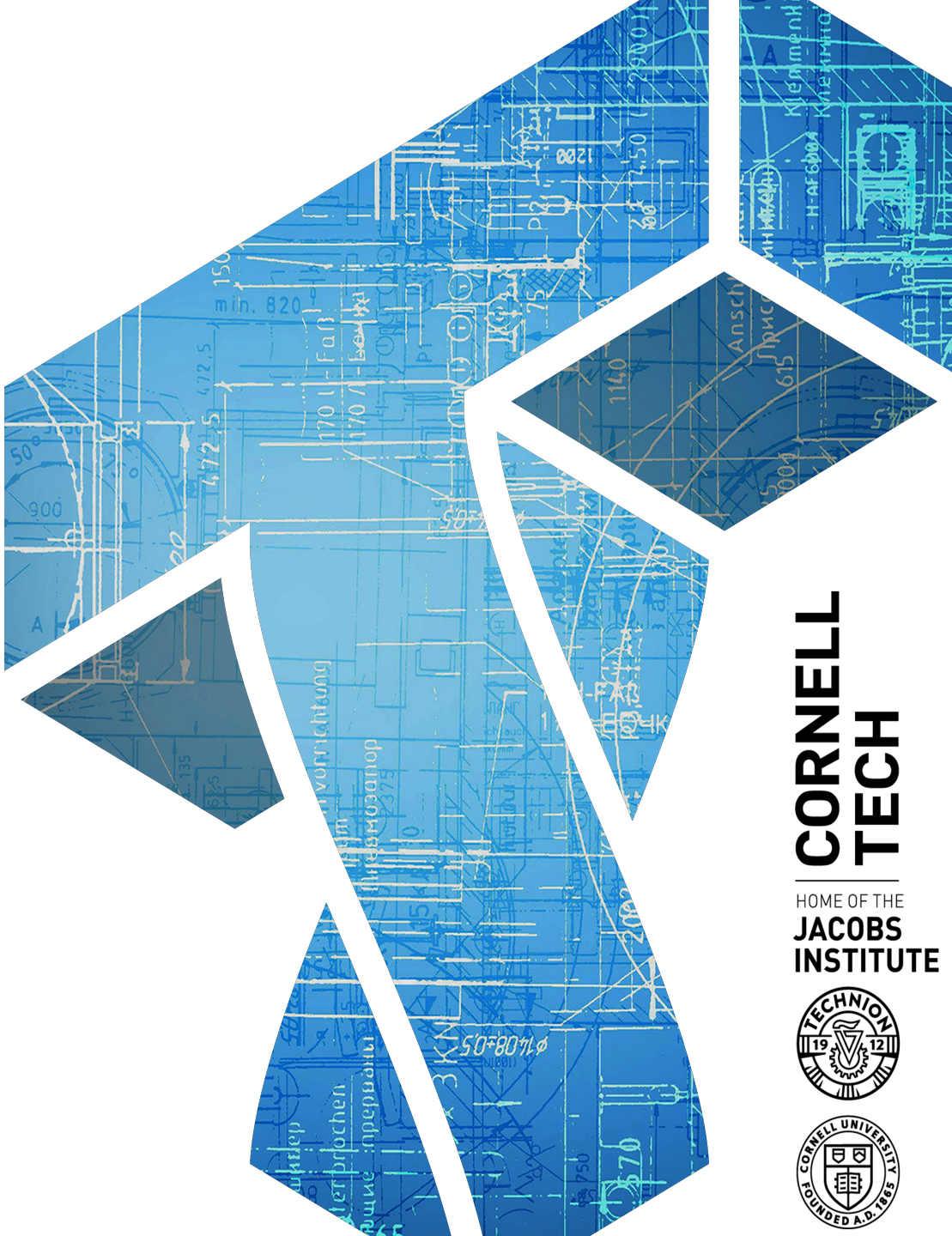
CS 5439:
Abusive messages and information disclosure

Tom Ristenpart



CORNELL TECH
HOME OF THE
JACOBS
INSTITUTE

Four categories of common attacks

Ownership-based

- Abuser owns device/account
- Shared account/device
- Buying children device

- Prevent use / destroy device
- Digitally control access
- Track location, monitor usage

Account/device compromise

- Physical access to unlocked device
- Force password / pin revelation
- Remotely “hack” via security questions / passwords

- Install spyware / “dual-use” app
- Track location, monitor victim
- Steal or delete info
- Lock victim out of account
- Impersonate victim

