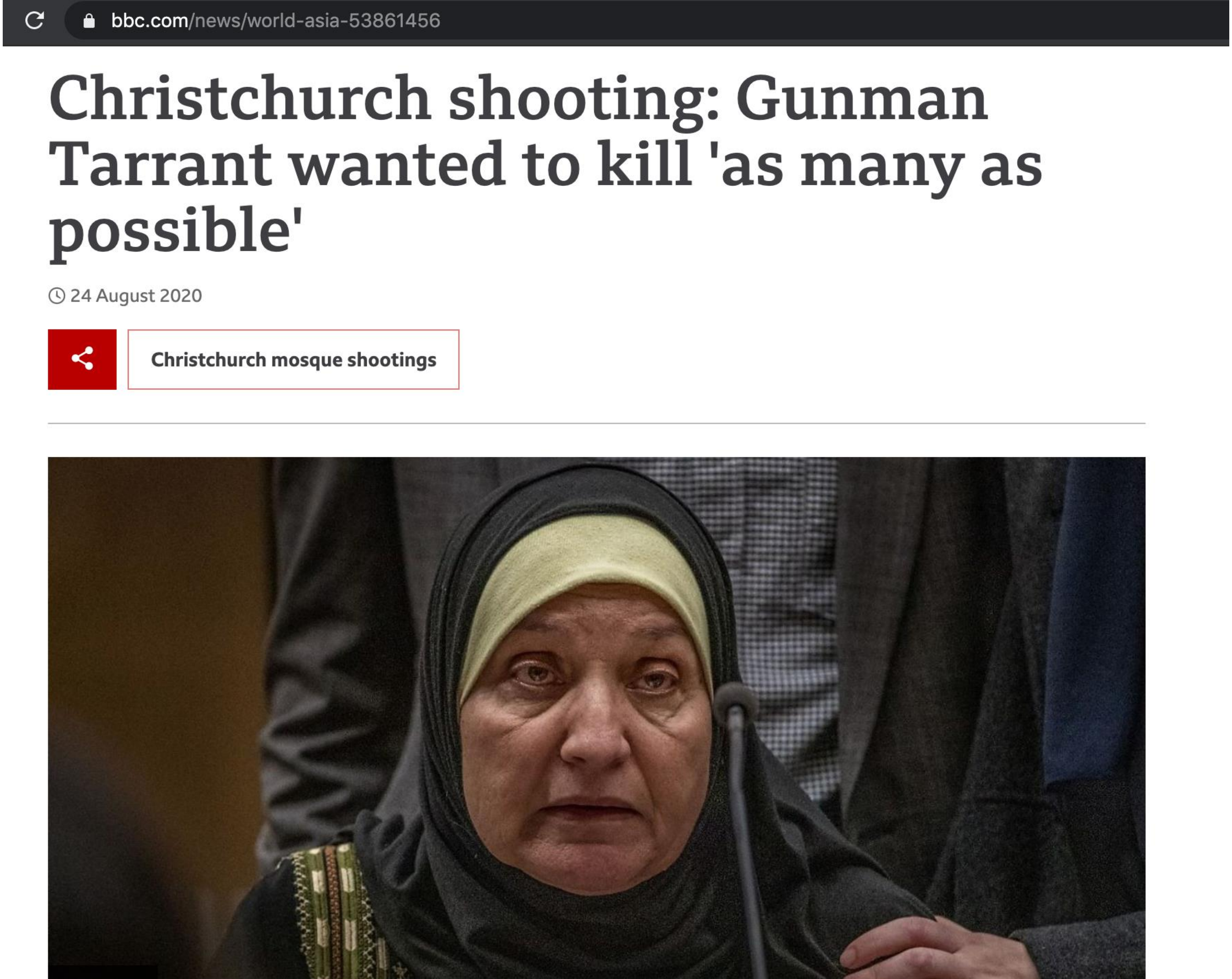
CS 6474/CS 4803 Social Computing: Online Content Moderation

Munmun De Choudhury

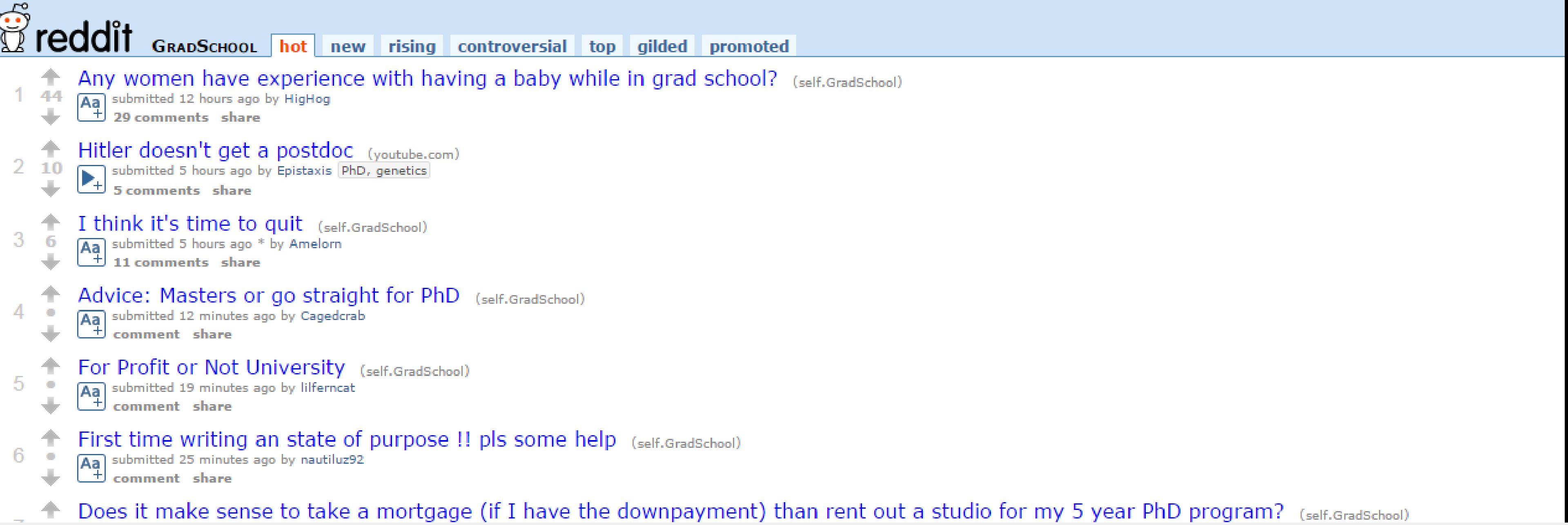
munmund@gatech.edu

Week 13 | March 31, 2025


bbc.com/news/world-asia-53861456
Christchurch mosque shootings

Christchurch shooting: Gunman Tarrant wanted to kill 'as many as possible'

24 August 2020



1 44 Any women have experience with having a baby while in grad school? (self.GradSchool)
submitted 12 hours ago by HigHog
29 comments share

2 10 Hitler doesn't get a postdoc (youtube.com)
submitted 5 hours ago by Epistaxis PhD, genetics
5 comments share

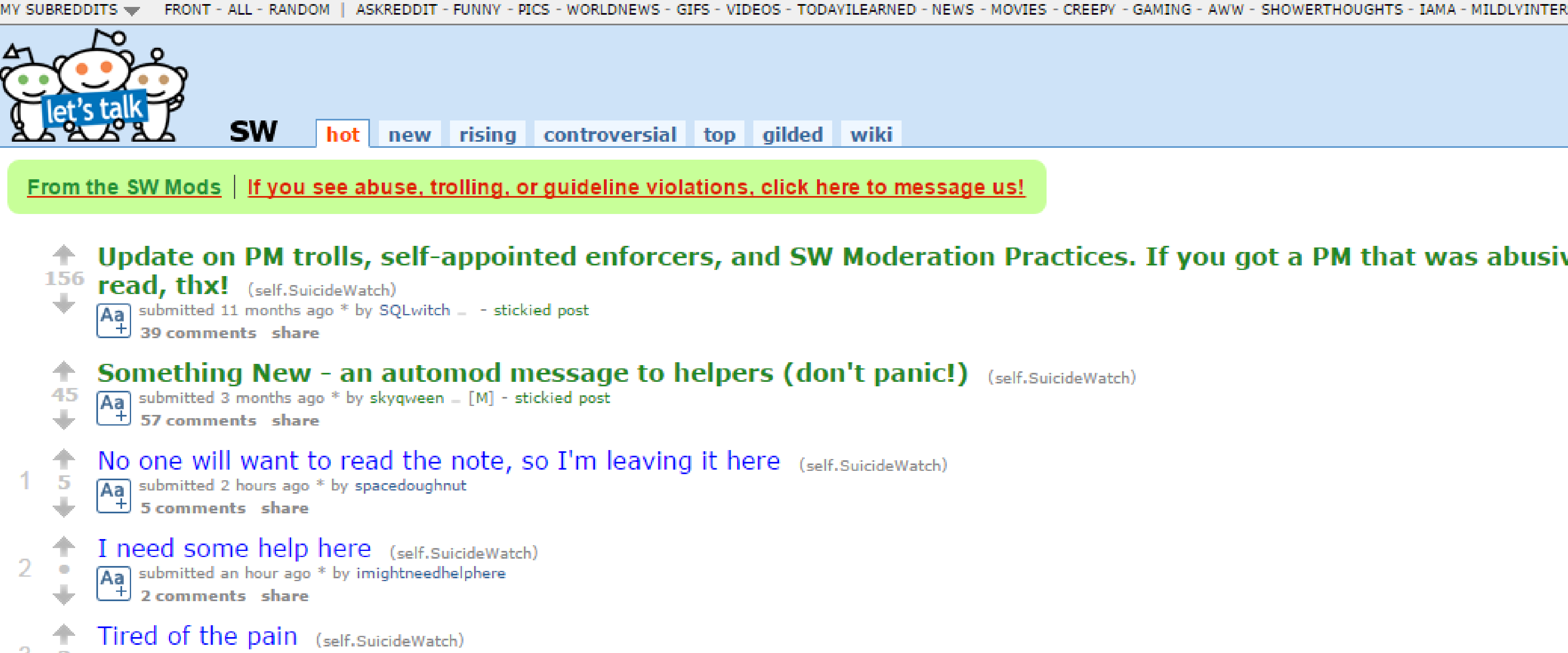
3 6 I think it's time to quit (self.GradSchool)
submitted 5 hours ago * by Amelorn
11 comments share

