

Computational Methods to Understand and Mitigate Online Hate



Gianluca Stringhini
Boston University

gian@bu.edu

[@gianluca_string](https://twitter.com/gianluca_string)

WARNING: Some of this content might be upsetting

The good'ole times...



Something went wrong...

How a racist, sexist hate
mob forced Leslie Jones
off

INTERNET 8 DECEMBER 2016

By Kristen V.

Pizzagate: How a 4Chan conspiracy went mainstream

The power of “meme magic” is changing the world as we know it.

Over the past 24 hours, Leslie Jones has been inundated with racist, hateful vitriol on Twitter.



Donald J. Trump
@realDonaldTrump



Follow

ews
ump"

Emerging problems



Anonymous (ID: [AMaV12jQ](#)) 🇺🇸 04/28/17(Fri)17:16:08 No.123210970 ▶ [>>123211305](#) >>[123211335](#) >>[123211559](#)

the email belongs to Gianluca Stringhini.

this is his Twitter: https://twitter.com/gianluca_string?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

one of his tweets: https://twitter.com/gianluca_string/status/786205595745550336

one of his projects: <https://arxiv.org/abs/1610.03452?platform=hootsuite>

youtube video: <https://www.youtube.com/watch?v=tS9Lkpj9DF4> [Embed]

if you dont think this is real, you are retarded.



As researchers, where do we even start?



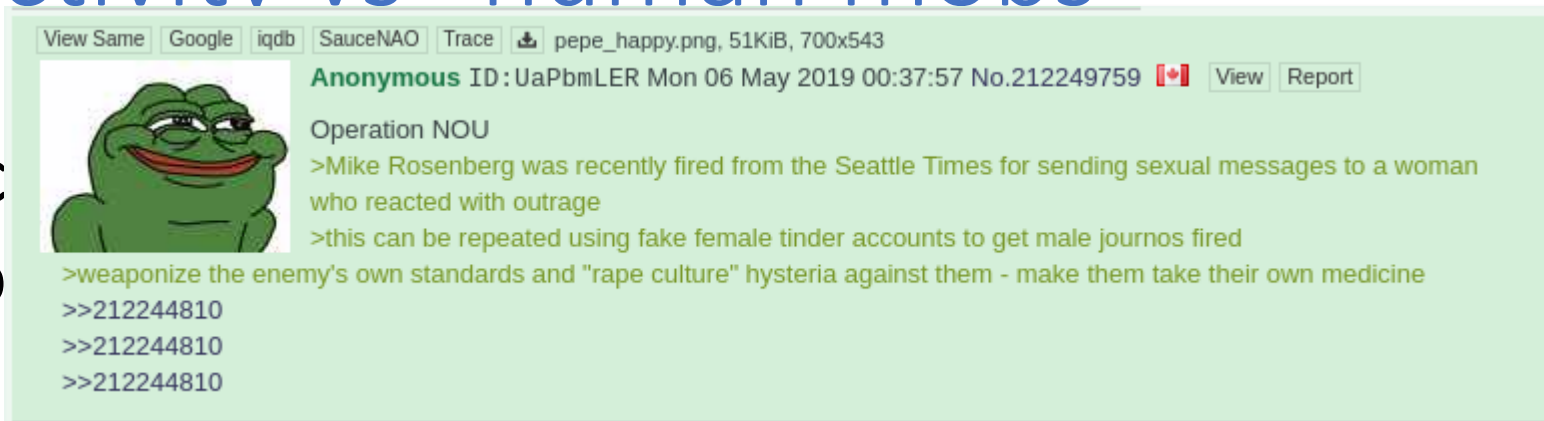
-  **gabs** @dinga_omandinga 2m
i wont pay for mcdonalds but ill eat it for free - so i did a mcdonalds survey here bit.ly/P7OTMs and...
tumblr.co/ZElCYQNwhsc
Expand
-  **Lex** @liddoelex 2m
i wont pay for mcdonalds but ill eat it for free - so i did a mcdonalds survey here bit.ly/P7OTfm and...
tumblr.co/ZTRxMyQNwgJ9
Expand
-  **Jasmine webber-james** @MingeXD 3m
i wont pay for mcdonalds but ill eat it for free - so i did a mcdonalds survey here bit.ly/P7OVnK and... tumblr.co/Z-LtoxQNwVOx
Expand
-  **XO**♥ @RosieCortez 3m
i wont pay for mcdonalds but ill eat it for free - so i did a mcdonalds survey here bit.ly/P7OTw8 and...
tumblr.co/ZFa1KxQNwRZF
Expand