Harmful messages or posts

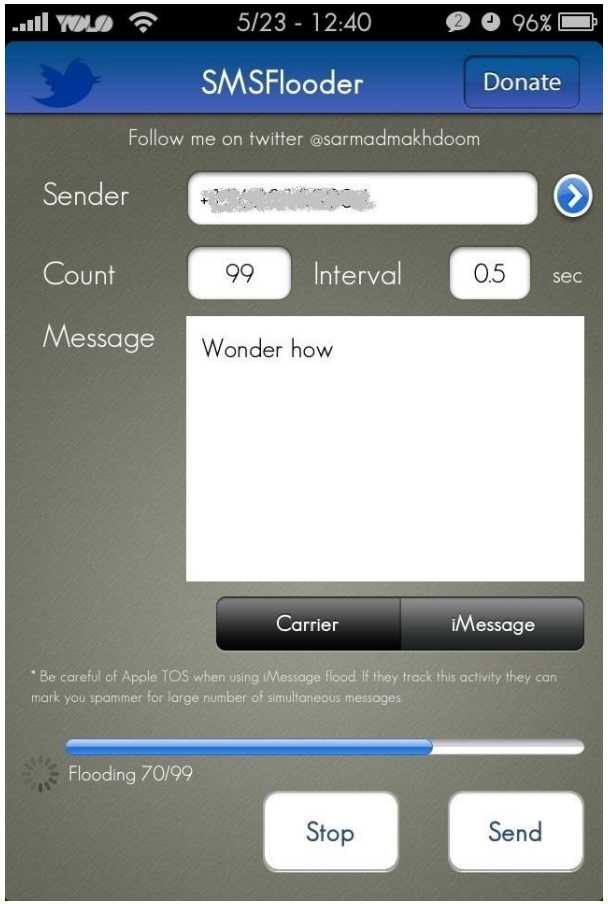
- Call/text/message victim (from spoofed account)
- Post harmful content (e.g., threaten violence)
- Harass victim’s friends/family
- Proxy harassment

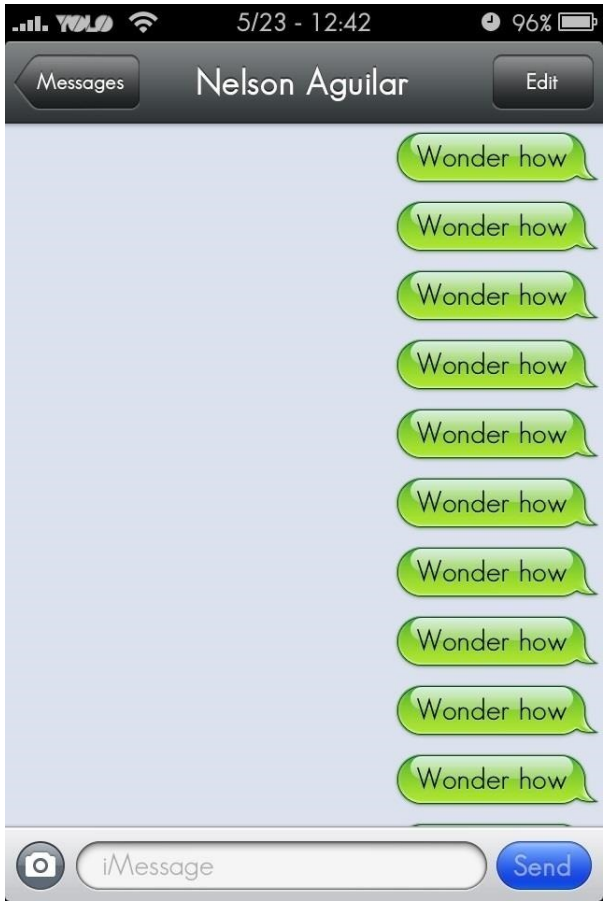
Exposure of private information

- Blackmail by threat of exposure
- “Doxxing” victim
- Non-consensual intimate images
- Fake profiles/advertisements of sexual services