4 • Advice: Masters or go straight for PhD (self.GradSchool)
submitted 12 minutes ago by Cagedcrab
comment share

5 • For Profit or Not University (self.GradSchool)
submitted 19 minutes ago by lilferncat
comment share

6 • First time writing an state of purpose !! pls some help (self.GradSchool)
submitted 25 minutes ago by nautiluz92
comment share

Does it make sense to take a mortgage (if I have the downpayment) than rent out a studio for my 5 year PhD program? (self.GradSchool)


MY SUBREDDITS
FRONT - ALL - RANDOM | ASKREDDIT - FUNNY - PICS - WORLDNEWS - GIFS - VIDEOS - TODAYILEARNED - NEWS - MOVIES - CREEPY - GAMING - AWW - SHOWERTHOUGHTS - IAMA - MILDLYINTER
let's talk
SW
hot
new
rising
controversial
top
gilded
wiki

From the SW Mods | If you see abuse, trolling, or guideline violations, click here to message us!

156 Update on PM trolls, self-appointed enforcers, and SW Moderation Practices. If you got a PM that was abusiv read, thx! (self.SuicideWatch)
submitted 11 months ago * by SQLwitch - stickied post
39 comments share

45 Something New - an automod message to helpers (don't panic!) (self.SuicideWatch)
submitted 3 months ago * by skyqween [M] - stickied post
57 comments share

1 5 No one will want to read the note, so I'm leaving it here (self.SuicideWatch)
submitted 2 hours ago * by spacedoughnut
5 comments share

2 • I need some help here (self.SuicideWatch)
submitted an hour ago * by imightneedhelphere
2 comments share

Tired of the pain (self.SuicideWatch)

Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

The Effect of Moderation on Online Mental Health Conversations

David Wadden, Tal August, Qisheng Li, and **Tim Althoff**
Paul G. Allen School of Computer Science & Engineering
University of Washington, Seattle, WA
{dwadden, taugust, liqs, althoff}@cs.washington.edu

Abstract

Many people struggling with mental health issues are unable to access adequate care due to high costs and a shortage of mental health professionals, leading to a global mental health crisis. Online mental health communities can help mitigate this crisis by offering a scalable, easily accessible alternative to in-person sessions with therapists or support groups. However, people seeking emotional or psychological support online may be especially vulnerable to the kinds of antisocial behavior that sometimes occur in online discussions. Moderation can improve online discourse quality, but we lack an understanding of its effects on online mental health conversations. In this work, we leveraged a natural experiment, occurring across 200,000 messages from 7,000 online mental health conversations, to evaluate the effects of moderation on online mental health discussions. We found that participation in group mental health discussions led to improvements in psychological perspective, and that these improvements were larger in moderated conversations. The presence of a moderator increased user engagement, encouraged users to discuss negative emotions more candidly, and dramatically reduced bad behavior among chat participants. Moderation also encouraged stronger linguistic coordination, which is indicative of trust building. In addition, moderators who remained active in conversations were especially successful in keeping conversations on topic. Our findings suggest that moderation can serve as a valuable tool to improve the efficacy and safety of online mental health conversations. Based on these findings, we discuss implications and trade-offs involved in designing effective online spaces for mental health support.

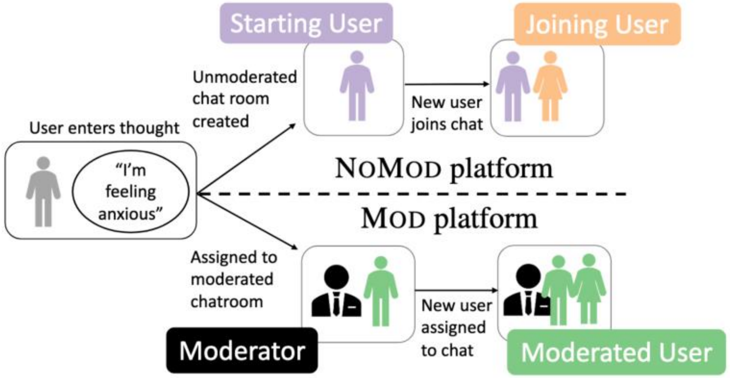
1 Introduction

Over 400 million people globally struggle with mental health challenges, with approximately 300 million experiencing depression (WHO 2018b). Depression leads to economic costs totalling more than $100 billion annually in the United States alone (Twenge et al. 2019). Rates of serious psychological distress – including suicidal ideation and suicide attempts – have increased 71% in adolescents and young adults since 2005 (Twenge et al. 2019). Although psychotherapy and social support can be effective treatments (Wampold and Imel 2015; WHO 2018a), vulnerable individuals often have limited access to therapy and counseling (Bose et al. 2018).

Instead, more and more people are turning to online mental health communities to express emotions, share stigmatized experiences, and receive helpful information (Eysenbach et al. 2004). These communities offer an accessible way for users to connect to a large network of peers experiencing similar challenges. Participants unable to access other treatment options can find social support and relief through these conversations (De Choudhury and De 2014; Sharma and De Choudhury 2018; Naslund et al. 2016). Recently, social support networks have begun to offer a more personalized experience by matching people sharing similar struggles in live, private conversations for support (Althoff, Clark, and Leskovec 2016).

While online mental health communities can provide a valuable setting for giving and receiving support, the quality of support provided by peers is less well-characterized. Can conversation participants temporarily assume the role of a psychological counselor to assist those in serious distress? In addition, the often unrestricted and anonymous environment of online discussions can become a platform for antisocial behavior, such as online abuse or harassment (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015; Zhang et al. 2018). Are these concerns relevant in the setting of an app designed expressly for mental health discussion? Perhaps users of this platform are more thoughtful and considerate than the average forum participant. On the other hand, if bad behavior is an issue, moderation has been shown to be effective tool to combat undesirable behavior in online discussions (Seering et al. 2019; Matias 2019; Lampe et al. 2014; Seo 2007). But little is known about the effectiveness of moderation in the context of mental health applications. Do moderators need to be highly *involved* to keep users safe? Or does simply the knowledge that a moderator is *present* influence behavior without active intervention? Furthermore, what roles do moderators assume in mental health discussions? Are they mostly discipline-keepers, or do they also act as counselors and facilitators?

