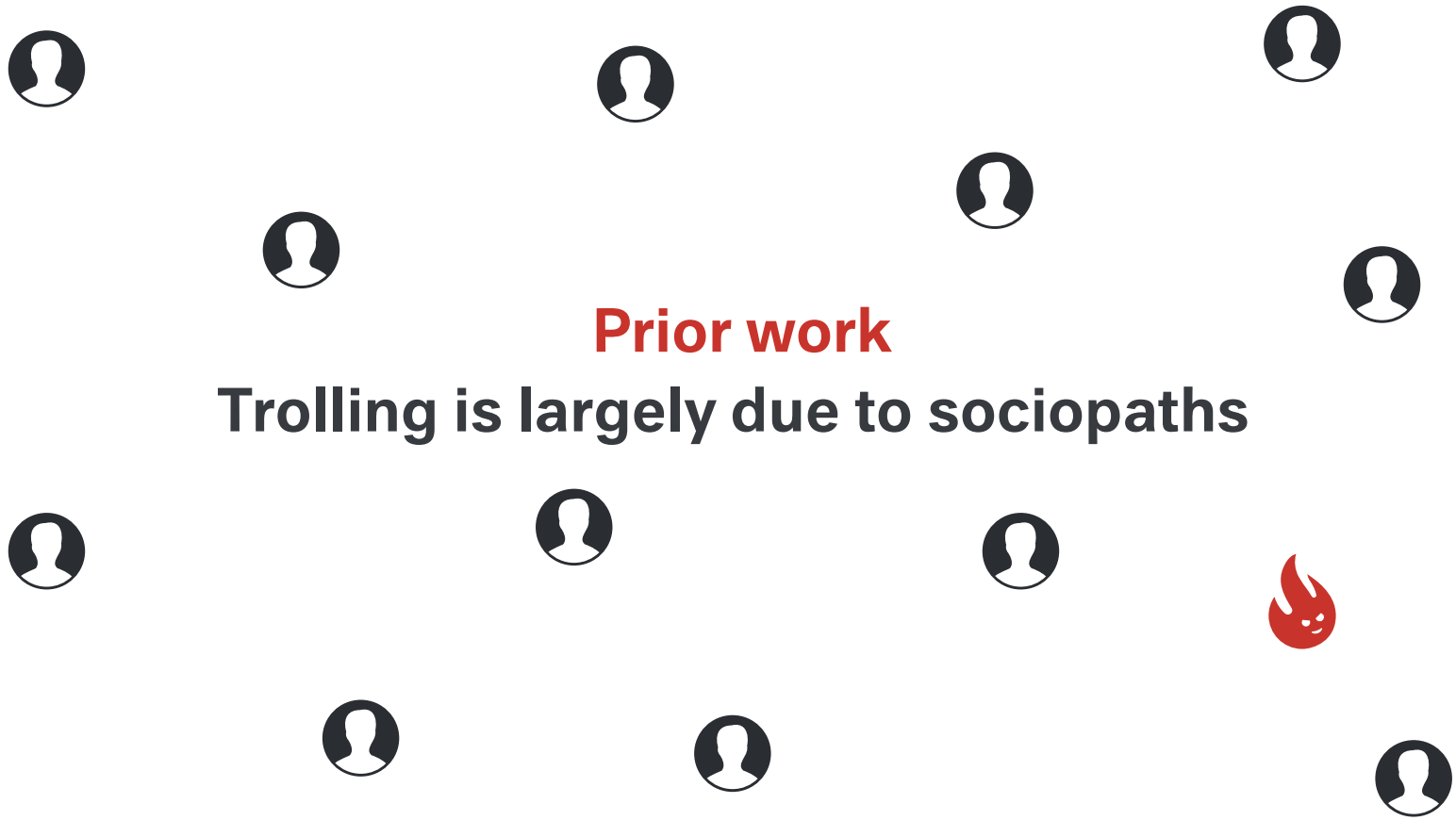
1. Engaging in negatively marked online behavior?

2. Not following the rules?

3. Taking pleasure in upsetting others?

Trolling is behavior outside community norms.