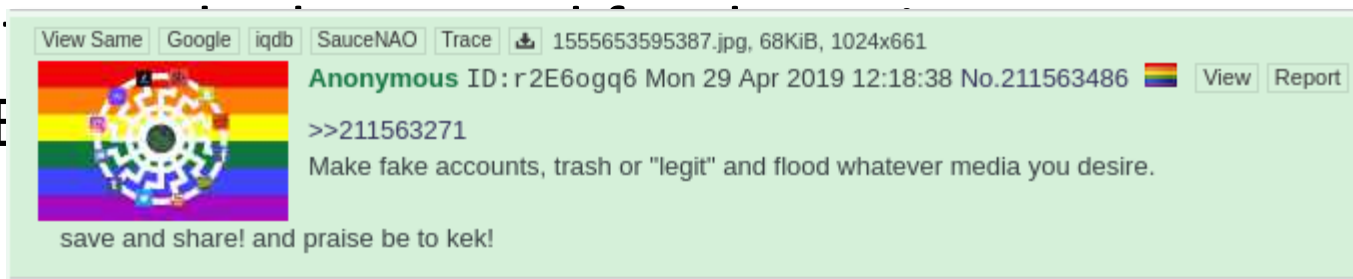
-  week ago
This gives me a lot of hope. This is what the world needs. People, a lot of young people, are poisoning their minds and souls on the internet these days. I was a dude posting on 4Chan back in the day before 'meme' was even a known term. Chaos, racism, sexism, homophobia-white supremacy, all of that nonsense so casually extolled. Then I grew up, and all of that time is just... well, it's a shame. To spend so much time on hatred. It's a waste of life.
Show less
Reply · 6 🍌 🍌 🍌
Hide replies ^
-  1 day ago
Kill yourself then, dont waste anymore life
Reply · 🍌 🍌 🍌
-  4 days ago
Italian immigrants... so the guy isn't even white
Reply · 2 🍌 🍌 🍌
Hide replies ^
-  3 days ago
(((ted)))
Reply · 🍌 🍌 🍌
-  3 days ago
Italians are white
Reply · 🍌 🍌 🍌
-  3 days ago
absolutely not. Way too much random Mediterranean stuff
Reply · 🍌 🍌 🍌
-  1 year ago (edited)
If you all have such a big problem with  opinion, why are you subscribed, Or watching his content?
Reply · 43 🍌 🍌 🍌
-  5 months ago
So the  knows his place.
Reply · 1 🍌 🍌 🍌

Bot activity vs “human mobs”

- Large scale
- Synchronous

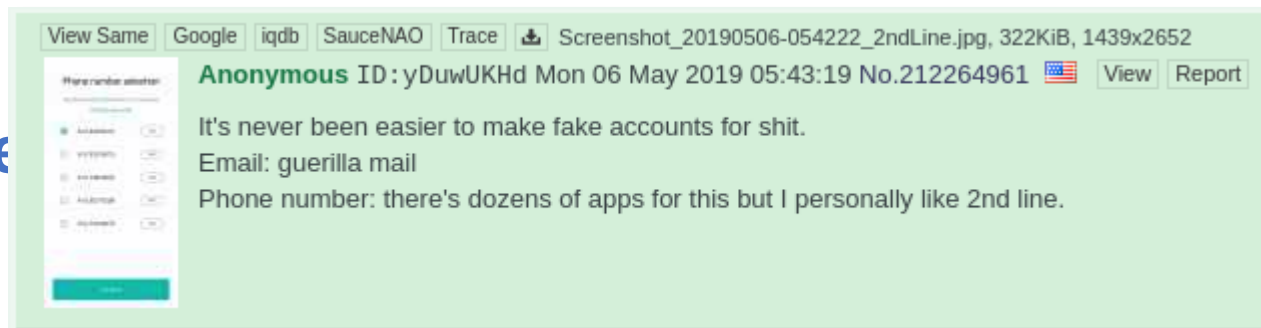


These elements
SynchroTrap, E



Human activity is less coordinated -> **characteristic traits stand out less**

We have a **loose**



, etc.