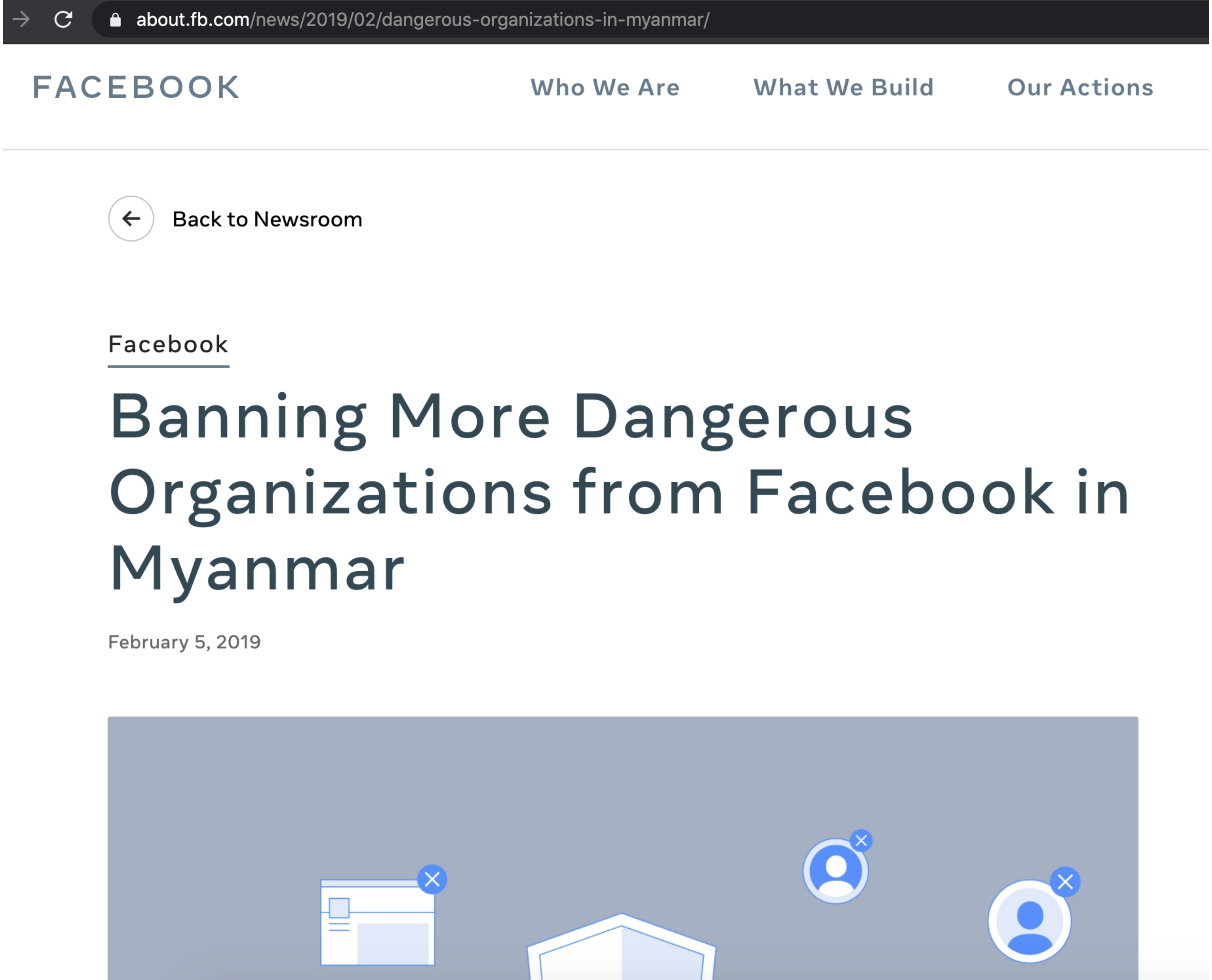
In this work, we investigated how moderation affected online mental health conversations by identifying a natural experiment (DiNardo 2016) occurring when the developers of an online platform hosting unmoderated mental health conversations discontinued the application, replacing it with a



ADRIAN CHEN BUSINESS 10.23.14 6:30 AM

THE LABORERS WHO KEEP DI ** PICS AND BEH ******* OUT OF YOUR FACEBOOK FEED

WIRED



about.fb.com/news/2019/02/dangerous-organizations-in-myanmar/

Facebook

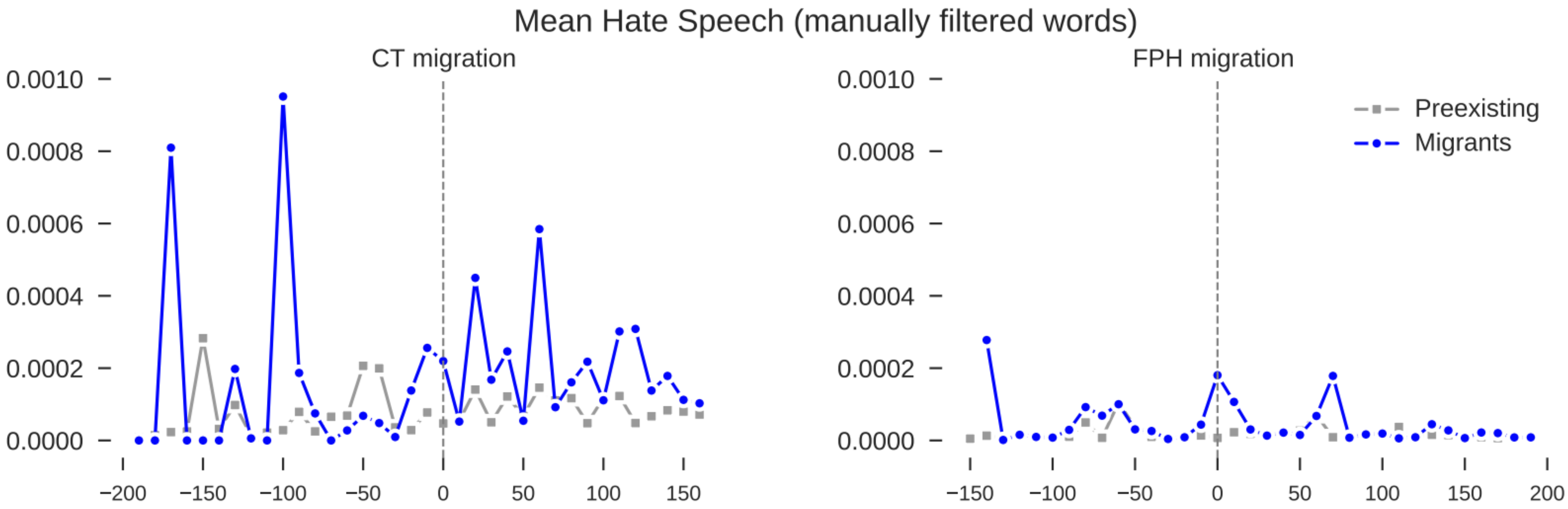
Banning More Dangerous Organizations from Facebook in Myanmar

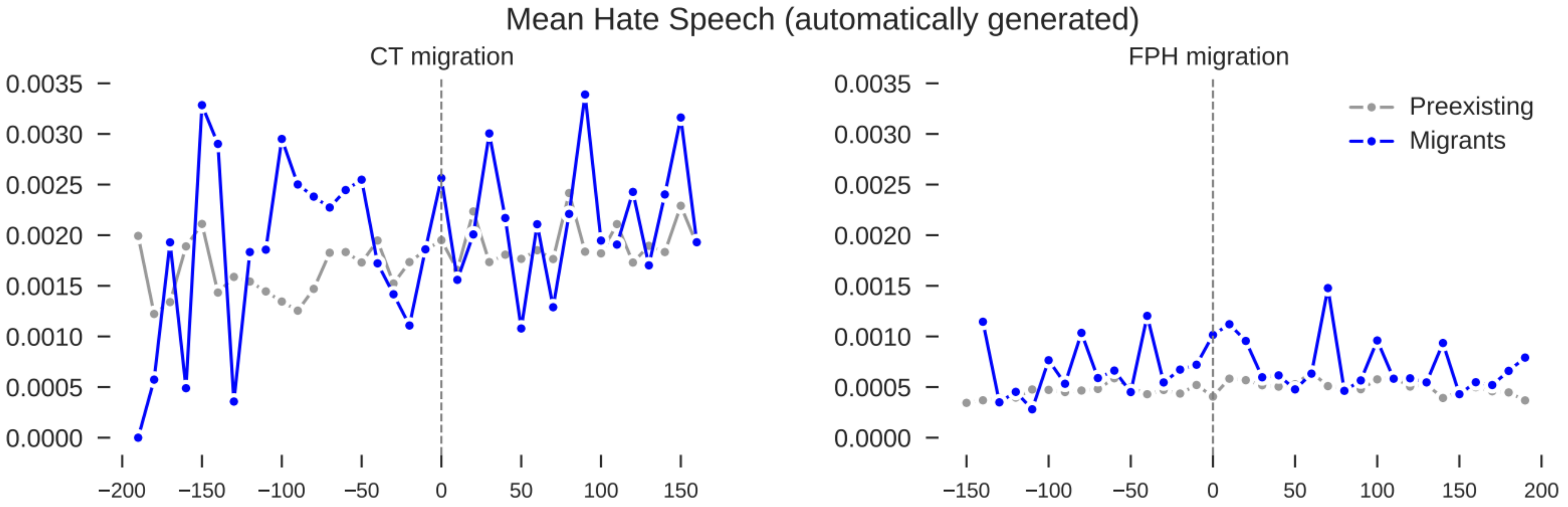
February 5, 2019

Deviance

- Behaviors that violate the norms of a group
  - Akers, 1977; Suler and Phillips 1998.
- Sociological concept
  - Classically comes from Durkeim's *Anomie* book
- Online content moderation and the connection to deviance

