SECURITY (COMP0141): HATE AND HARASSMENT

FINANCIALLY MOTIVATED ABUSE

Social media: spammer accounts, promotional accounts

Messaging apps: spammer accounts, promotional accounts

App stores: SEO, bad apps

Airbnb: scam rentals

TripAdvisor: fake reviews, fake restaurants

Uber: colluding drivers

Dating apps: spammer accounts, romance scams

PREVENTING ABUSE

Content analysis like spam filters

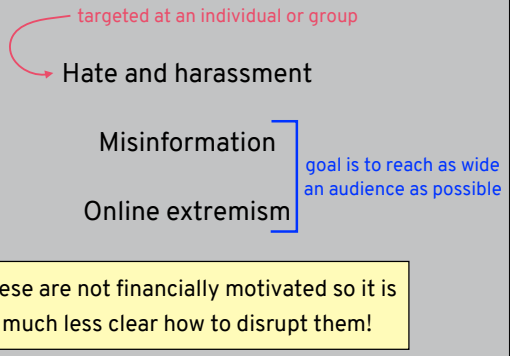
Identify bad accounts
- Correlate bad actions across accounts

Identify bad devices
- Use device cookies to correlate different accesses
- IP blacklists

Target infrastructure
- Botnet takedowns
- Close bank accounts

We've already seen a variety of ways to try to prevent / detect this financially motivated abuse

OTHER FORMS OF ABUSE



We'll focus only on hate and harassment in this lecture, but the other two are important problems as well. Some of this material may be challenging for some of you, as it discusses topics like toxic content (bullying, hate speech, etc.), stalking, and intimate partner violence.

THREAT MODEL

overall goal: inflict emotional harm

other possible goals:

- silence the target
- damage their reputation
- reduce safety (physical, sexual, etc.)
- coerce the target

amplification: ability to spread attack more widely
privileged access: know the target personally
deception: of an authority or broader audience
visible: to the target / broader audience

See a diverse set of possible motivations and capabilities here

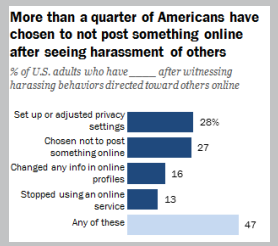
TOXIC CONTENT

Toxic content is designed to offend the target

Attack on availability: prevent target from taking part

Examples include:

- Bullying
- Trolling
- Hate speech
- Threats of violence
- Sexual harassment
- Unwanted explicit content



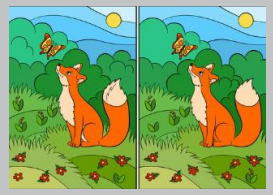
Strategies here tend to focus more on detection than prevention

Personal strategies tend to focus on just trying to avoid toxic content, not prevent it from happening

DETECTING TOXIC CONTENT

Can also develop classifiers (like for spam)

Perceptual hashing can check similarity to known bad content



$H(image_1)$ close to $H(image_2)$ if $image_1$ close to $image_2$

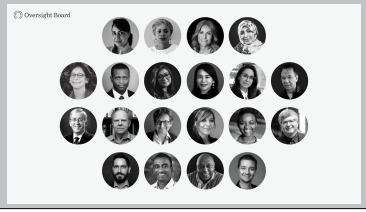
Perceptual hashing is sort of the opposite of cryptographic hashing – instead of uniformity we have that hashes are similar if content is similar

DETECTING TOXIC CONTENT

Can also develop classifiers (like for spam)

Perceptual hashing can check similarity to known bad content

Most platforms have (large) teams of contracted human moderators who remove content that lies outside the platform's terms of use



Human moderation is limited because terms of use are narrowly focused and definition of toxic content is subjective / varies across cultures – it also takes a significant toll on the moderators themselves. Facebook has even gone so far as to create an oversight board to make these decisions

RAIDING: MY OWN EXPERIENCE

Raiding (also called brigading) is when people join together and coordinate on one platform to harass a target on another, the term comes from coordinated attacks in an MMORPG

RAIDING: MY OWN EXPERIENCE

1 Nov 2016: give talk

RAIDING: MY OWN EXPERIENCE

1 Nov 2016: give talk

11 Nov: talk posted to YouTube

RAIDING: MY OWN EXPERIENCE

1 Nov 2016: give talk

11 Nov: talk posted to YouTube

17 Dec: Reddit thread created

RAIDING: MY OWN EXPERIENCE

1 Nov 2016: give talk

11 Nov: talk posted to YouTube

17 Dec: Reddit thread created

19 Dec: YouTube comments disabled

"Never has someone more uninformed been near a microphone."

"Lady, you do not know what the f*** you are talking about. You're just an attention seeker who randomly pick a topic and run your mouth about it. Cut the crap, you're only making yourself look stupid."