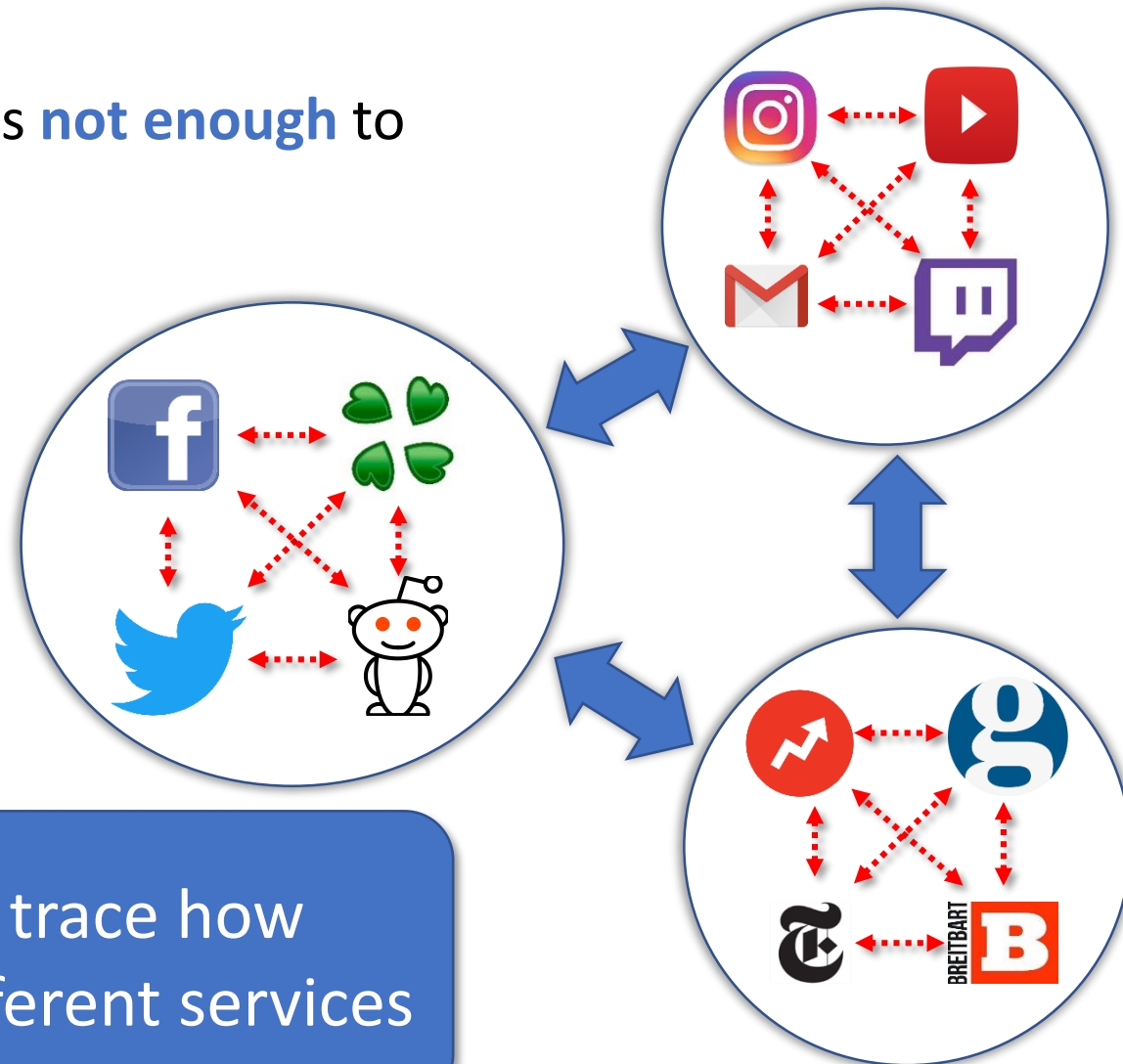
Information is not only textual



Online services do not exist in a vacuum

Looking at a single service at a time is **not enough** to capture online hate dynamics

There is anecdotal evidence that “fringe” Web communities have a strong influence with regards to coordinating hate attacks against users on other platforms



We lack tools to effectively trace how information spreads across different services

Collecting and labeling hate data

Hate data is hard to label



Dataset (ICWSM 2018)