Persistent messaging / calling

“upwards of 200 text messages a day; upwards of 160 calls a day.”
- Client

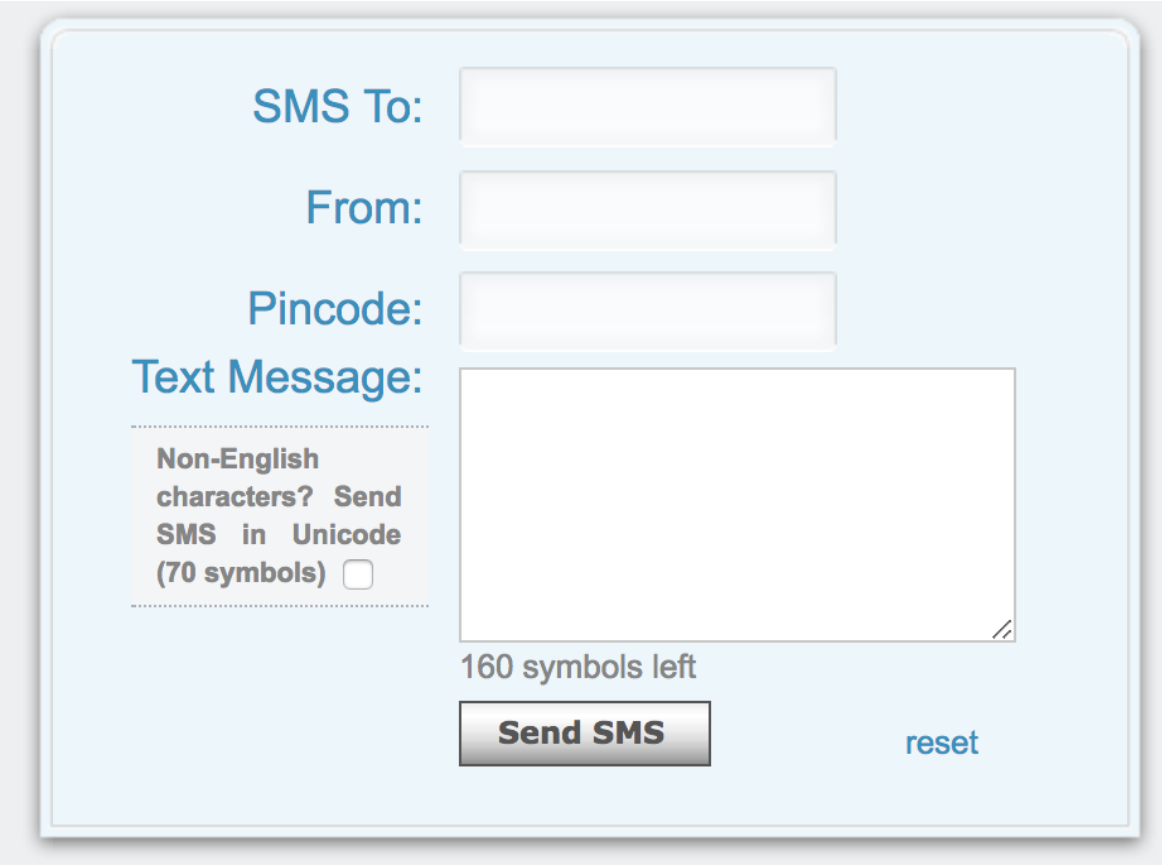




Spoofing to avoid blocks

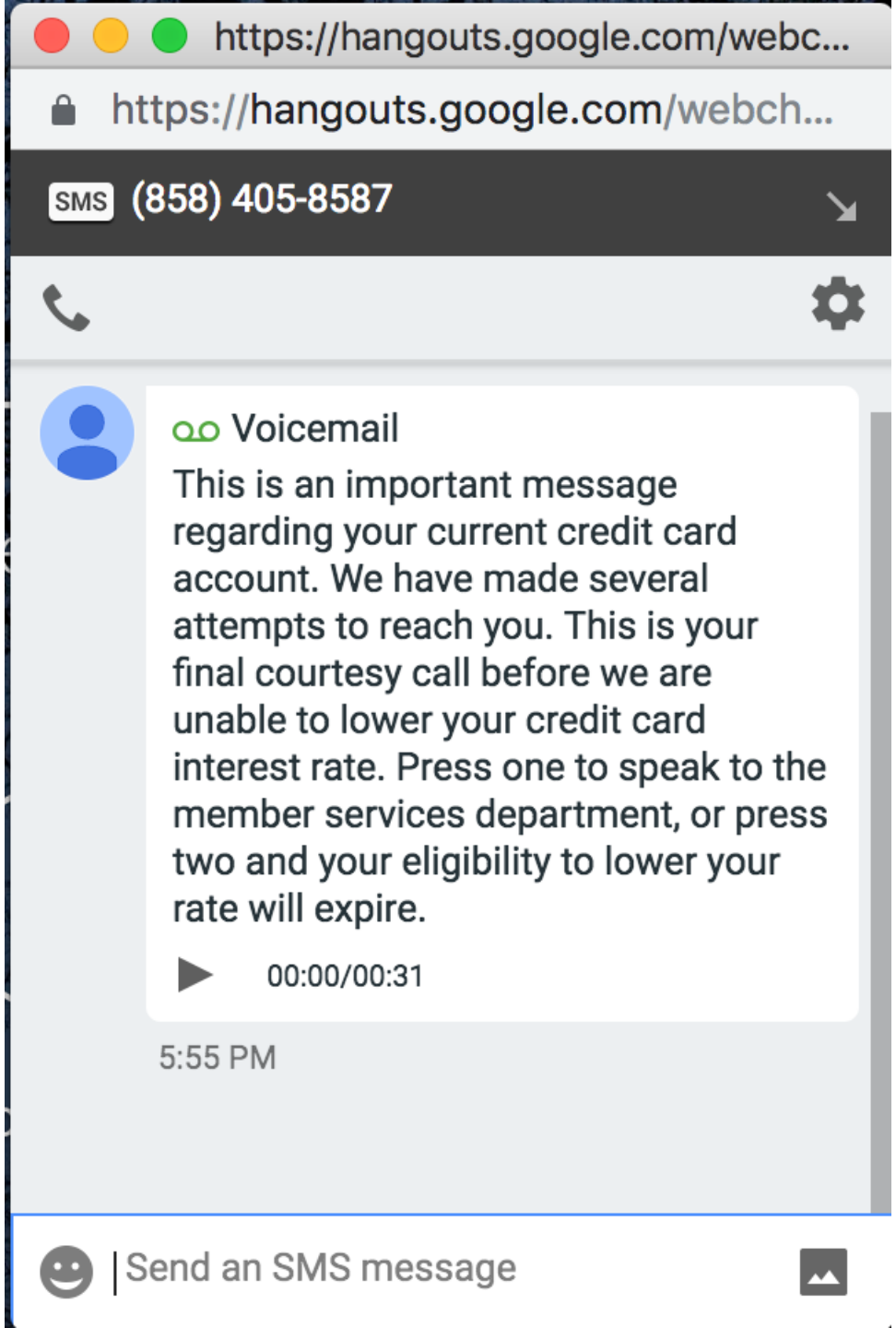
"You can put a fake number. So when you call someone, that fake number's going to come up. So they won't know it's you . . . Same thing with texting. They can text from a fake number or an online number . . . There's no way of tracking that back."

- Case manager



Twilio (SMS and phone call API provider) requires proof of ownership of sending number

US/Canada laws restrict spoofing