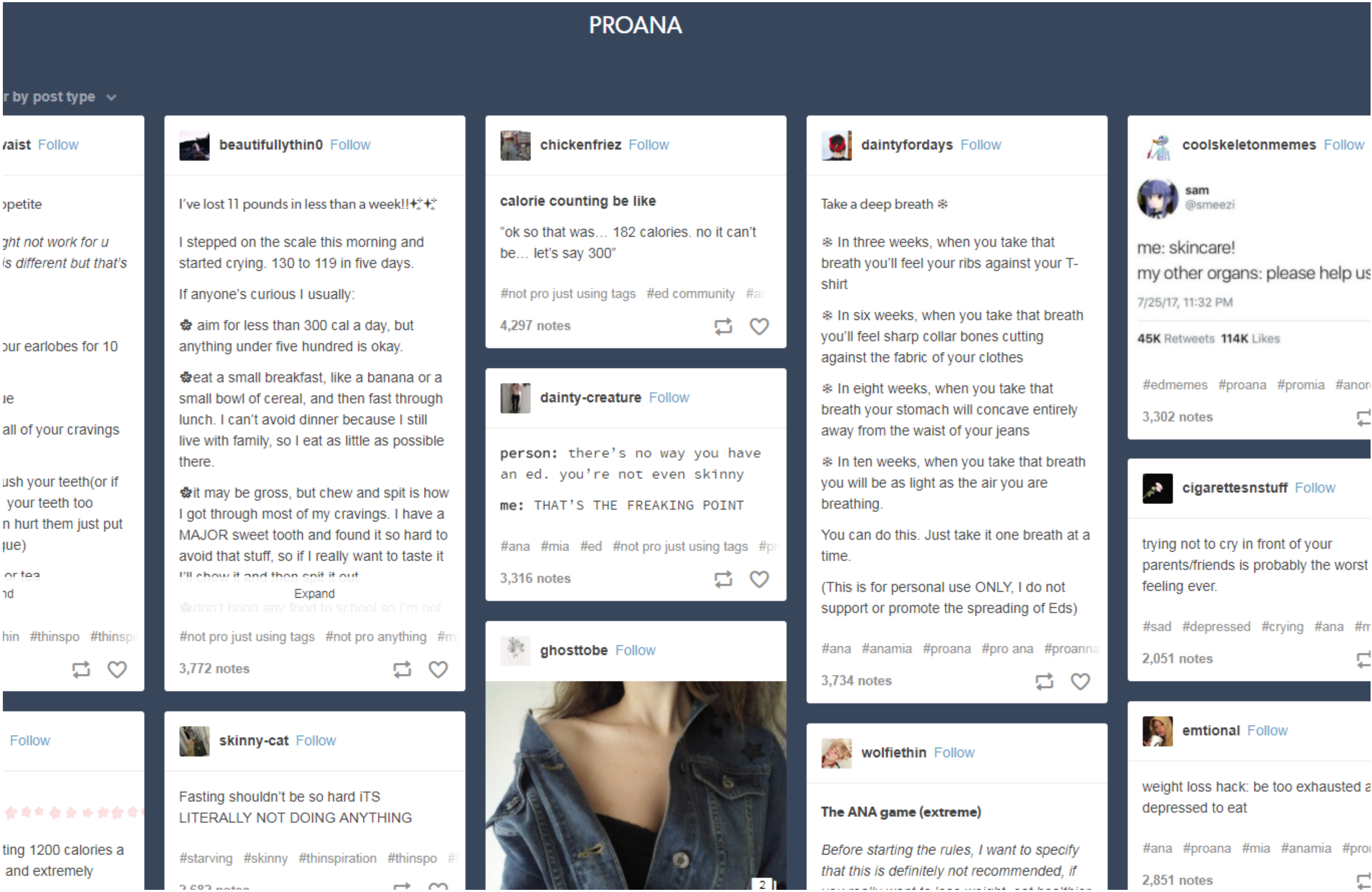
You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech


Mean Hate Speech (manually filtered words)
CT migration
FPH migration
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
−200
−150
−100
−50
0
50
100
150
200
Preexisting
Migrants


Mean Hate Speech (automatically generated)
CT migration
FPH migration
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
−200
−150
−100
−50
0
50
100
150
200
Preexisting
Migrants

Stringent moderation such as deplatforming works. But does it always?

#thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities



PROANA

r by post type

'aist Follow

ɔpetite

ght not work for u
is different but that's

ɔur earlobes for 10

ɪe

all of your cravings

ush your teeth(or if
your teeth too
n hurt them just put
ɟue)

or tea
ɪd

hin #thinspo #thinspi

Follow

ting 1200 calories a
and extremely

beautifullythin0 Follow

I've lost 11 pounds in less than a week!!✨✨

I stepped on the scale this morning and started crying. 130 to 119 in five days.

If anyone's curious I usually:

❁ aim for less than 300 cal a day, but anything under five hundred is okay.

❁eat a small breakfast, like a banana or a small bowl of cereal, and then fast through lunch. I can't avoid dinner because I still live with family, so I eat as little as possible there.

❁it may be gross, but chew and spit is how I got through most of my cravings. I have a MAJOR sweet tooth and found it so hard to avoid that stuff, so if I really want to taste it I'll chew it and then spit it out.

Expand

❁don't bring any food to school so I'm not

#not pro just using tags #not pro anything #m

3,772 notes

skinny-cat Follow

Fasting shouldn't be so hard iTS LITERALLY NOT DOING ANYTHING

#starving #skinny #thinspiration #thinspo #t

chickenfriez Follow

calorie counting be like

"ok so that was… 182 calories. no it can't be… let's say 300"

#not pro just using tags #ed community #a

4,297 notes

dainty-creature Follow

```
person: there's no way you have
an ed. you're not even skinny

me: THAT'S THE FREAKING POINT
```

#ana #mia #ed #not pro just using tags #p

3,316 notes

ghosttobe Follow

2

daintyfordays Follow

Take a deep breath ❄

❄ In three weeks, when you take that breath you'll feel your ribs against your T-shirt

❄ In six weeks, when you take that breath you'll feel sharp collar bones cutting against the fabric of your clothes

❄ In eight weeks, when you take that breath your stomach will concave entirely away from the waist of your jeans

❄ In ten weeks, when you take that breath you will be as light as the air you are breathing.

You can do this. Just take it one breath at a time.

(This is for personal use ONLY, I do not support or promote the spreading of Eds)

#ana #anamia #proana #pro ana #proanna

3,734 notes

wolfiethin Follow

The ANA game (extreme)

Before starting the rules, I want to specify that this is definitely not recommended, if

coolskeletonmemes Follow

sam
@smeezi

me: skincare!
my other organs: please help us

7/25/17, 11:32 PM

45K Retweets 114K Likes

#edmemes #proana #promia #anor

3,302 notes

cigarettesnstuff Follow

trying not to cry in front of your parents/friends is probably the worst feeling ever.