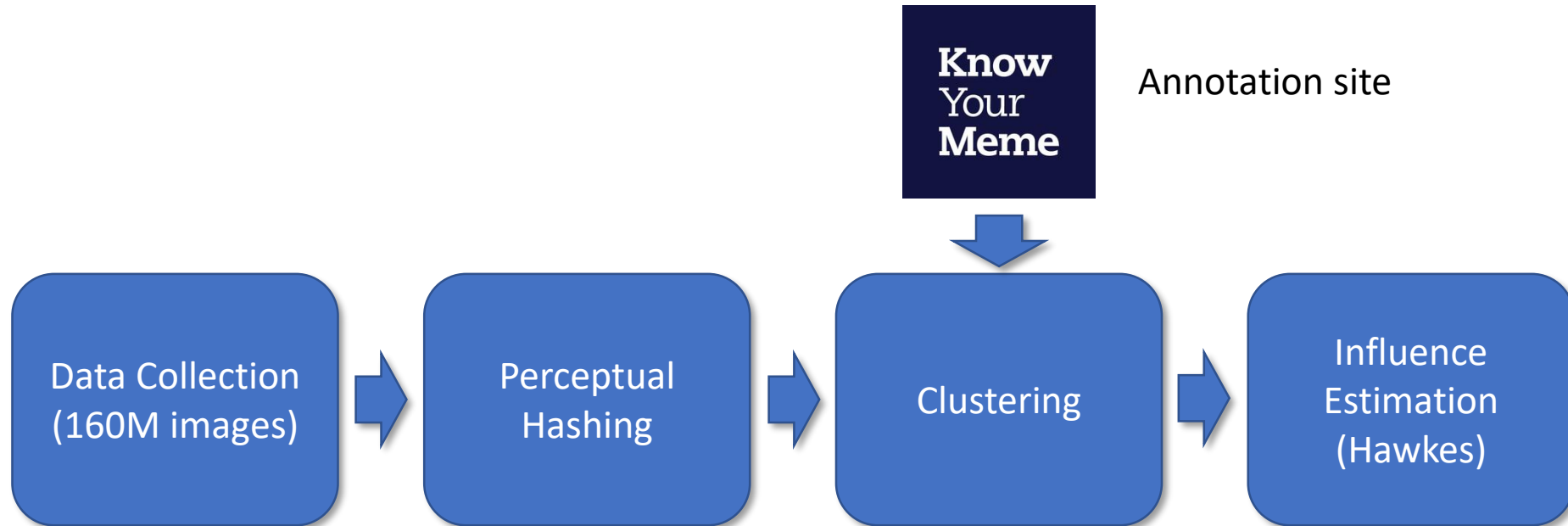
We leveraged crowd workers to label 80k tweets

Issues:

- Expensive: how do we find feeds with **high toxicity**?
- Hard to agree: what exactly is hateful content?
 - Having many categories doesn't help (bullying, aggression, hate, ...)
 - Hate is often very content dependent

Measuring “meme magic”

Measuring memes at scale



Code for our analysis pipeline (and data) available at https://github.com/memespaper/memes_pipeline



Examples of clusters



Top memes per social network



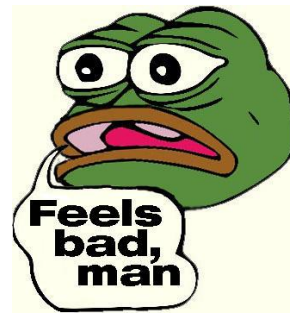
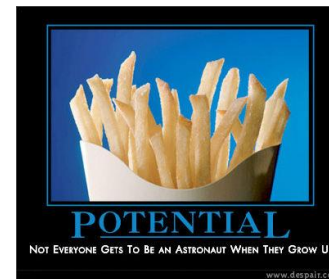
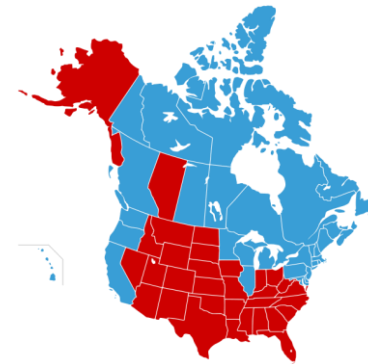
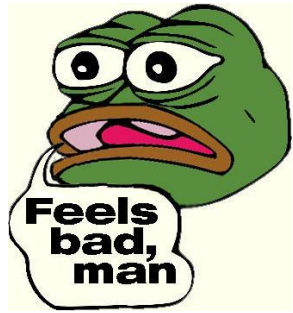
Paper (IMC 2018)

4chan (/pol/)

Reddit

Gab


Twitter




Measuring online raids

Real life story: weird emails in my inbox

Alt-Right

 Emilee Reid <emileereid4@gmail.com> 👍 ↻ Reply all | ▾

You are a retarded f████t

 Joel Dunn <joelcdunn@gmail.com> 👍 ↻ Reply all | ▾
Fri 09/06, 23:20
Stringhini, Gianluca ▾

Inbox

Flag for follow up. Completed on 10 June 2017.