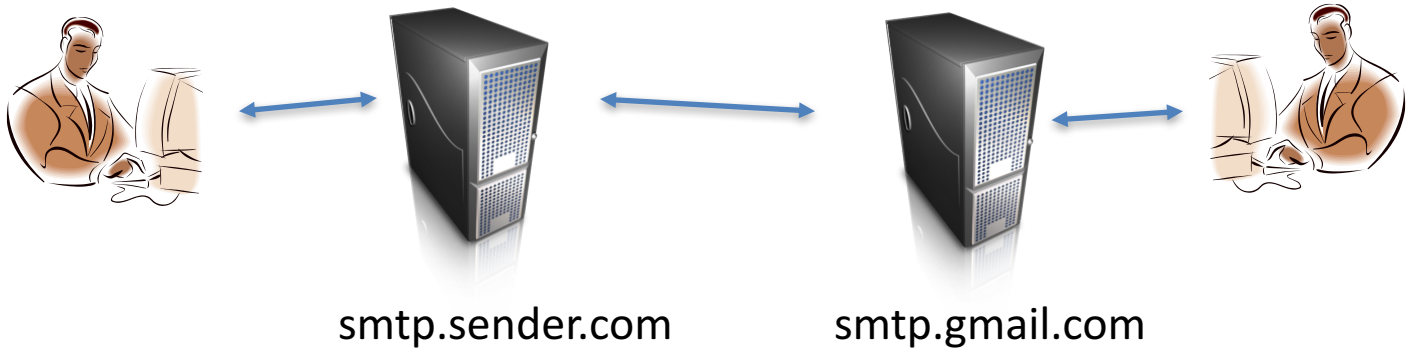
"Under the Truth in Caller ID Act, FCC rules prohibit any person or entity from transmitting misleading or inaccurate caller ID information with the intent to defraud, cause harm, or wrongly obtain anything of value. If no harm is intended or caused, spoofing is not illegal."

http://transition.fcc.gov/cgb/consumerfacts/callerid.pdf

Email spoofing

Send email as if it were from someone else's email address

Originally: email servers trusted whatever sender said sender email is



smtp.sender.com smtp.gmail.com

Example from Wikipedia:

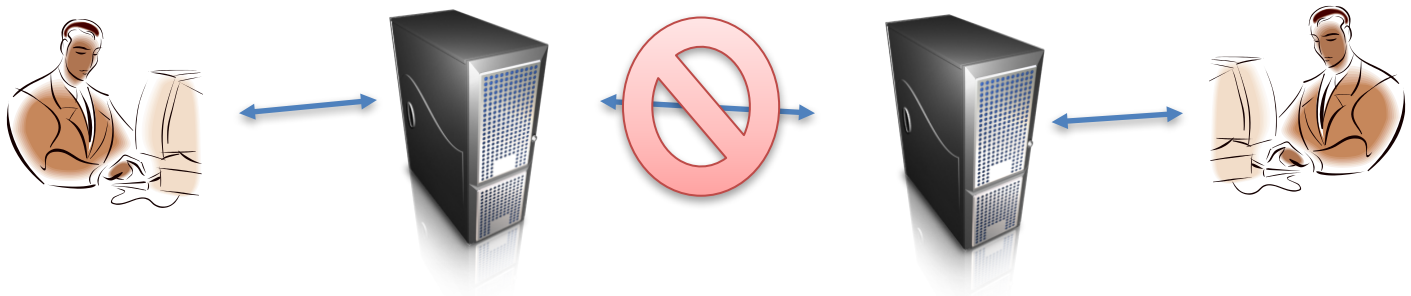
S: 220 smtp.example.com ESMTP Postfix
C: HELO relay.example.com
S: 250 smtp.example.com, I am glad to meet you
C: MAIL FROM:<bob@example.com>
S: 250 Ok
C: RCPT TO:<alice@example.com>
S: 250 Ok
C: RCPT TO:<theboss@example.com>
S: 250 Ok
C: DATA
S: 354 End data with <CR><LF>.<CR><LF>
C: From: "Bob Example" <bob@example.com>
C: To: Alice Example <alice@example.com>
C: Cc: theboss@example.com
C: Date: Tue, 15 January 2008 16:02:43 -0500
C: Subject: Test message
C:
C: Hello Alice

Email spoofing

Send email as if it were from someone else's email address

Originally: email servers trusted whatever sender said sender email is

Now: email servers use various mechanisms to check sender validity. Sender policy framework, DMARC, DNSBLs, TLS



smtp.sender.com smtp.gmail.com

Example from Wikipedia:

S: 220 smtp.example.com ESMTP Postfix
C: HELO relay.example.com
S: 250 smtp.example.com, I am glad to meet you
C: MAIL FROM:<bob@example.com>
S: 250 Ok
C: RCPT TO:<alice@example.com>
S: 250 Ok
C: RCPT TO:<theboss@example.com>
S: 250 Ok
C: DATA
S: 354 End data with <CR><LF>.<CR><LF>
C: From: "Bob Example" <bob@example.com>
C: To: Alice Example <alice@example.com>
C: Cc: theboss@example.com
C: Date: Tue, 15 January 2008 16:02:43 -0500
C: Subject: Test message
C:
C: Hello Alice

Spoofing used in various ways

- Avoid blocks
- Circumvents orders of protection
- Incriminate victim
  - Spoof as victim's phone # / email, send abusive messages to abuser

Social media abuse

"Facebook is really a stalker's paradise."

- Derogatory or otherwise harassing content on Facebook, Instagram, etc.
- Many subtleties arise:
  - Shared social circles & proxy abuse
  - Personalized vs. socially normative harassment
  - Subtleties of blocking and privacy defaults
  - Fake accounts and account reporting

Shared social circles & proxy abuse

Complicated social networks make disconnecting from abuser difficult

> *"A lot of times there's an order of protection, but if we have mutual friends and you commented on your friend's photo, I have every right to comment on your friend's photo too. I'm not talking to you, but we're on the same feed. So it gets really confusing."* - Case manager

Proxy abuse easy to mount on social media:

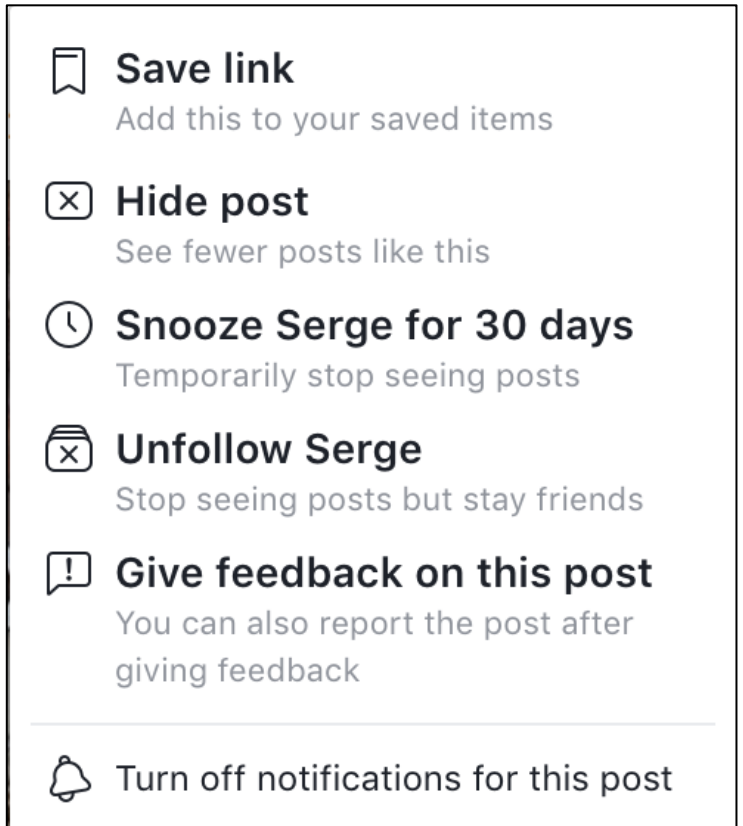
> "*I've had survivors where people have contacted them through [social media] . . . it's like, the new girlfriend of the abuser who's threatening. So that often happens, where there are threats made by third parties, sort of as a representative of the abuser*." - Attorney

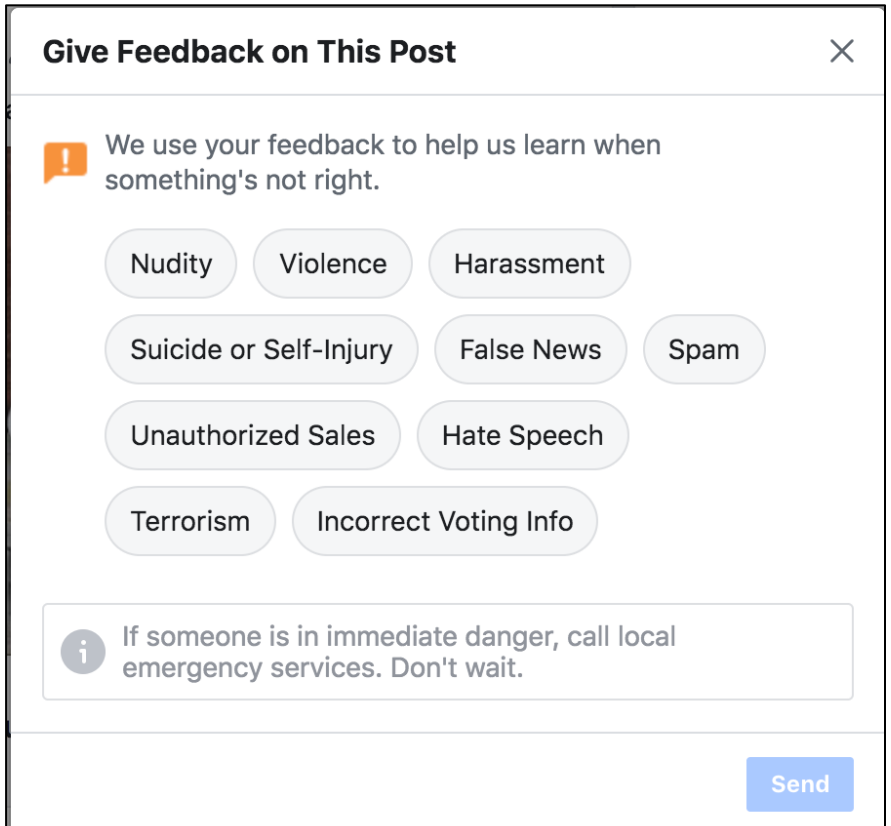
Personalized vs. socially normative abuse

Abusers sends personalized messages that don't look like abuse to moderators

> *"[The abuser] will post something on [social media], sometimes in code language . . . They'll say things that they know is a threat, but you might not think it's a threat at first if you're just looking at it."* - Attorney

So content can be reported, but won't appear to be abuse to platform





Report Harassment or Bullying on Instagram

Fill out this form to report photos, videos, comments or profiles on Instagram that are bullying or harassing others. Please provide as many details as possible to help us review this issue.

Do you have an Instagram account?

- ( ) Yes
- (•) No

Your email address

[ ]

Where is this happening?