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

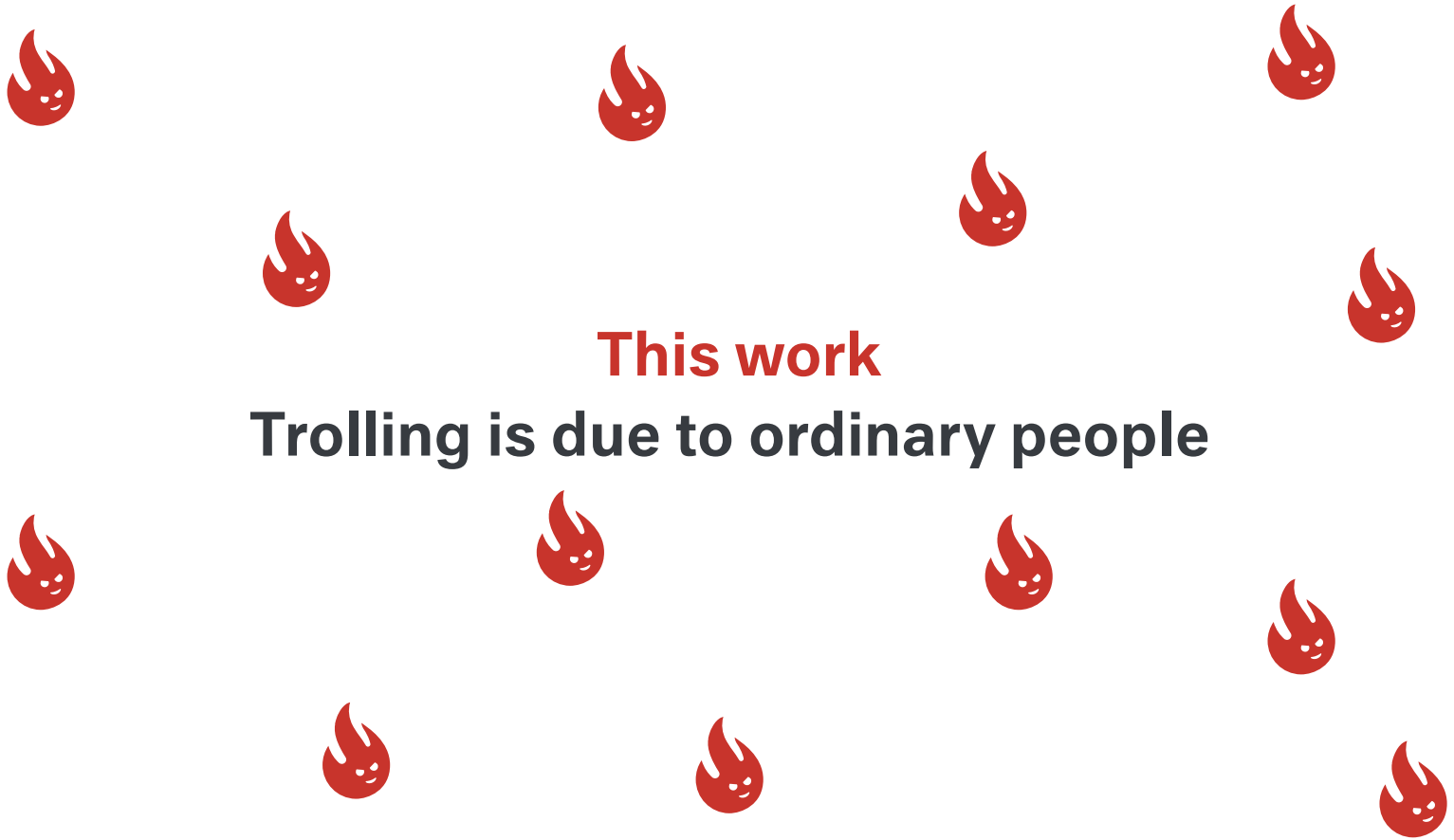


Prior work

Trolling is largely due to sociopaths

Donath (1999); Hardaker (2010); Buckels, et al. (2014)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil



This work

Trolling is due to ordinary people

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Data

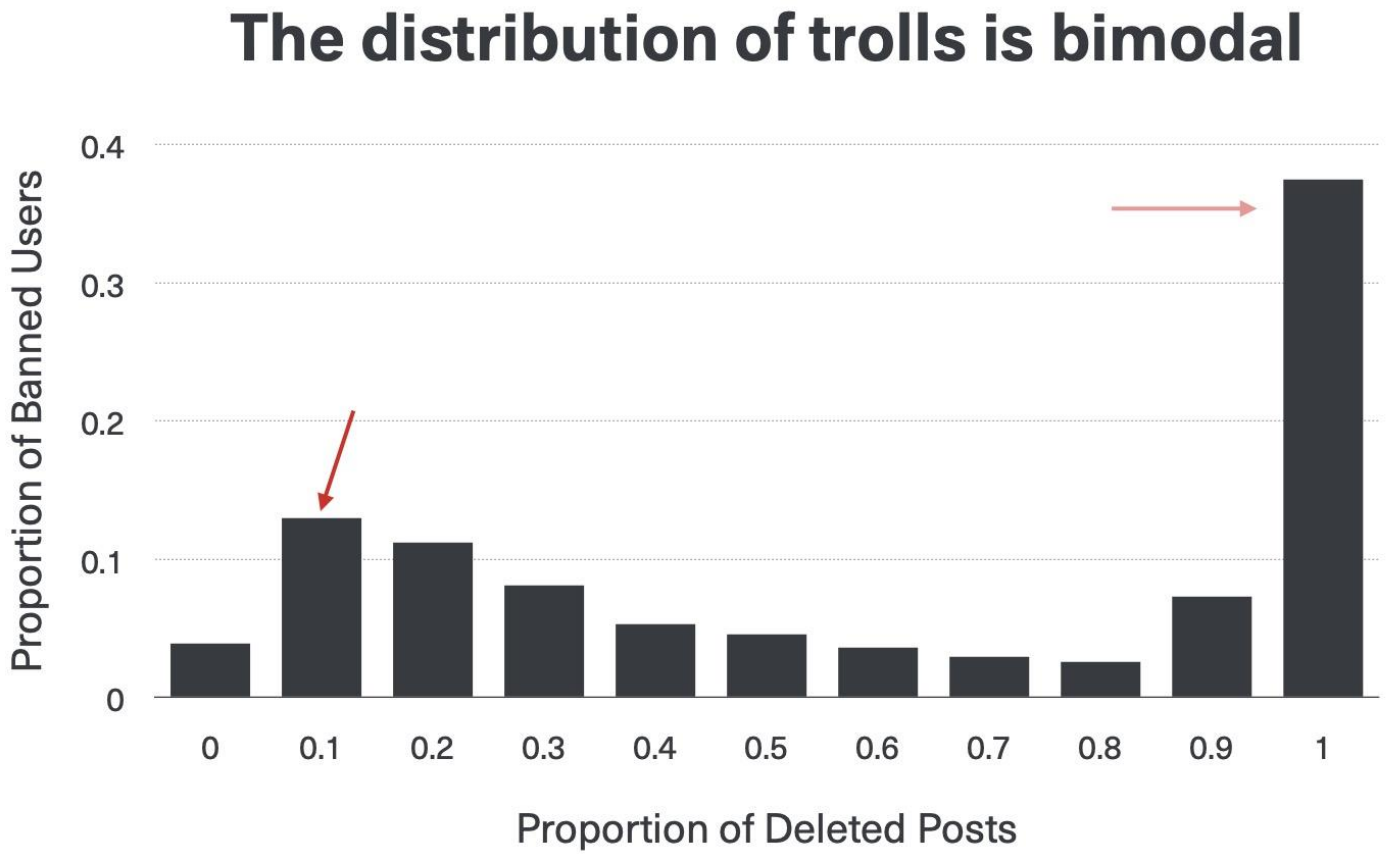
16M posts on 16K articles from 

.com

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

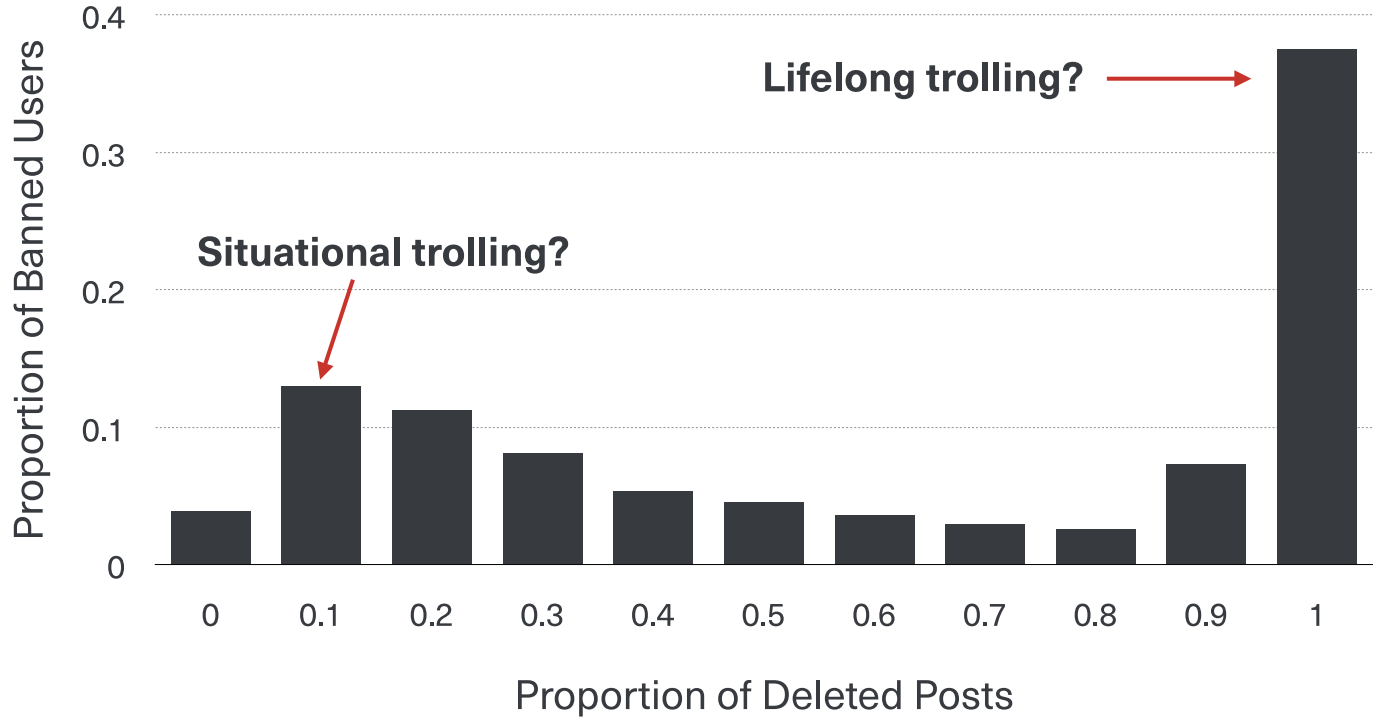
How much do trolls troll?

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Are there two types of trolls?



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

How to show that trolling is situational?

Observational data isn't causal

Experiments are hard to generalize

Solution: online experiment + observational study

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Online Experiment Overview

Online experiment simulating a discussion forum

Complete a quiz, then participate in a discussion

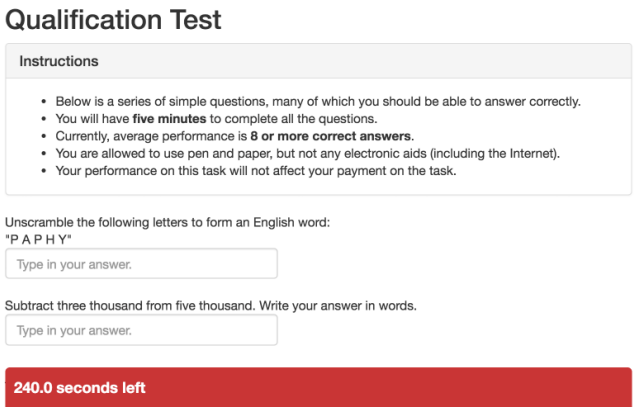
We manipulated quiz difficulty and discussion context

The quiz was either easy or difficult

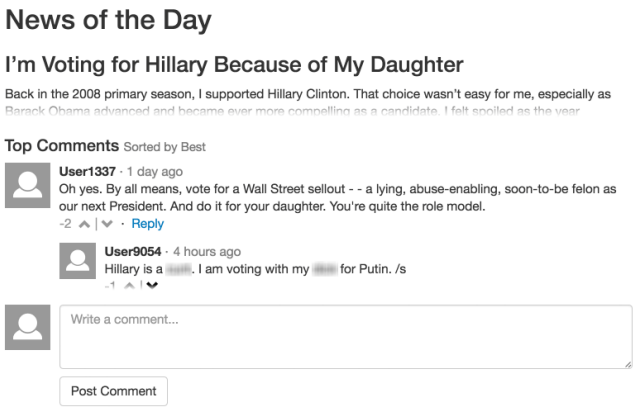
Discussion context was either positive or negative

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Online Experiment Overview



(a)



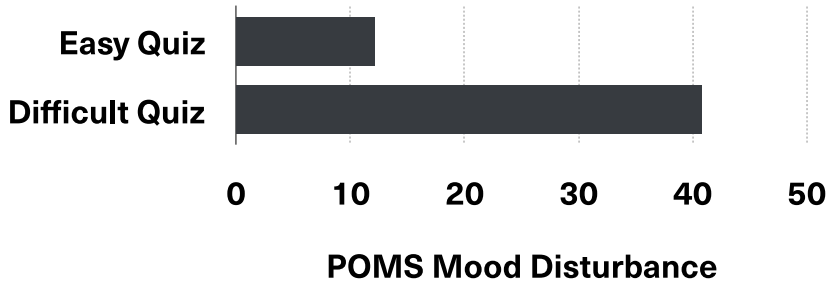
(b)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Data Analysis: Understanding Mood

Manipulation checks

People were in a worse mood after the difficult quiz



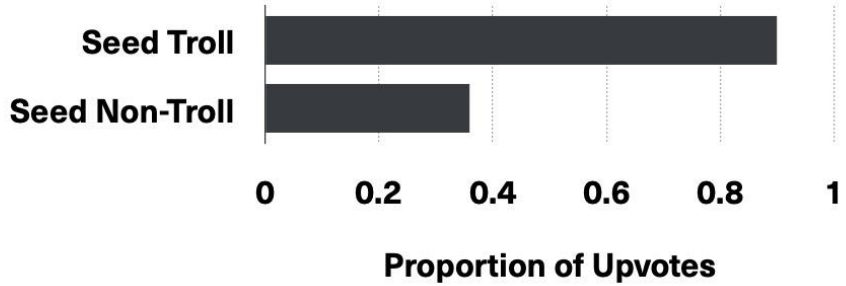
Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Data Analysis: Understanding Discussion Context

Manipulation checks

People were in a worse mood after the difficult quiz

People also perceived seed troll posts as worse



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Trolling almost doubles in the negative mood and context condition

% Troll Posts

	Positive Mood	Negative Mood
Positive Context	35%	49%
Negative Context	47%	**68%**

($p < 0.05$ using a mixed effects logistic regression model)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Negative affect also triples

% Neg. Affect Words (LIWC)

	Positive Mood	Negative Mood
Positive Context	1.1%	1.4%
Negative Context	2.3%	2.9%