#sad #depressed #crying #ana #m

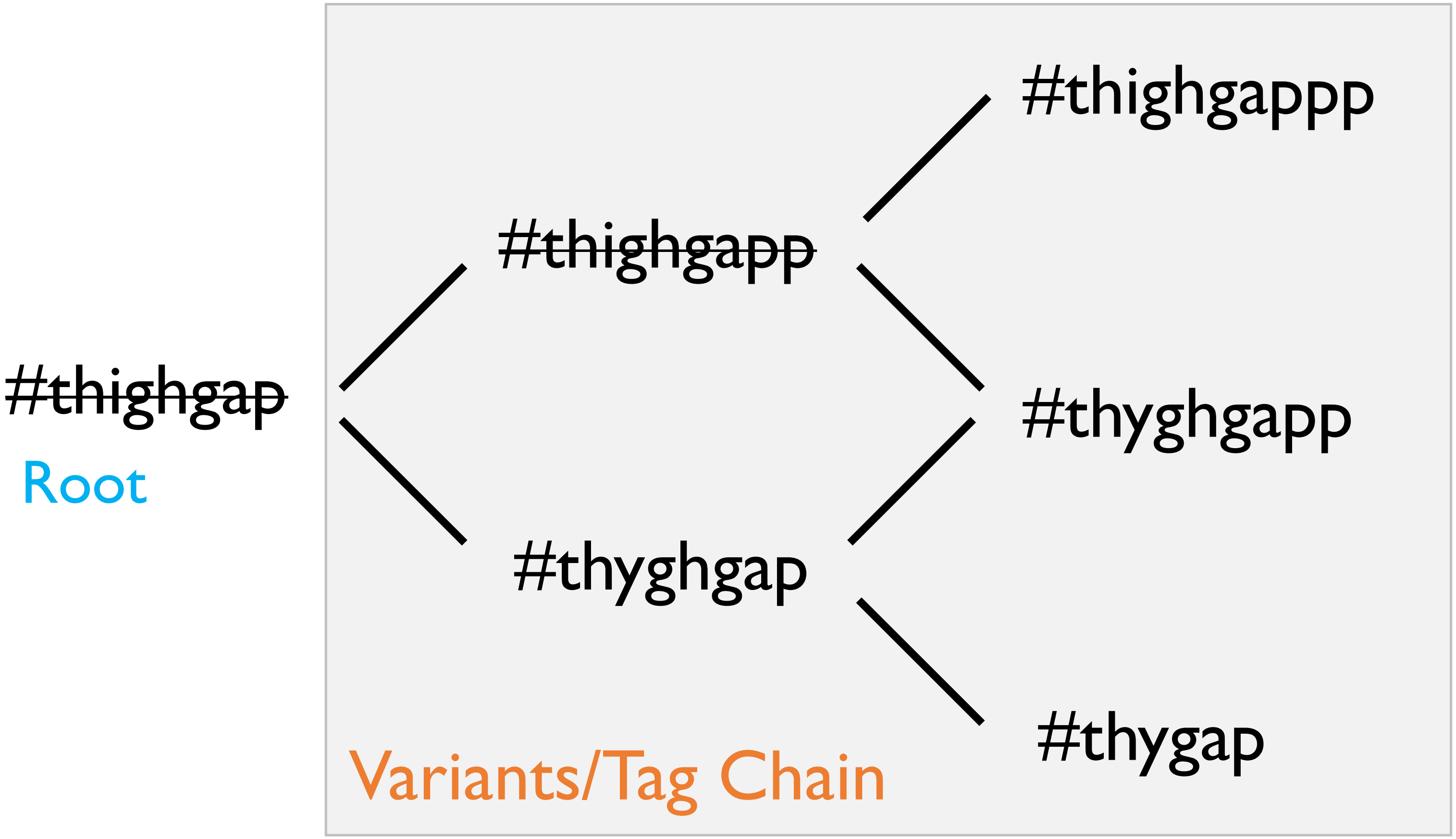
2,051 notes

emtional Follow

weight loss hack: be too exhausted a depressed to eat

#ana #proana #mia #anamia #pro

2,851 notes



~~#thighgap~~

Root

~~#thighgapp~~

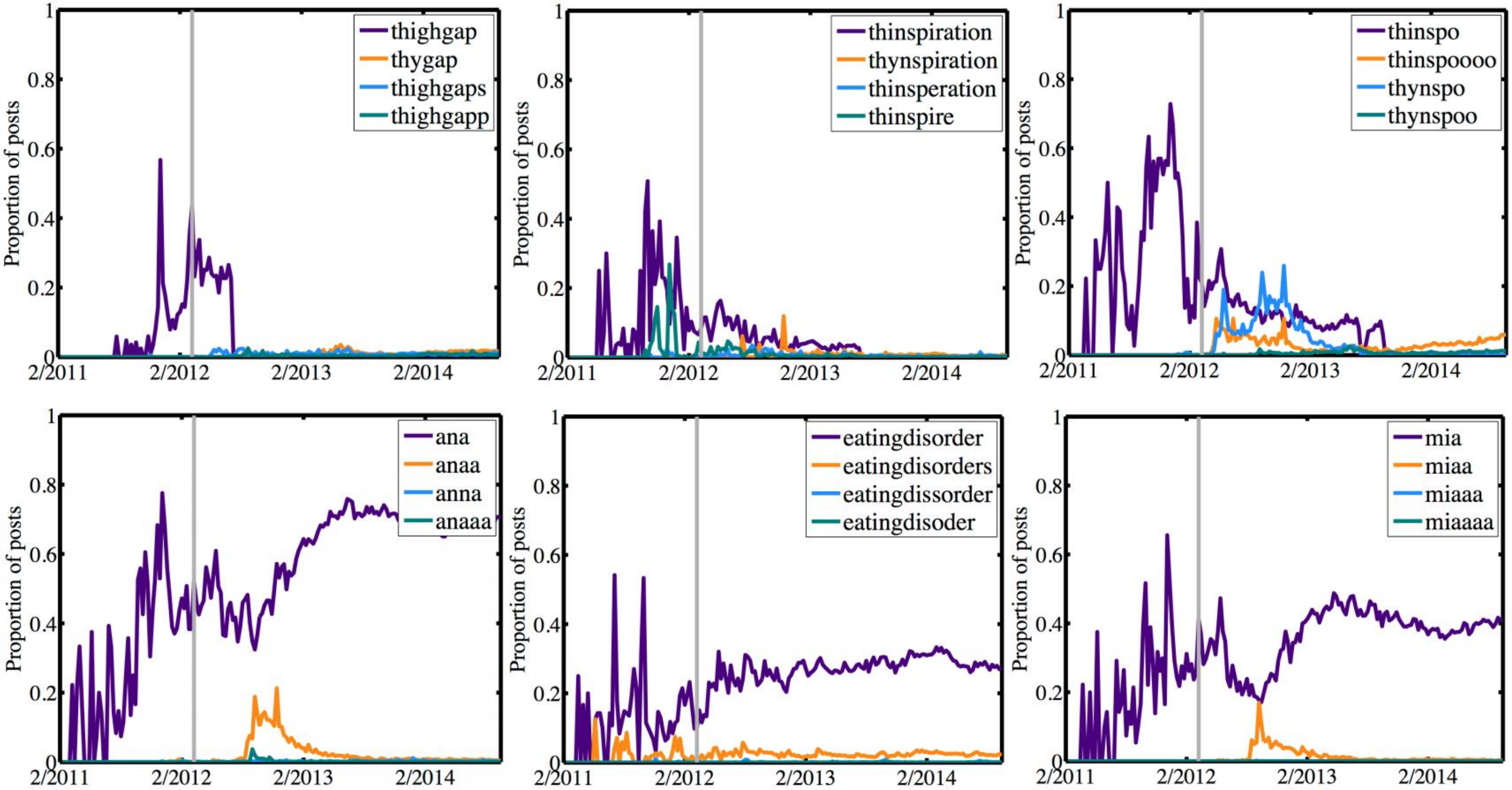
#thyghgap

#thighgappp

#thyghgapp

#thygap

Variants/Tag Chain

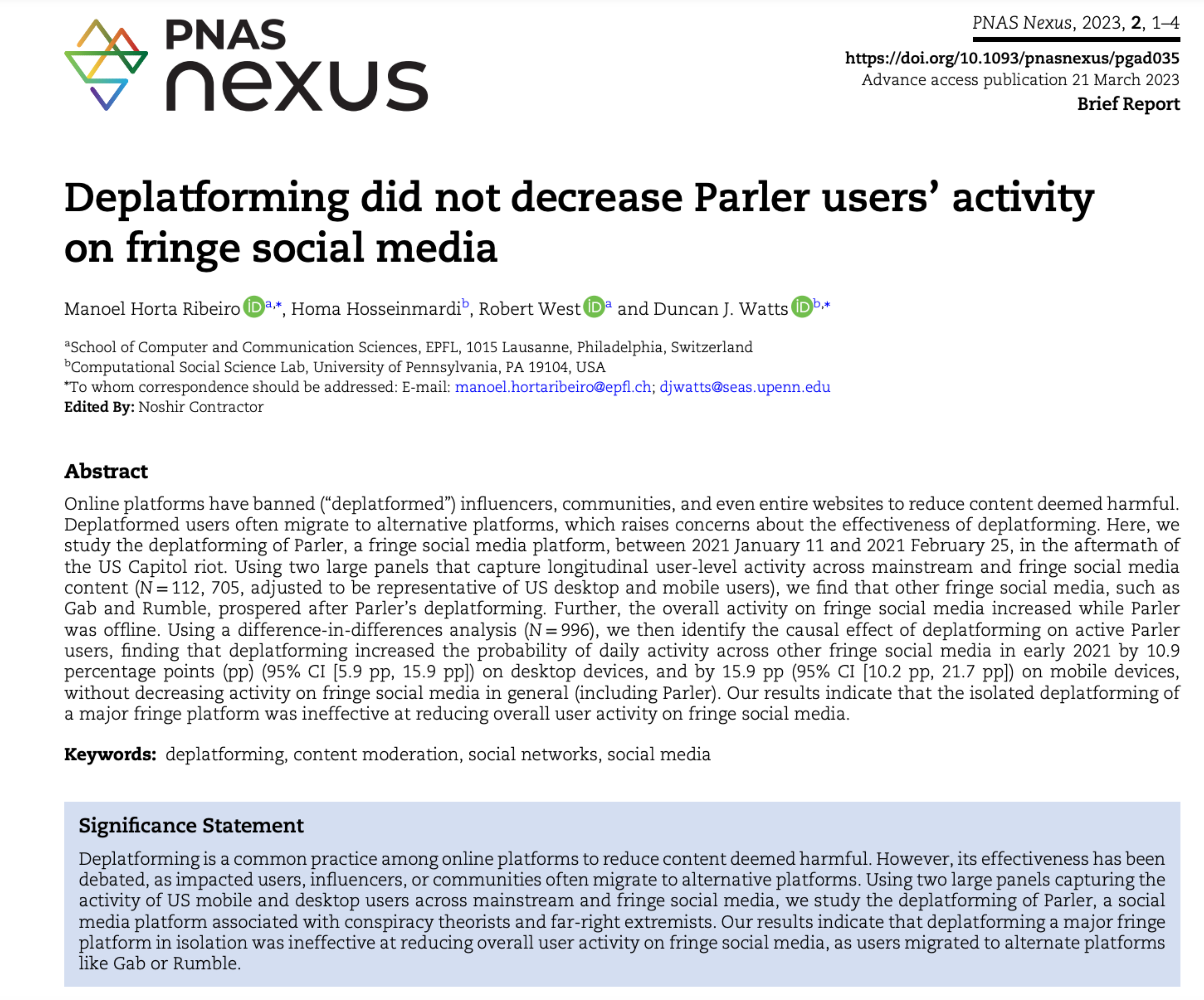


Proportion of posts
0
0.2
0.4
0.6
0.8
1
2/2011
2/2012
2/2013
2/2014
thighgap
thygap
thighgaps
thighgapp
thinspiration
thynspiration
thinsperation
thinspire
thinspo
thinspoooo
thynspo
thynspoo
ana
anaa
anna
anaaa
eatingdisorder
eatingdisorders
eatingdissorder
eatingdisoder
mia
miaa
miaaa
miaaaa