Choose one:

- ( ) In a photo
- ( ) In a video
- ( ) In comments on a photo or video
- ( ) An entire profile is abusive

Send

Blocking & privacy defaults subtleties

Nuances significant in abuse settings

> *"Something that I see often is survivors being harassed by abusers through social media when there is an order of protection . . . a lot of fake Facebook profiles very obviously belong to the abuser but [the survivors] have no way to prove it.* ***Often it's because that person writes them a message, which you can do when you're not friends with someone on Facebook***." - Counselor

Privacy Settings and Tools

Your Activity	Who can see your future posts?	**Friends**	Edit
	Review all your posts and things you're tagged in		Use Activity Log
	Limit the audience for posts you've shared with friends of friends or Public?		Limit Past Posts
How People Find and Contact You	Who can send you friend requests?	**Everyone**	Edit
	Who can see your friends list?	**Public**	Edit
	Who can look you up using the email address you provided?	**Everyone**	Edit
	Who can look you up using the phone number you provided?	**Everyone**	Edit
	Do you want search engines outside of Facebook to link to your profile?	**Yes**	Edit

Block users

Once you block someone, that person can no longer see things you post on your timeline, tag you, invite you to events or groups, start a conversation with you, or add you as a friend. Note: Does not include apps, games or groups you both participate in.

Block users Add name or email **Block**

You haven't added anyone to your block list.

Block messages

If you block messages and video calls from someone here, they won't be able to contact you in the Messenger app either. Unless you block someone's profile, they may be able to post on your timeline, tag you, and comment on your posts or comments. Learn more.

Block messages from Type the name of a friend...

Featured Answer

Walter M Facebook Help Team

Hi Laura,

At this time, you do not have the ability to preventing people you do not know on Facebook from messaging you.

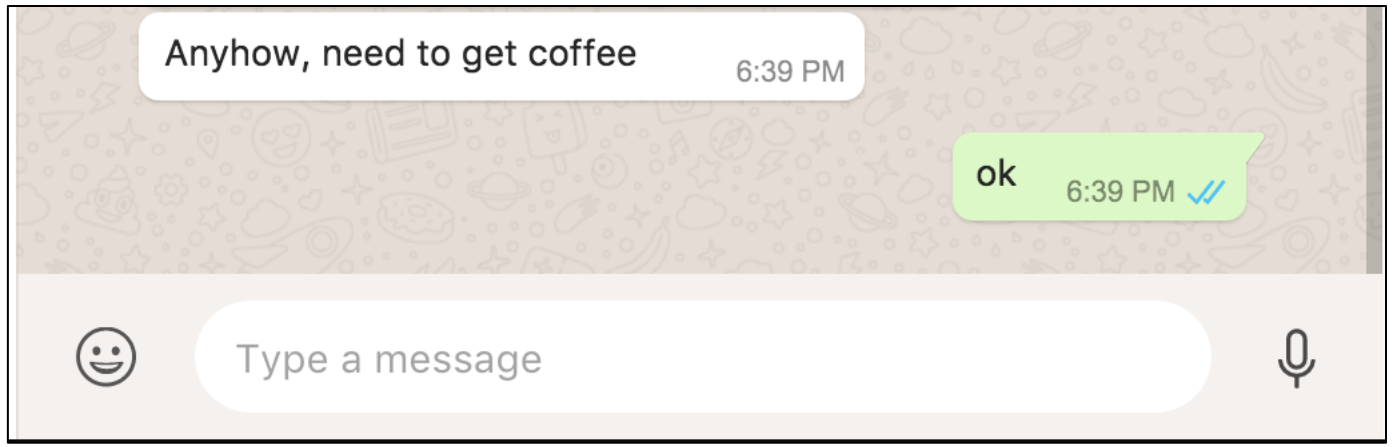
You'll get messages in your inbox from people who you're friends with on Facebook. If you get a message from someone who we think you might know, you'll get a message request. Spam messages are filtered out of your requests. To learn more about Message Requests, please visit our Help Center:

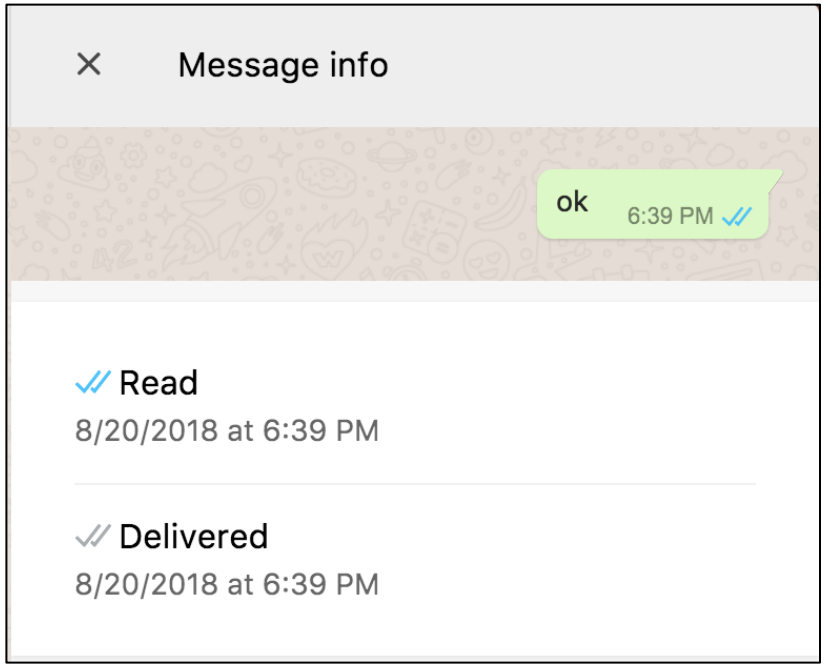
https://www.facebook.com/help/907368596013605/?ref=u2u