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

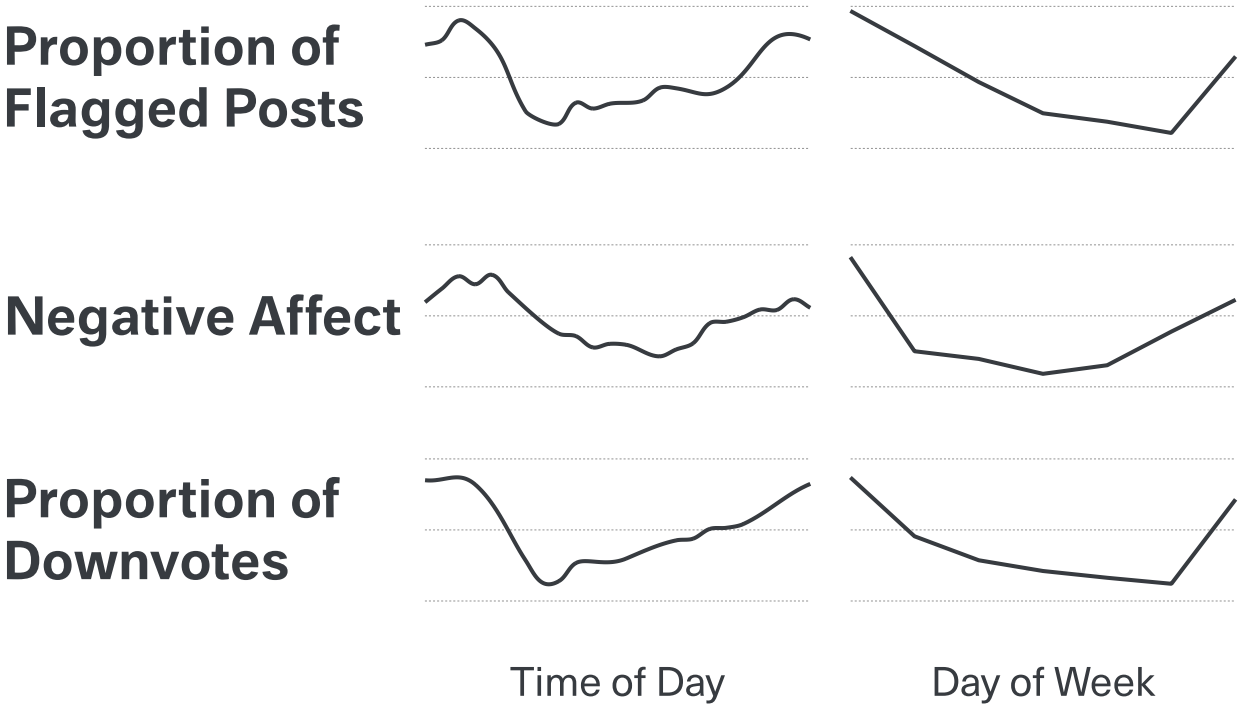
Bad mood and **negative context** increase trolling

But does this generalize?

Online experiment + **Observational study**

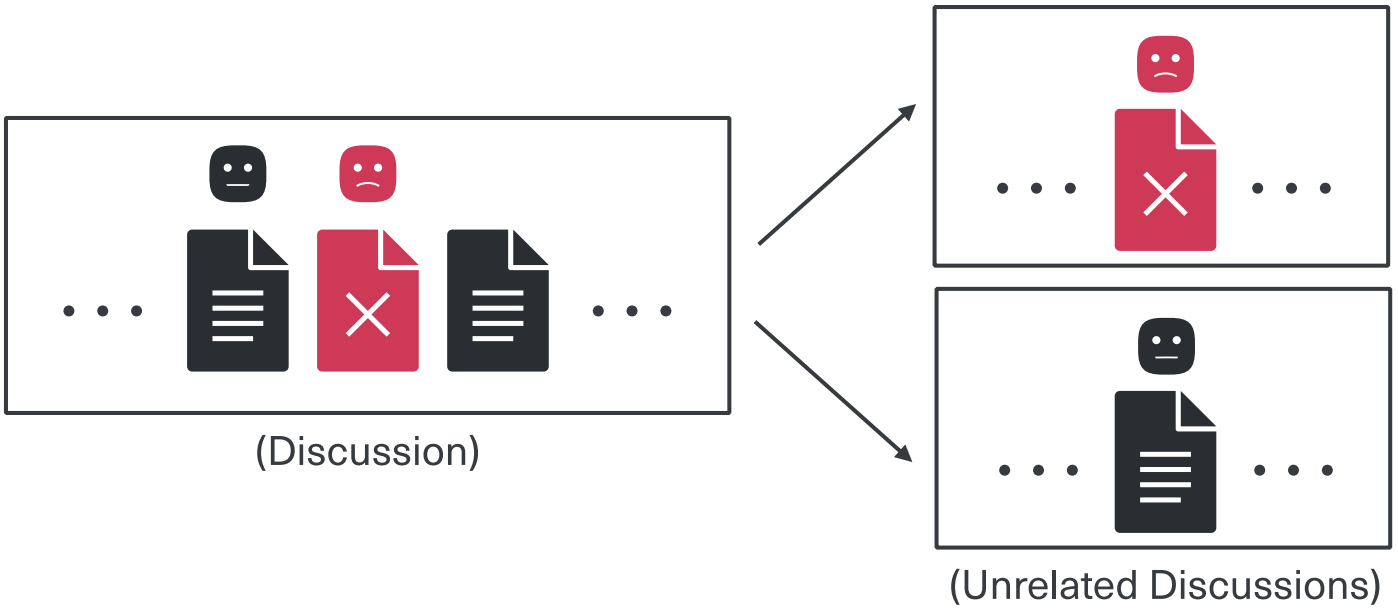
Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Trolling peaks when moods are worse