A) you're in IT but are trying to be a sociology major.

B) you're clearly trying to overblow and be uber-dramatic about what 4chan is.

C) 4chan is people having discussions amongst themselves. Since you don't like or agree with the subject matter or the views of those who spend time communicating on 4chan (because you're an obvious HIV+ fa████ you are attempting to persuade people to believe that free speech is bad, racism is a completely invalid human reaction to danger, etc.

D) the harder people like you try to silence those who you disagree with the more obvious you make it to people that it is in fact you who are wrong.

E) F████ you. You're an idiot. Enjoy wasting all of your time wrapped and coddled by academia and then getting raped and murdered by a n████.

You're a terrible educator.

Can you please direct me further?

Thank you,
Emilee Reid

What was happening?

Alt-Right

Emilee Reid <emileereid4@gmail.com>
Yesterday, 16:45
Stringhini, Gianluca

Flag for follow up. Completed on 28 April 2017.

Hello,

This recent election has been a s...
lining, it has compelled the rest

As a Jew, it's been especially sta...
Muslim, and anti-whomever.

The coincidence of timing between this trend and the...
House play directly upon, and stoke further, the sorts of prejudices that can take violent forms and that may be manifested in overturned gravestones in a cemetery.

We must support our Jewish community, LGBT brothers and sisters, and all others who will be victimized by the Trump regime.

With that being said...I am reaching out to find out if any universities are doing work to help stop the rise of hate speech online?? We have to start somewhere. I'd like to explore joining a group of activists who can explore fighting against this hate speech.

Can you please direct me further?

Thank you,
Emilee Reid

This is what we call a raid

Anonymous >>123205 That's not UCL.



4chan

What is 4chan?

- An image board
- Conversations grouped into threads
- Fixed number of threads alive at a given time
- New replies bump a thread to the top of the page
- **Anonymous**
- **Ephemeral** (median thread lifetime: 47 min)

What is 4chan?

4chan is a simple image-based bulletin board where anyone can post comments and share images. There are boards for a variety of topics, from Japanese animation and culture to videogames, music, and photography. Users do not need to register before participating in the community. Feel free to click on a board below that interests you and jump right in!

Be sure to familiarize yourself with the [Rules](#) before posting, and read the [FAQ](#) if you wish to learn more about how to use the site.

Boards

<u>Japanese Culture</u>	<u>Interests</u>	<u>Creative</u>	<u>Other</u>	<u>Adult (NSFW)</u>
Anime & Manga	Comics & Cartoons	Oekaki	Business & Finance	Sexy Be
Anime/Cute	Technology	Papercraft & Origami	Travel	Hardcore
Anime/Wallpapers	Television & Film	Photography	Fitness	Handson
Mecha	Weapons	Food & Cooking	Paranormal	Hentai
Cosplay & EGL	Auto	Artwork/Critique	Advice	Ecchi
Cute/Male	Animals & Nature	Wallpapers/General	LGBT	Yuri
Flash	Traditional Games	Literature	Pony	Hentai/A
Transportation	Sports	Music	Current News	Yaoi
Otaku Culture	Alternative Sports	Fashion	Worksafe Requests	Torrents
<u>Video Games</u>	Science & Math	3DCG	Very Important Posts	High Res
Video Games	History & Humanities	Graphic Design	<u>Misc. (NSFW)</u>	Adult GIF
Video Game Generals	International	Do-It-Yourself	Random	Adult Ca
Pokémon	Outdoors	Worksafe GIF	ROBOT9001	Adult Re
Retro Games	Toys	Quests	Politically Incorrect	
			International/Random	
			Cams & Meetups	
			Shit 4chan Says	

Data Collection



Paper (from ICWSM 2017)