"You are obviously confused, or mixed up about the technical aspects of bitcoin. It's actually scary how you are explaining something to people you clearly do not understand at it's most fundamental level. WOW oh WOW oh WOW The longer I watch the more problems I hear, it's actually getting funny to hear how someone can miss-interpret so many different aspects of bitcoin. If you spent half as much time thinking about what you are saying instead of making these cool little graphics you'd be a lot further. Oh yeah, maybe check out some other technical bitcoin papers to avoid looking like a moron in the future."

"You should be ashamed of yourself! Your knowledge is extremely flawed and misleading."

RAIDING: GAMERGATE

Gamergate was a sustained harassment campaign against female and non-binary gaming developers (and people defending them)

- Started with false claims made by an ex-boyfriend
- Threats of rape and murder, doxing, leak of sensitive images

Campaign was organised on anonymous message boards...

4chan → 8chan

...and carried out via sock-puppet and other accounts on Twitter

OVERLOADING

Raiding is an example of overloading, in which an attack is amplified by many participants (or by technical means)

Attack on availability: hard to find the “real” content

Other examples include:

- DDoS (saw in Week 5)
- Flooding a review site with negative reviews
- Notification bombing

Overloading is an attack on availability (as we already explored when we looked at DDoS in Week 5)

PREVENTING OVERLOADING

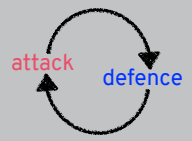
We already saw that preventing DDoS is hard! Approaches here are similarly heuristic

For raiding/brigading:

- Identify features of YouTube videos that are likely to be raided, proactively disable comments
- Identify raid as it’s happening and disable comments

For notification bombing:

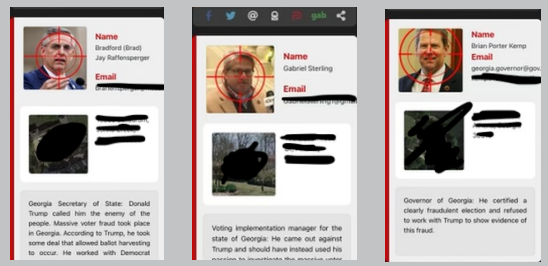
- Don’t allow these apps in the app store



Banning apps from the app store is hard, they are marketed as pranks – again, this is a shifting target

DOXING

In the 2020 US election, public officials in Georgia and other states had their personal details revealed, including email addresses, phone numbers, and home addresses



17

Doxing is often carried out as a type of revenge, or attempt to coerce the target

CONTENT LEAKAGE

Doxing is an example of content leakage, in which an attacker leaks (or threatens to leak) sensitive information about the target

Attack on confidentiality

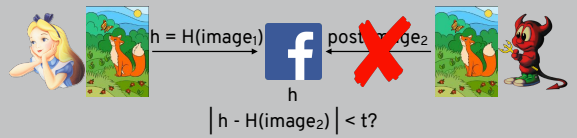
Other examples include:

- Non-consensual image exposure (“revenge porn”)
- Outing and deadnaming
- Sextortion

18

PREVENTING CONTENT LEAKAGE

Can rely again on perceptual hashing



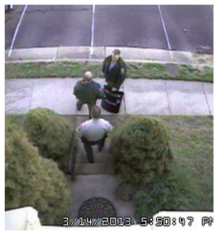
Also rely again on human moderation to remove content that seems to fit into this category

Platforms could also stop users from forwarding sensitive material

SWATTING

A prominent security journalist, Brian Krebs, had a SWAT team called to his house in 2013 (also had heroin sent to him and was subjected to a DDoS attack in the same year)

When I opened the door to peel the rest of the tape off, I heard someone yell, "Don't move! Put your hands in the air." Glancing up from my squat, I saw a Fairfax County Police officer leaning over the trunk of a squad car, both arms extended and pointing a handgun at me. As I very slowly turned my head to the left, I observed about a half-dozen other squad cars, lights flashing, and more officers pointing firearms in my direction, including a shotgun and a semi-automatic rifle. I was instructed to face the house, back down my front steps and walk backwards into the adjoining parking area, after which point I was handcuffed and walked up to the top of the street.