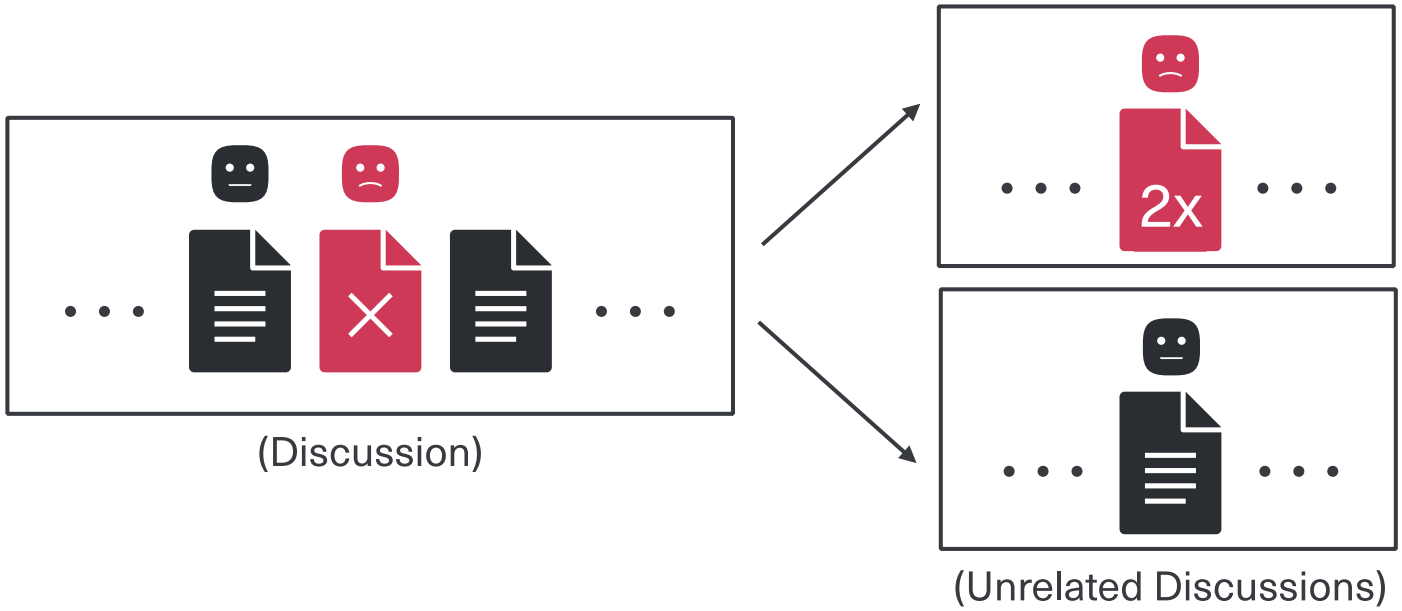
Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Mood spills over from prior discussions



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Trolling is twice as likely in unrelated discussions



($p < 0.01$)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

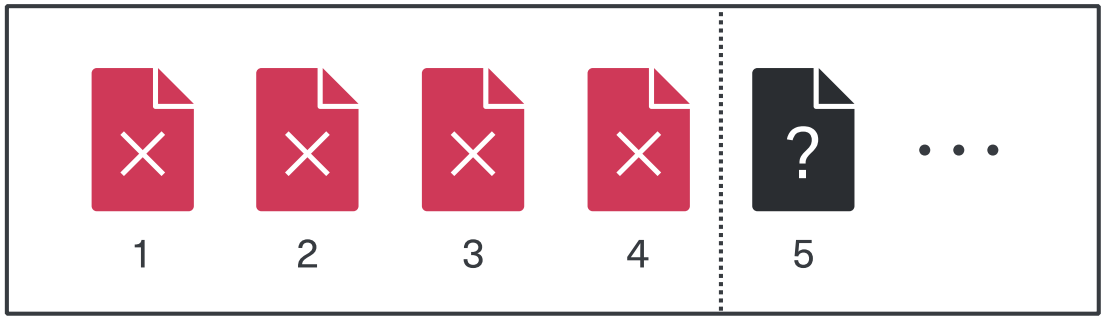
An initial post increases later trolling by over 1.5x

(Separate discussions of same article)

($p < 0.01$)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

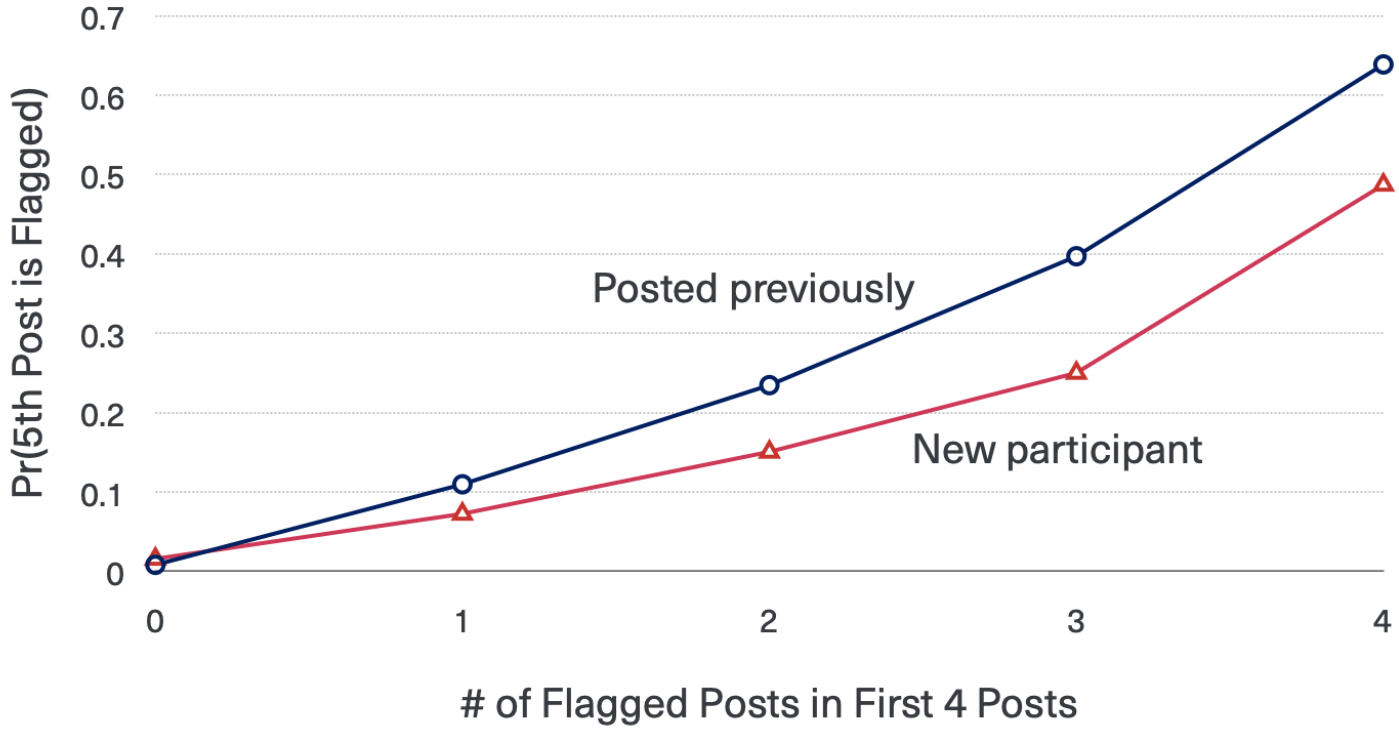
Does increased trolling have an additive effect?