Likes						
Tag Chain	Mean (Root)		Mean (Variants)		z	
eatingdisorder	53	±55.28	44	±72.87	-36.21	***
mia	44	±46.37	56	±46.42	32.79	***
thighgap	36	±39.02	52	±49.00	38.55	***
thinspiration	31	±26.35	58	±57.86	64.12	***
thinspo	33	±34.47	53	±50.58	87.16	***
Change in #likes in variant posts vs. root posts					30.6%	
Comments						
Tag Chain	Mean - Root		Mean – Variant		t Stat.	
eatingdisorder	2	±4.80	2	±4.01	-23.76	***
thighgap	1	±3.05	2	±3.97	27.85	***
thinspiration	1	±3.01	1	±3.62	24.50	***
thinspo	1	±3.22	2	±3.95	38.54	***
Change in #comments in variant posts vs. root posts					15.1%	

Tags co-occurrent w/ roots		Tags co-occurrent w/ variants	
Topic I	Topic II	Topic I	Topic II
alone	bodycheck	suicide	smoke
alwayssad	nofood	selfharrrrm	failure
lifesucks	bones	selfmutilation	depression
pain	flatstomach	cutaddict	depressedquote
unhappy	collarbones	cuts	deadinside
emptyfeeling	skinnyangels	harmingmyself	notgoodenough
anaxiety	thinstagram	scar	addiction
broken	mustbesmaller	razor	wishiweredead
emogirl	fat	bloodsecret123	abandon
sad	tiny	blades	paranoid
sadstagram	assbutt	cutting	callmemistaken
sadsmile	fatty	beautifulpain	useless
anxiety	hipbones	slicemywrists	letmeleave
sorry	beautiful	blood	lost
im_not_okay	pale	die	crying

Other examples where stringent content moderation and banning didn’t help



PNAS Nexus

PNAS Nexus, 2023, **2**, 1–4

https://doi.org/10.1093/pnasnexus/pgad035

Advance access publication 21 March 2023

Brief Report

Deplatforming did not decrease Parler users' activity on fringe social media

Manoel Horta Ribeiro[a,*], Homa Hosseinmardi[b], Robert West[a] and Duncan J. Watts[b,*]

[a]School of Computer and Communication Sciences, EPFL, 1015 Lausanne, Philadelphia, Switzerland
[b]Computational Social Science Lab, University of Pennsylvania, PA 19104, USA
*To whom correspondence should be addressed: E-mail: manoel.hortaribeiro@epfl.ch; djwatts@seas.upenn.edu
Edited By: Noshir Contractor

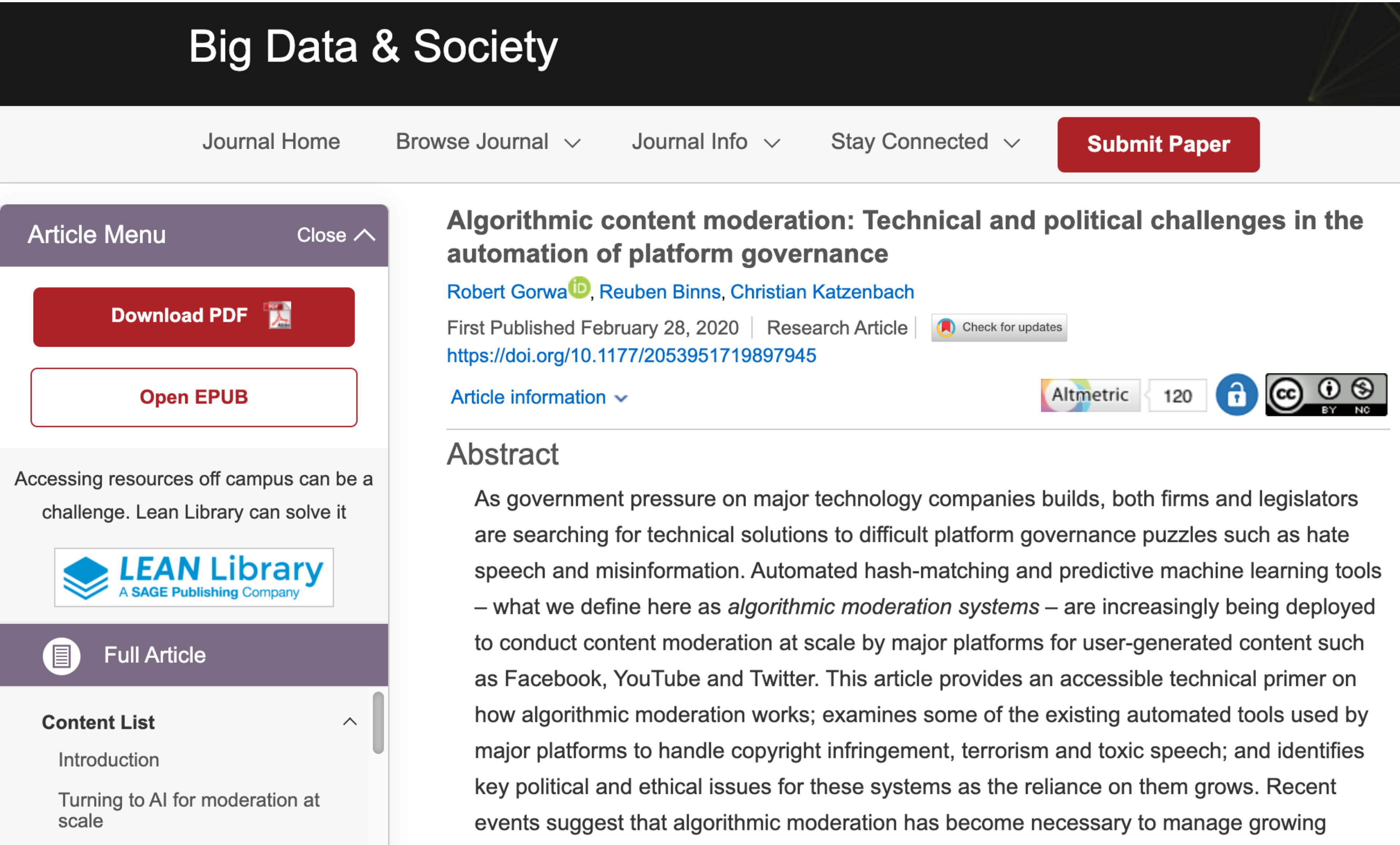
Abstract

Online platforms have banned ("deplatformed") influencers, communities, and even entire websites to reduce content deemed harmful. Deplatformed users often migrate to alternative platforms, which raises concerns about the effectiveness of deplatforming. Here, we study the deplatforming of Parler, a fringe social media platform, between 2021 January 11 and 2021 February 25, in the aftermath of the US Capitol riot. Using two large panels that capture longitudinal user-level activity across mainstream and fringe social media content ($N = 112, 705$, adjusted to be representative of US desktop and mobile users), we find that other fringe social media, such as Gab and Rumble, prospered after Parler's deplatforming. Further, the overall activity on fringe social media increased while Parler was offline. Using a difference-in-differences analysis ($N = 996$), we then identify the causal effect of deplatforming on active Parler users, finding that deplatforming increased the probability of daily activity across other fringe social media in early 2021 by 10.9 percentage points (pp) (95% CI [5.9 pp, 15.9 pp]) on desktop devices, and by 15.9 pp (95% CI [10.2 pp, 21.7 pp]) on mobile devices, without decreasing activity on fringe social media in general (including Parler). Our results indicate that the isolated deplatforming of a major fringe platform was ineffective at reducing overall user activity on fringe social media.

Keywords: deplatforming, content moderation, social networks, social media

Significance Statement

Deplatforming is a common practice among online platforms to reduce content deemed harmful. However, its effectiveness has been debated, as impacted users, influencers, or communities often migrate to alternative platforms. Using two large panels capturing the activity of US mobile and desktop users across mainstream and fringe social media, we study the deplatforming of Parler, a social media platform associated with conspiracy theorists and far-right extremists. Our results indicate that deplatforming a major fringe platform in isolation was ineffective at reducing overall user activity on fringe social media, as users migrated to alternate platforms like Gab or Rumble.