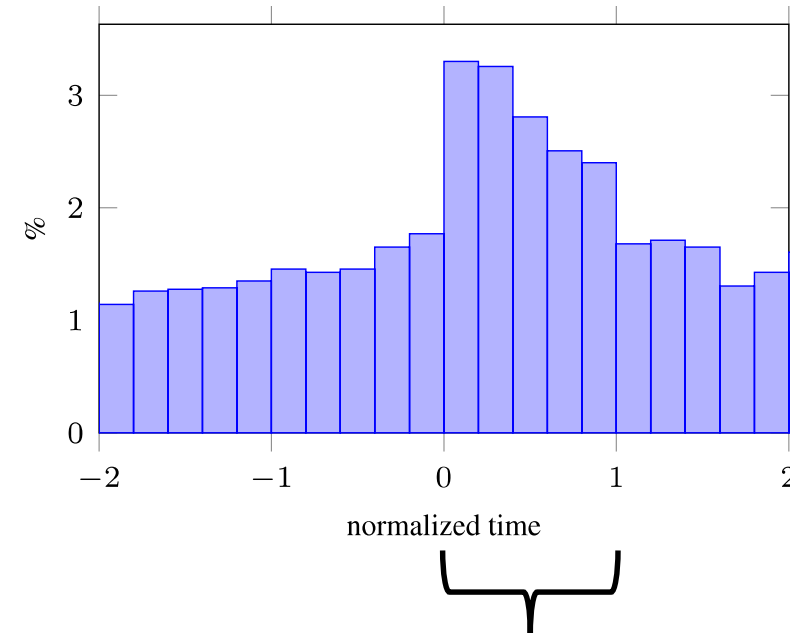
4chan is divided in boards, we focus on /pol/, the “politically incorrect” board

	/pol/	/sp/	/int/	Total
Threads	217K	14.4K	24.9K	256K
Posts	8.3M	1.2M	1.4M	10.9M

Raids against YouTube videos

YouTube is the domain most often linked on /pol/
(93k in our dataset)

Anecdotally, we know of many raids directed from
/pol/ towards YouTube

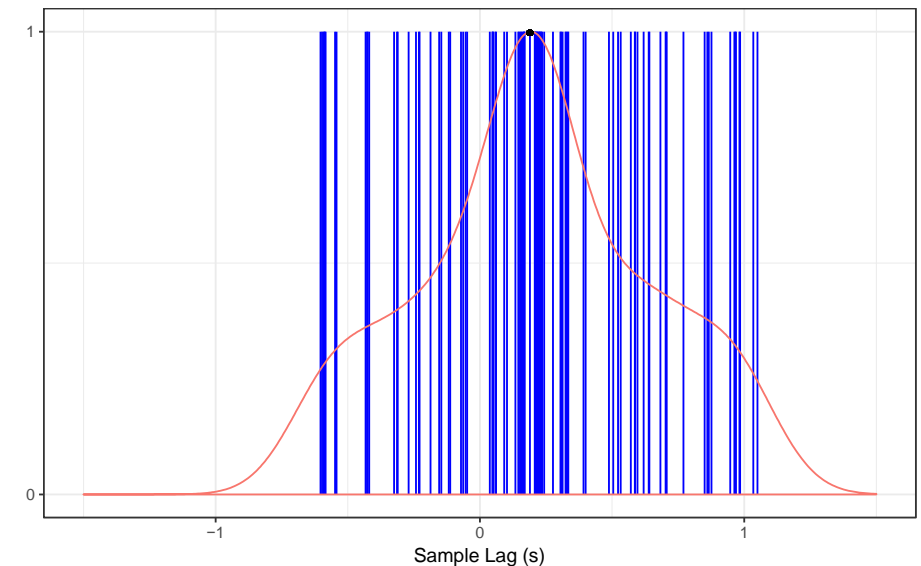
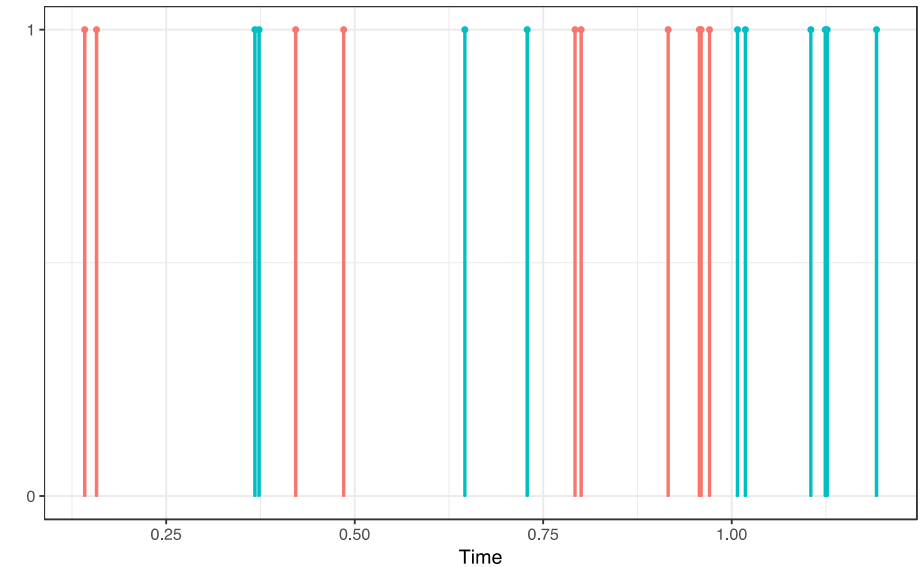