(Discussion)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

More initial trolling means more future trolling

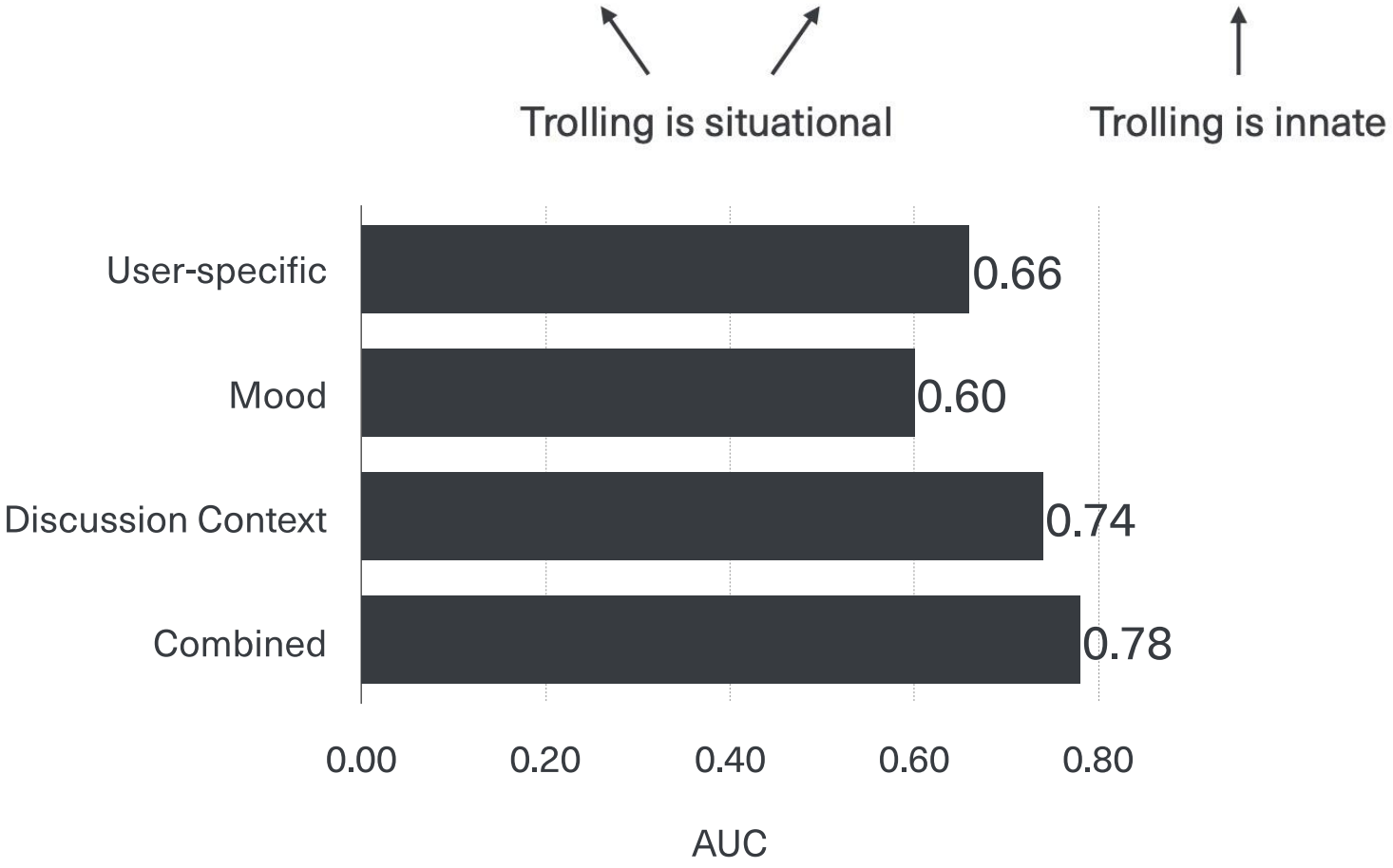


Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Can we predict trolling before it happens?

Logistic regression on 120K posts

Three sets of features: mood, context, or user-specific



Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Implications for Designing Better Discussion Platforms

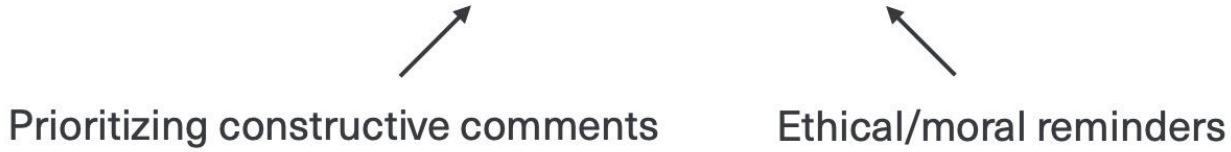
Couldn't we just ban trolls?

But many "trolls" are ordinary people!

Important to also curb situational trolling:

Through inferring mood…

Or altering the context of a discussion.



Mazar, Amir, & Ariely (2008)

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Class Discussion

Should platforms proactively alter the context of discussions? Who is the right stakeholder to do so?

How can the context of discussion be altered (technically) without taking away the essence of the discourse?

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Class Discussion

Should the same policy apply to all trolls, or the innately antisocial users need a different moderation policy?

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

Class Discussion

How should we determine who is a troll and when are they trolling?

Slide courtesy: Justin Cheng, borrowed from Danescu-Niculescu-Mizil

When Online Harassment Is Perceived as Justified

Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, Cliff Lampe
University of Michigan School of Information
{lindsay.blackwell, cchent, sarita.schoenebeck, cacl}@umich.edu

Abstract

Most models of criminal justice seek to identify and punish offenders. However, these models break down in online environments, where offenders can hide behind anonymity and lagging legal systems. As a result, people turn to their own moral codes to sanction perceived offenses. Unfortunately, this vigilante justice is motivated by retribution, often resulting in personal attacks, public shaming, and doxing—behaviors known as online harassment. We conducted two online experiments (n=160; n=432) to test the relationship between retribution and the perception of online harassment as appropriate, justified, and deserved. Study 1 tested attitudes about online harassment when directed toward a woman who has stolen from an elderly couple. Study 2 tested the effects of social conformity and bystander intervention. We find that people believe online harassment is more deserved and more justified—but not more appropriate—when the target has committed some offense. Promisingly, we find that exposure to a bystander intervention reduces this perception. We discuss alternative approaches and designs for responding to harassment online.