Accessing resources off campus can be a challenge. Lean Library can solve it

Algorithmic content moderation: Technical and political challenges in the automation of platform governance

Robert Gorwa, Reuben Binns, Christian Katzenbach

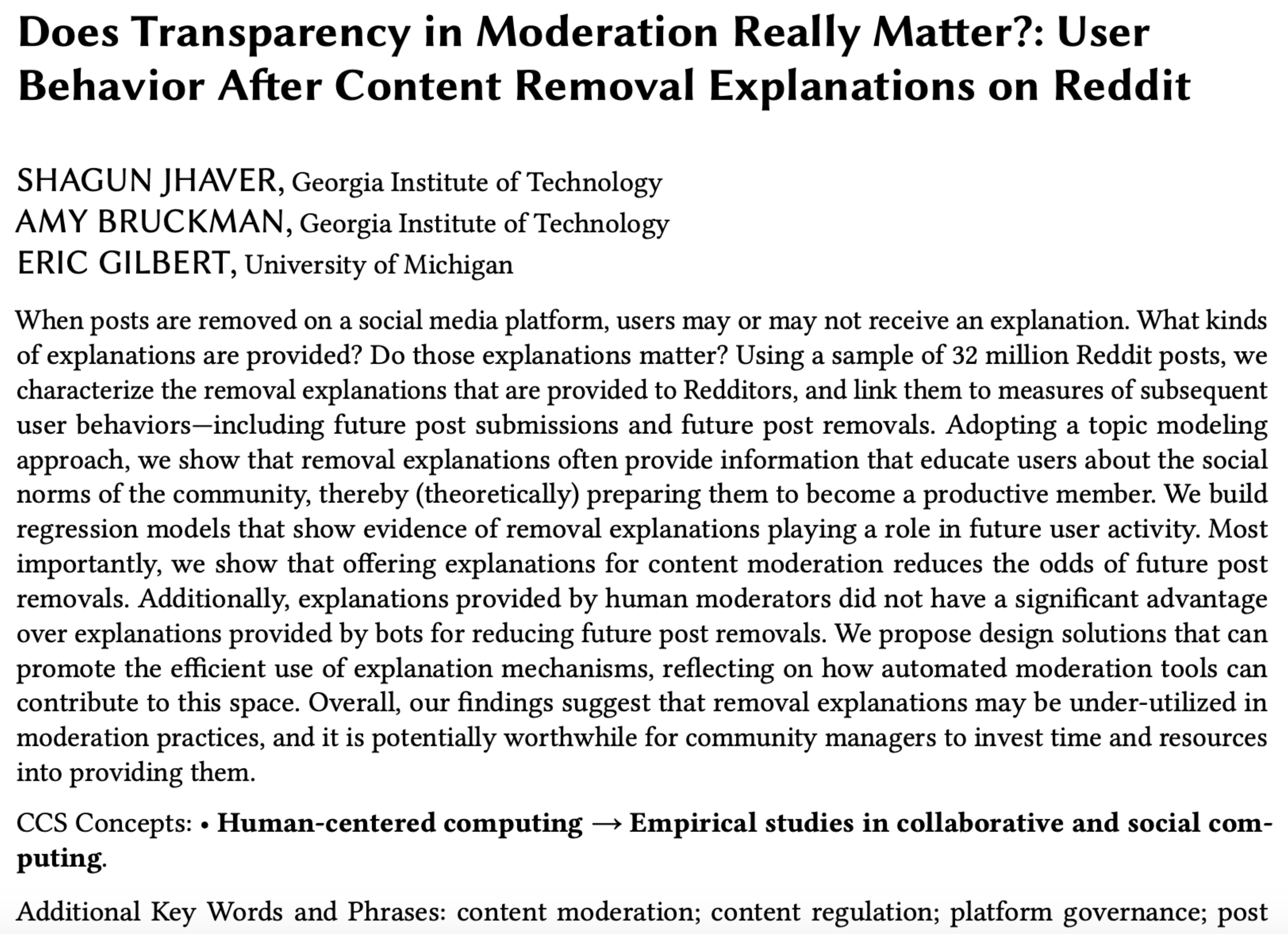
First Published February 28, 2020 | Research Article

https://doi.org/10.1177/2053951719897945

Abstract

As government pressure on major technology companies builds, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation. Automated hash-matching and predictive machine learning tools – what we define here as *algorithmic moderation systems* – are increasingly being deployed to conduct content moderation at scale by major platforms for user-generated content such as Facebook, YouTube and Twitter. This article provides an accessible technical primer on how algorithmic moderation works; examines some of the existing automated tools used by major platforms to handle copyright infringement, terrorism and toxic speech; and identifies key political and ethical issues for these systems as the reliance on them grows. Recent events suggest that algorithmic moderation has become necessary to manage growing

Challenges of reliance on AI moderation tools



Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit

SHAGUN JHAVER, Georgia Institute of Technology
AMY BRUCKMAN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

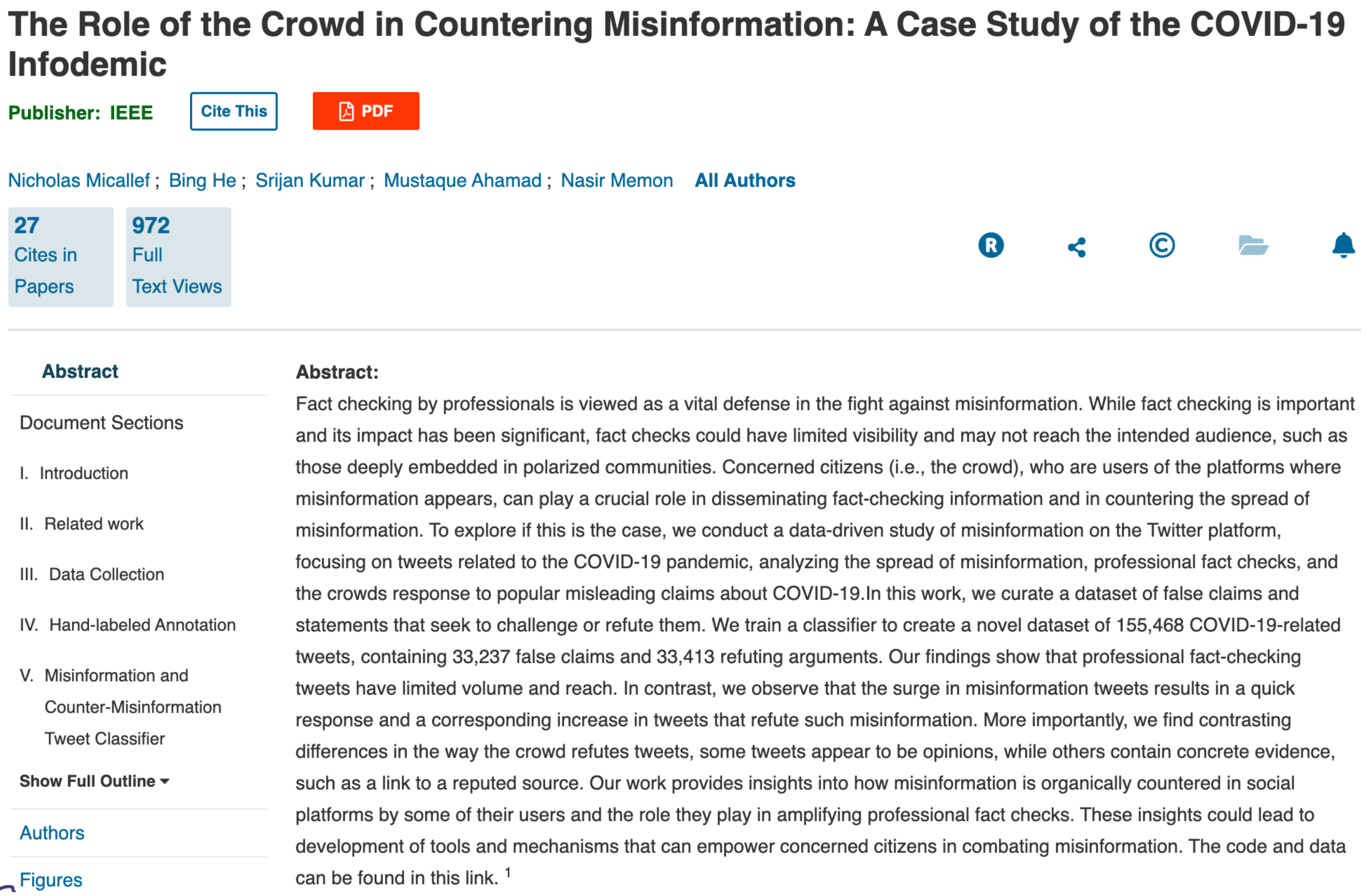
When posts are removed on a social media platform, users may or may not receive an explanation. What kinds of explanations are provided? Do those explanations matter? Using a sample of 32 million Reddit posts, we characterize the removal explanations that are provided to Redditors, and link them to measures of subsequent user behaviors—including future post submissions and future post removals. Adopting a topic modeling approach, we show that removal explanations often provide information that educate users about the social norms of the community, thereby (theoretically) preparing them to become a productive member. We build regression models that show evidence of removal explanations playing a role in future user activity. Most importantly, we show that offering explanations for content moderation reduces the odds of future post removals. Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots for reducing future post removals. We propose design solutions that can promote the efficient use of explanation mechanisms, reflecting on how automated moderation tools can contribute to this space. Overall, our findings suggest that removal explanations may be under-utilized in moderation practices, and it is potentially worthwhile for community managers to invest time and resources into providing them.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation; content regulation; platform governance; post

Class Exercise 1 – what could be alternative strategies for content moderation that do not involve outright banning or deplatforming?

Class Exercise 2 –What could be the potential consequences of over-moderation and under-moderation on user engagement and platform credibility?