Fairfax County Police outside my home on 3/14/13

Full report at https://krebsonsecurity.com/2013/03/the-world-has-no-room-for-cowards/

FALSE REPORTING

Swatting is an example of false reporting, in which an attacker falsely accuses a target of abusive behaviour

Attack on integrity

Other examples include:

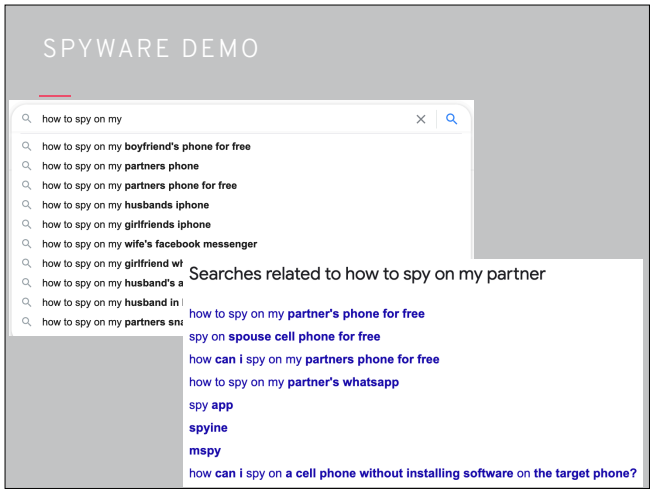
- Falsified flagging of abusive content
- Falsified abuse report

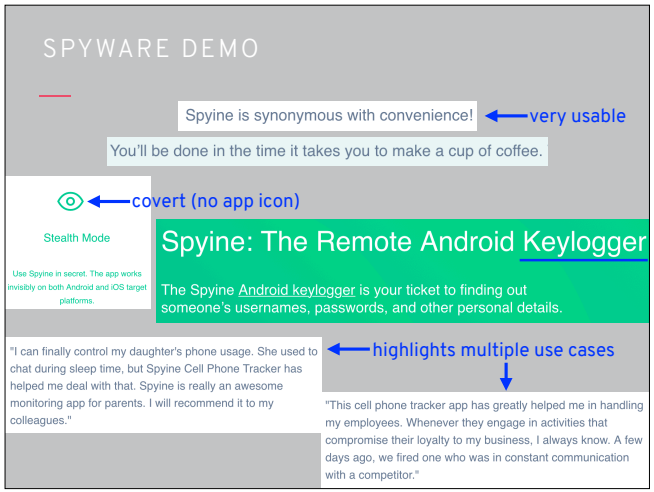
PREVENTING FALSE REPORTING

Classifiers for detection (like for spam)

Targets can proactively contact law enforcement / platforms

Take reputation into account when reacting to flags / reports


SPYWARE DEMO
how to spy on my
how to spy on my boyfriend's phone for free
how to spy on my partners phone
how to spy on my partners phone for free
how to spy on my husbands iphone
how to spy on my girlfriends iphone
how to spy on my wife's facebook messenger
how to spy on my girlfriend wh
how to spy on my husband's a
how to spy on my husband in
how to spy on my partners sna
Searches related to how to spy on my partner
how to spy on my partner's phone for free
spy on spouse cell phone for free
how can i spy on my partners phone for free
how to spy on my partner's whatsapp
spy app
spyine
mspy
how can i spy on a cell phone without installing software on the target phone?


SPYWARE DEMO
Spyine is synonymous with convenience!
very usable
You'll be done in the time it takes you to make a cup of coffee.
covert (no app icon)
Stealth Mode
Use Spyine in secret. The app works invisibly on both Android and iOS target platforms.
Spyine: The Remote Android Keylogger
The Spyine Android keylogger is your ticket to finding out someone's usernames, passwords, and other personal details.
"I can finally control my daughter's phone usage. She used to chat during sleep time, but Spyine Cell Phone Tracker has helped me deal with that. Spyine is really an awesome monitoring app for parents. I will recommend it to my colleagues."
highlights multiple use cases
"This cell phone tracker app has greatly helped me in handling my employees. Whenever they engage in activities that compromise their loyalty to my business, I always know. A few days ago, we fired one who was in constant communication with a competitor."

SURVEILLANCE / LOCKOUT

Device monitoring is an example of **surveillance**, in which an attacker uses their privileged access to the target to monitor them

Attack on confidentiality

Other examples include:

- IoT monitoring
- Browser monitoring
- Stalking

Lockout and control is a more active attack in which the attacker performs modifications (deletes emails, changes settings, etc.)

PREVENTING SURVEILLANCE / LOCKOUT

Very difficult when the threat is your parent or intimate partner!

Send **warnings** / reminders to users about delegated access to their devices, location, photos, etc.

Show **indicators** when sensitive resources are being accessed

Don't allow these apps in the app store

TENSIONS

Free speech vs. hate speech

Empowering vs. burdening targets

Moderation vs. filter bubbles

Moderator review vs. well-being

Privacy vs. accountability

This is an emergent area, solutions here need to take many different tensions into account

ABUSE VS. COMPUTER SECURITY

Computer security: exploiting “unintended” functionality
Abuse: using system’s functionality

Where is the line between the two?

Traditional threat models don’t take into account many of the possible attackers here (parent, intimate partner, friend, ex-partner, etc.)

Platforms can be consciously designed with these threats in mind

This is very much part of security, and computer science more generally!