14% of videos see peak
commenting activity during
the /pol/ thread lifetime

Measuring synchronization

We use *cross-correlation* to estimate the **lag** between the two threads

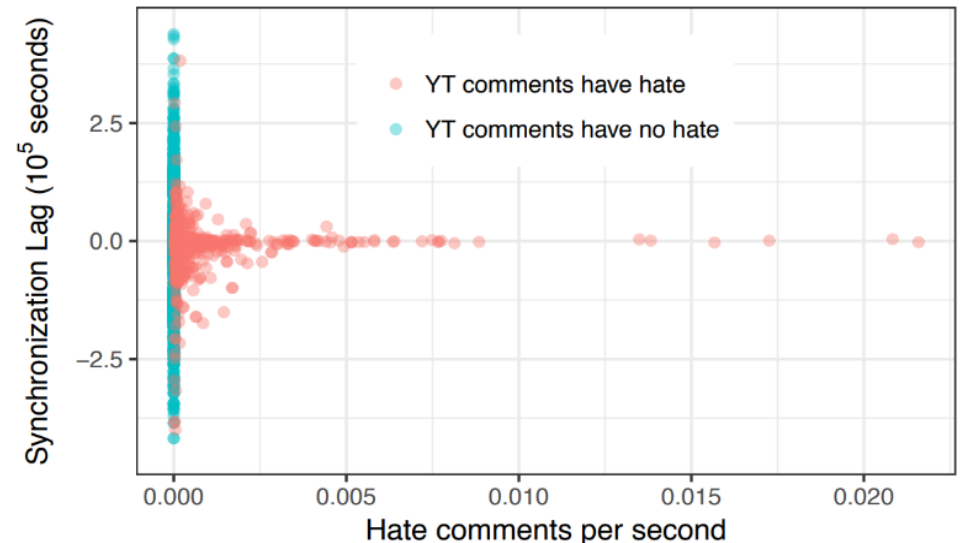
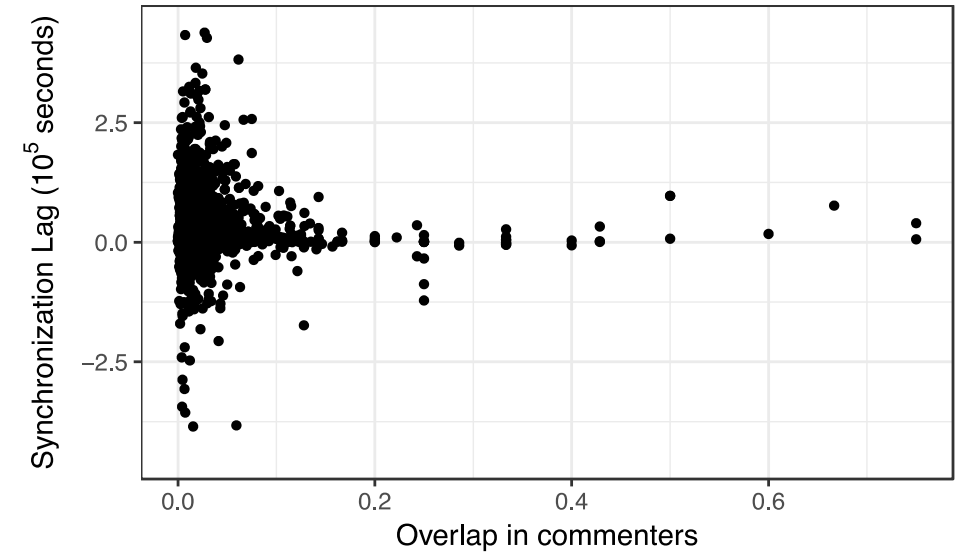
- We model the comments on /pol/ and YouTube as two Dirac Combs
- We slide a signal on the other for every possible lag value and calculate the dot product between the two signals
- We calculate the PDF of the dot products for all possible lags
- The estimated lag value is the one that maximizes this PDF