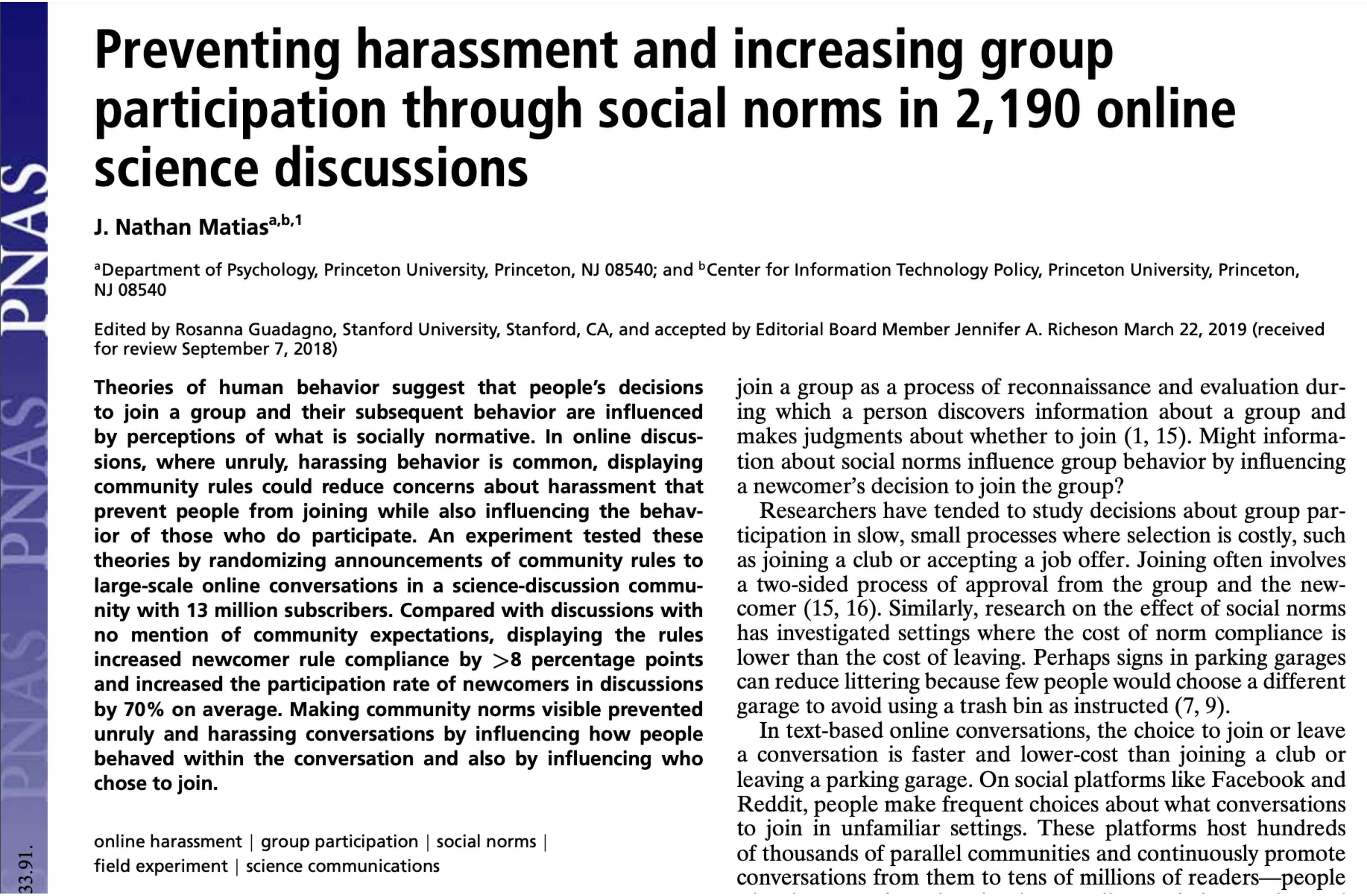
The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic

Publisher: IEEE

Nicholas Micallef ; Bing He ; Srijan Kumar ; Mustaque Ahamad ; Nasir Memon All Authors

Abstract:

Fact checking by professionals is viewed as a vital defense in the fight against misinformation. While fact checking is important and its impact has been significant, fact checks could have limited visibility and may not reach the intended audience, such as those deeply embedded in polarized communities. Concerned citizens (i.e., the crowd), who are users of the platforms where misinformation appears, can play a crucial role in disseminating fact-checking information and in countering the spread of misinformation. To explore if this is the case, we conduct a data-driven study of misinformation on the Twitter platform, focusing on tweets related to the COVID-19 pandemic, analyzing the spread of misinformation, professional fact checks, and the crowds response to popular misleading claims about COVID-19.In this work, we curate a dataset of false claims and statements that seek to challenge or refute them. We train a classifier to create a novel dataset of 155,468 COVID-19-related tweets, containing 33,237 false claims and 33,413 refuting arguments. Our findings show that professional fact-checking tweets have limited volume and reach. In contrast, we observe that the surge in misinformation tweets results in a quick response and a corresponding increase in tweets that refute such misinformation. More importantly, we find contrasting differences in the way the crowd refutes tweets, some tweets appear to be opinions, while others contain concrete evidence, such as a link to a reputed source. Our work provides insights into how misinformation is organically countered in social platforms by some of their users and the role they play in amplifying professional fact checks. These insights could lead to development of tools and mechanisms that can empower concerned citizens in combating misinformation. The code and data can be found in this link. [1]



Preventing harassment and increasing group participation through social norms in 2,190 online science discussions

J. Nathan Matias[a,b,1]

[a]Department of Psychology, Princeton University, Princeton, NJ 08540; and [b]Center for Information Technology Policy, Princeton University, Princeton, NJ 08540

Edited by Rosanna Guadagno, Stanford University, Stanford, CA, and accepted by Editorial Board Member Jennifer A. Richeson March 22, 2019 (received for review September 7, 2018)

Theories of human behavior suggest that people's decisions to join a group and their subsequent behavior are influenced by perceptions of what is socially normative. In online discussions, where unruly, harassing behavior is common, displaying community rules could reduce concerns about harassment that prevent people from joining while also influencing the behavior of those who do participate. An experiment tested these theories by randomizing announcements of community rules to large-scale online conversations in a science-discussion community with 13 million subscribers. Compared with discussions with no mention of community expectations, displaying the rules increased newcomer rule compliance by >8 percentage points and increased the participation rate of newcomers in discussions by 70% on average. Making community norms visible prevented unruly and harassing conversations by influencing how people behaved within the conversation and also by influencing who chose to join.

online harassment | group participation | social norms | field experiment | science communications

join a group as a process of reconnaissance and evaluation during which a person discovers information about a group and makes judgments about whether to join (1, 15). Might information about social norms influence group behavior by influencing a newcomer's decision to join the group?

Researchers have tended to study decisions about group participation in slow, small processes where selection is costly, such as joining a club or accepting a job offer. Joining often involves a two-sided process of approval from the group and the newcomer (15, 16). Similarly, research on the effect of social norms has investigated settings where the cost of norm compliance is lower than the cost of leaving. Perhaps signs in parking garages can reduce littering because few people would choose a different garage to avoid using a trash bin as instructed (7, 9).

In text-based online conversations, the choice to join or leave a conversation is faster and lower-cost than joining a club or leaving a parking garage. On social platforms like Facebook and Reddit, people make frequent choices about what conversations to join in unfamiliar settings. These platforms host hundreds of thousands of parallel communities and continuously promote conversations from them to tens of millions of readers—people

Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support

CHARLOTTE SCHLUGER, Cornell University, USA
JONATHAN P. CHANG, Cornell University, USA
CRISTIAN DANESCU-NICULESCU-MIZIL, Cornell University, USA
KAREN LEVY, Cornell University, USA

To address the widespread problem of uncivil behavior, many online discussion platforms employ human moderators to take action against objectionable content, such as removing it or placing sanctions on its authors. This *reactive* paradigm of taking action against already-posted antisocial content is currently the most common form of moderation, and has accordingly underpinned many recent efforts at introducing automation into the moderation process. Comparatively less work has been done to understand other moderation paradigms—such as *proactively* discouraging the emergence of antisocial behavior rather than reacting to it—and the role algorithmic support can play in these paradigms. In this work, we investigate such a proactive framework for moderation in a case study of a collaborative setting: Wikipedia Talk Pages. We employ a mixed methods approach, combining qualitative and design components for a holistic analysis. Through interviews with moderators, we find that despite a lack of technical and social support, moderators already engage in a number of proactive moderation behaviors, such as preemptively intervening in conversations to keep them on track. Further, we explore how automation could assist with this existing proactive moderation workflow by building a prototype tool, presenting it to moderators, and examining how the assistance it provides might fit into their workflow. The resulting feedback uncovers both strengths and drawbacks of the prototype tool and suggests concrete steps towards further developing such assisting technology so it can most effectively support moderators in their existing proactive moderation workflow.

Decentralized platform governance

Platform Governance outside of the US

A Multi-Stakeholder Approach to Content Moderation