Introduction

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users. This work is motivated by recent examples of harassment in online contexts that, although broadly viewed as harmful, are considered by some as justifiable responses to perceived social norm violations—a controversial form of social sanctioning. This "retributive harassment" can take many forms: high-profile examples include the 2013 public shaming of public relations executive Justine Sacco, the 2015 release of 40 million Ashley Madison users' personal and financial information, or the 2017 doxing of people who attended a white supremacist rally in Charlottesville, Virginia. Retributive harassment is especially widespread on social media sites such as Facebook and Twitter; however, why it happens and how to prevent it remain unknown.

Historically, abusive behavior online has been relegated to fringe cases—"narcissists, psychopaths, and sadists" (Buckels, Trapnell, and Paulhus 2014) who are either exceptions themselves, or inhabit atypical parts of the internet. Today, however, almost half of adult internet users in the U.S. have personally experienced online harassment, and a majority of users have witnessed others being harassed online (Duggan 2014; Duggan 2017; Lenhart et al. 2016; Rainie, Anderson, and Albright 2017). Although policies, reporting tools, and moderation strategies are improving (e.g., Perez 2017), most online platforms have failed to effectively curb harassing behaviors (Lenhart et al. 2016; Rainie, Anderson, and Albright 2017), and internet users and experts alike believe the problem is only getting worse (Rainie, Anderson, and Albright 2017).

This research aims to understand online harassment using a *retributive justice* framework. Retributive justice refers to a theory of punishment in which individuals who knowingly commit an act deemed to be morally wrong receive a proportional punishment for their misdeeds, sometimes referred to as "an eye for an eye" (Carlsmith and Darley 2008; Walen 2015). Retributive justice relies upon the assumption that everyday citizens possess intuitive judgments of "deservingness" that accurately and consistently express the degree of moral wrongdoing of others' acts. The integration of theories about justice and punishment with existing knowledge about social deviance and sanctioning has the potential to transform our current understanding of misbehavior in online spaces—in particular, when an instance of online harassment is perceived to be justified.

We conducted two online experiments to test the relationship between retributive justice and the perception of online harassment as justified or deserved. The first experiment tested whether exposure to a retributive prime—i.e., that the person being harassed had committed a crime—increases the belief that harassment is justified, deserved, or appropriate. The second experiment tested the effects of social influence on online harassment; specifically, whether conformity increases the belief that harassment is justified, deserved, or appropriate, and whether or not the presence of a bystander intervention would reduce these beliefs.

Investigating the relationship between orientations of justice and the perception of harassing behaviors online is an important step in better understanding what may motivate users to perpetrate online harassment—as well as what

Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis

Building the COVID-HATE Dataset

- Over 206 million tweets collected.
- Utilization of specific keywords for data collection
- Formation of a social network with more than 127 million nodes.
- Duration: January 15, 2020, to March 26, 2021.

Category	Keywords
COVID-19 Hate keywords	coronavirus, covid 19, covid-19, covid19, corona virus #CCPVirus, #ChinaDidThis, #ChinaLiedPeopleDied, #ChinaVirus, #ChineseVirus, chinese virus, #ChineseBioterrorism, #FuckChina, #KungFlu, #MakeChinaPay, #wuhanflu, #wuhanvirus, wuhan virus, chink, chinky, chonky, churka, cina, cokin, communistvirus, coolie, dink, niakoué, pastel de flango, slant, slant eye, slopehead, ting tong, yokel
Counterspeech keywords	#IAmNotAVirus, #WashTheHate, #RacismIsAVirus, #IAmNotCovid19, #BeCool2Asians, #StopAAPIHate, #ActToChange, #HateIsAVirus

Annotating Tweets: Hate, Counterspeech, Neutral

- Definition of hate, counterspeech, and neutral tweets.
- Process and results of hand-annotating 3355 tweets.

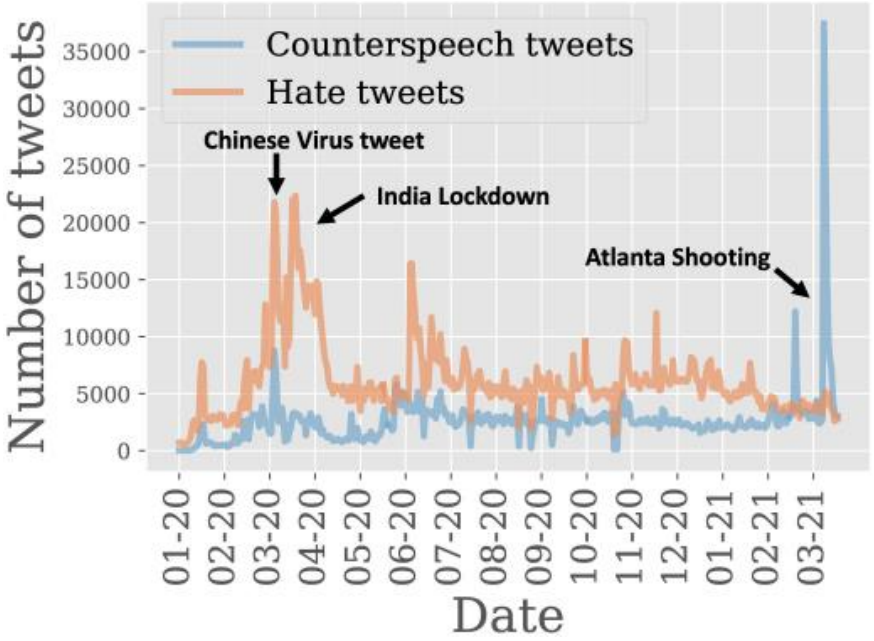
Classification

- Text classifier using BERT embeddings.
- Classifier's performance metrics.