Identifying raided videos on YouTube

Videos showing a high degree of synchronization with /pol/ threads

- Tend to receive comments from common sets of accounts (*sock puppets*)
- Tend to attract comments that contain more hate speech (calculated through the *Hatebase API*)



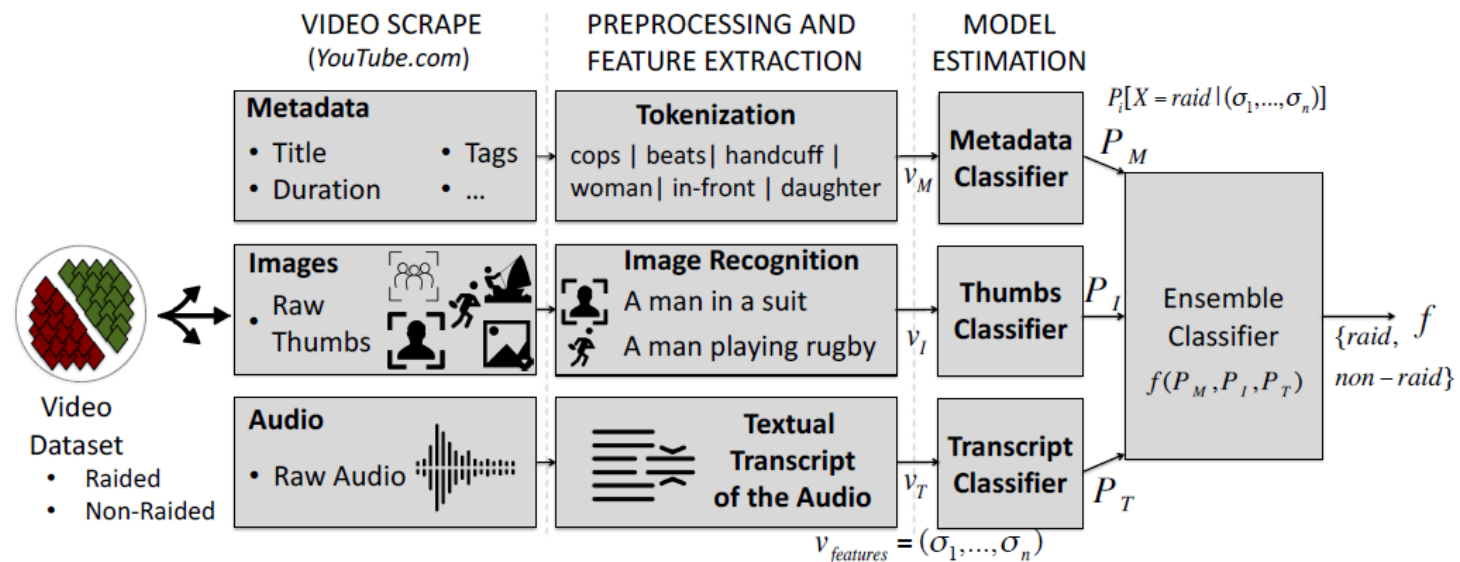


(Preprint)

What to do about raids?

Our measurement allowed us to collect a ground truth dataset of raided videos

Can we predict which videos will get raided? (Work in progress)



Sometimes simple solutions are the most effective: **disable comments for the lifetime of a /pol/ thread?**

There is so much more to do

- The Web is a big place, there are many communities
- How do we automatically extract meaning from images?
- There is more to text and images (videos, streaming)

What are the best countermeasures against online hate?

Conclusions

It is difficult to measure online hate

- Lots of heterogeneous communities
- Lots of different types of content
- Hard to automate – domain knowledge required

Established techniques to mitigate bot activity are not enough

- Loose coordination
- Human component

Credits



H2020 Marie Curie RISE program

ENCASE grant



Jeremy Blackburn
University of Alabama at
Birmingham



Emiliano De Cristofaro
University College London



Guillermo Suarez-Tangil
King's College London



Nicolas Kourtellis
Telefonica Research



Ilias Leontiadis
Telefonica Research



Michael Sirivianos
Cyprus University of
Technology



Tristan Caulfield
University College London



Despoina Chatzakou
Aristotle University of
Thessaloniki



Antigoni Founta
Aristotle University of
Thessaloniki



Gabriel Hine
Roma3 University



Jeremiah Onaolapo
University College London



Savvas Zannettou
Cyprus University of
Technology



Questions?

gian@bu.edu

[@gianluca_string](https://twitter.com/gianluca_string)