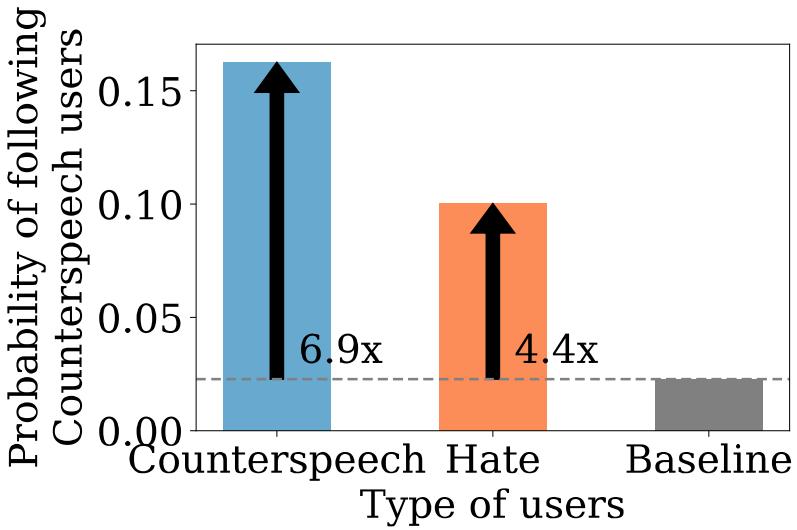
Feature set	Precision	Recall	F1 score
Anti-Asian hate tweet detection			
Linguistic	0.541	0.233	0.323
Hashtag	0.100	0.002	0.005
BERT	0.765	0.760	0.762
Counterspeech tweet detection			
Linguistic	0.483	0.189	0.267
Hashtag	0.800	0.029	0.056
BERT	0.839	0.868	0.853
Neutral tweet detection			
Linguistic	0.632	0.891	0.739
Hashtag	0.591	0.999	0.743
BERT	0.886	0.874	0.880

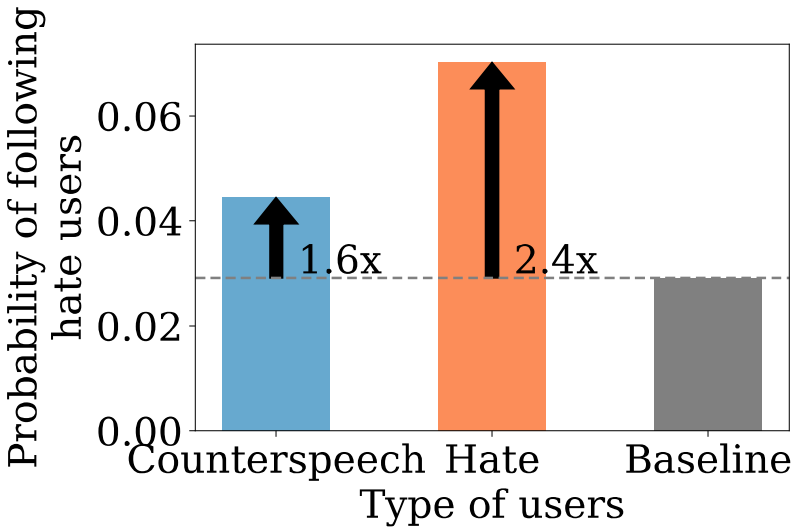
Trends Over Time: Hate vs. Counterspeech

- Analysis of hate and counterspeech tweet volumes over 14 months.
- Impact of specific events on tweet volumes.

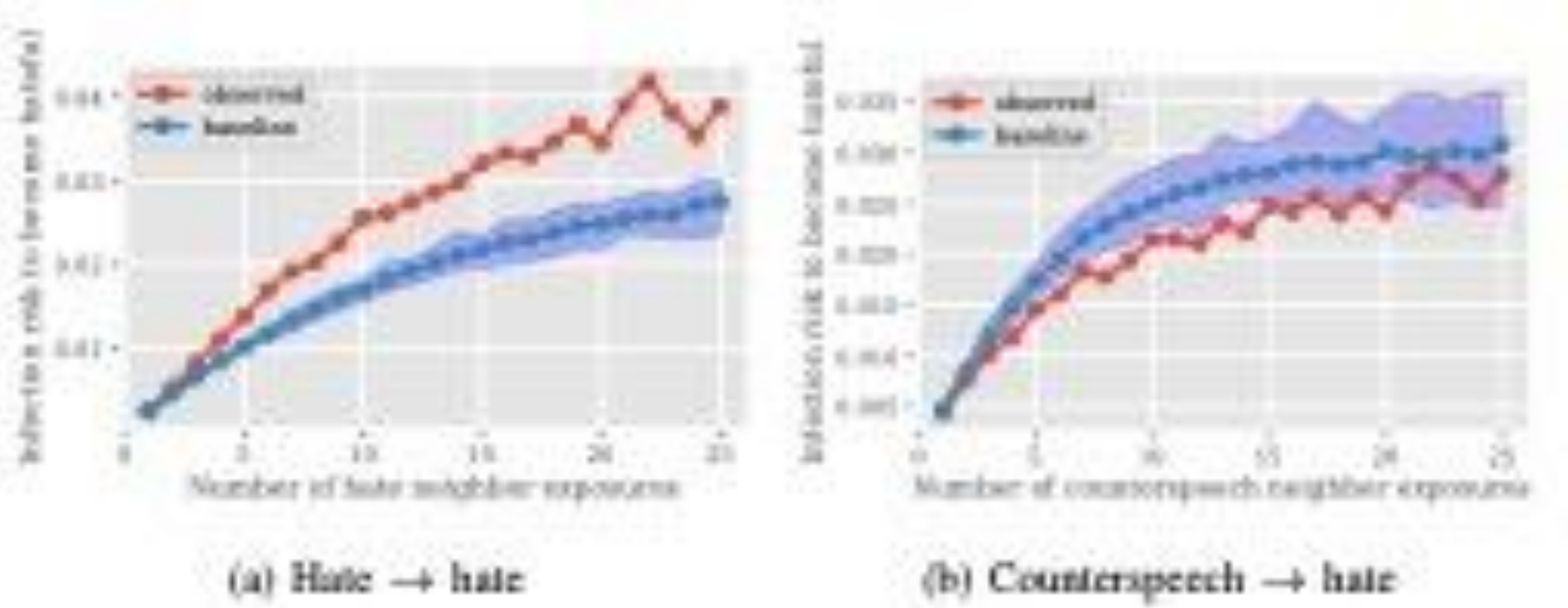


Do hateful and counterspeech users form polarized communities?





Counterspeech's Influence on Hate Spread



(a) Hate → hate

(b) Counterspeech → hate

Class Discussion

The study highlights the potential of counterspeech as a tool for combating online hate, underscoring the importance of supportive and opposing voices in online communities.

Can counterspeech be encouraged? If so, how? Who should be responsible for it?

Class Discussion

Engaging in counterspeech can expose individuals to toxic content and potentially lead to psychological harm. What are the long-term effects on those who regularly engage in counterspeech, and how can they be supported?

Class Discussion

Can AI be effectively used to identify opportunities for counterspeech or even generate counterspeech responses? If so, what are the ethical and practical implications of AI-facilitated counterspeech?

Class Discussion

What connection do you see between the two studies?