Blocking & privacy defaults subtleties

Nuances significant in abuse settings

> *"It was WhatsApp. The survivor didn't know how to turn off the notification to the other party about whether it's read or not . . . [the abuser] would go on a rant like, "I know you read this, I know you saw it" . . . which prompted more harassment."*
>
> - Case manager





Blocking & privacy defaults subtleties

Nuances significant in abuse settings

> *"It was WhatsApp. The survivor didn't know how to turn off the notification to the other party about whether it's read or not . . . [the abuser] would go on a rant like, "I know you read this, I know you saw it" . . . which prompted more harassment."*
>
> - Case manager

Turning off Read Receipts

To turn off your read receipts, go to **Menu button** > **Settings** > **Account** > **Privacy** and uncheck **Read receipts**.

Note: This won't disable the read receipts for group chats or play receipts for voice messages. There's no way to turn these settings off.

https://faq.whatsapp.com/en/android/28000015/?category=5245250

Blocking & privacy defaults subtleties

Nuances significant in abuse settings

"It was WhatsApp. The survivor didn't know how to turn off the notification to the other party about whether it's read or not . . . [the abuser] would go on a rant like, "I know you read this, I know you saw it" . . . which prompted more harassment."

- Case manager

How do I know if someone has seen a message I sent in Messenger?

Messenger uses different icons to let you know when your messages have been sent, delivered and read:

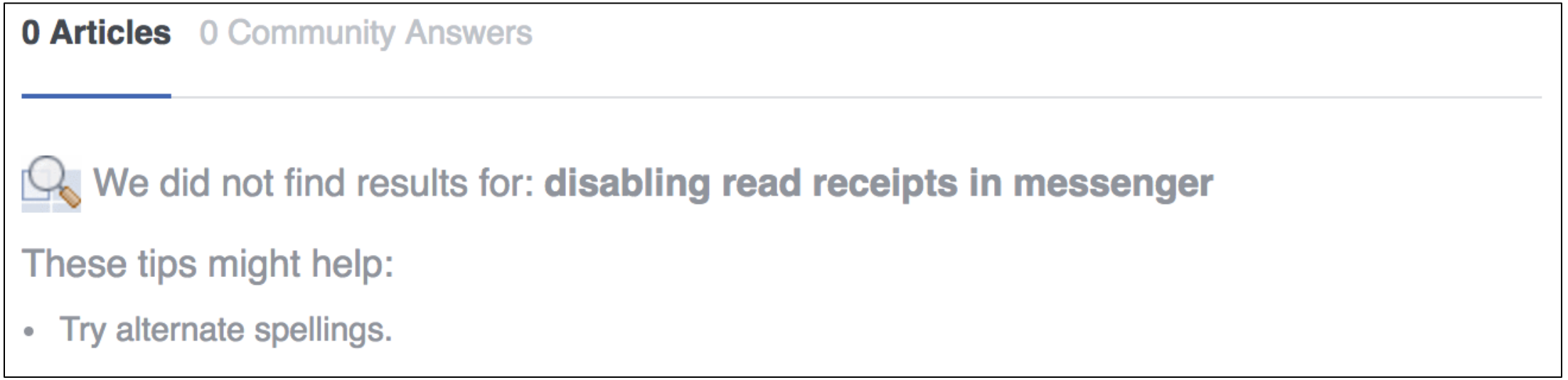
- : A blue circle means that your message is sending
- : A blue circle with a check means that your message has been sent
- : A filled-in blue circle with a check means that your message has been delivered
- : A small version of your friend or contact's photo will pop up below the message when they've read it

https://www.facebook.com/help/messenger-app/iphone/926389207386625/

Blocking & privacy defaults subtleties

Nuances significant in abuse settings

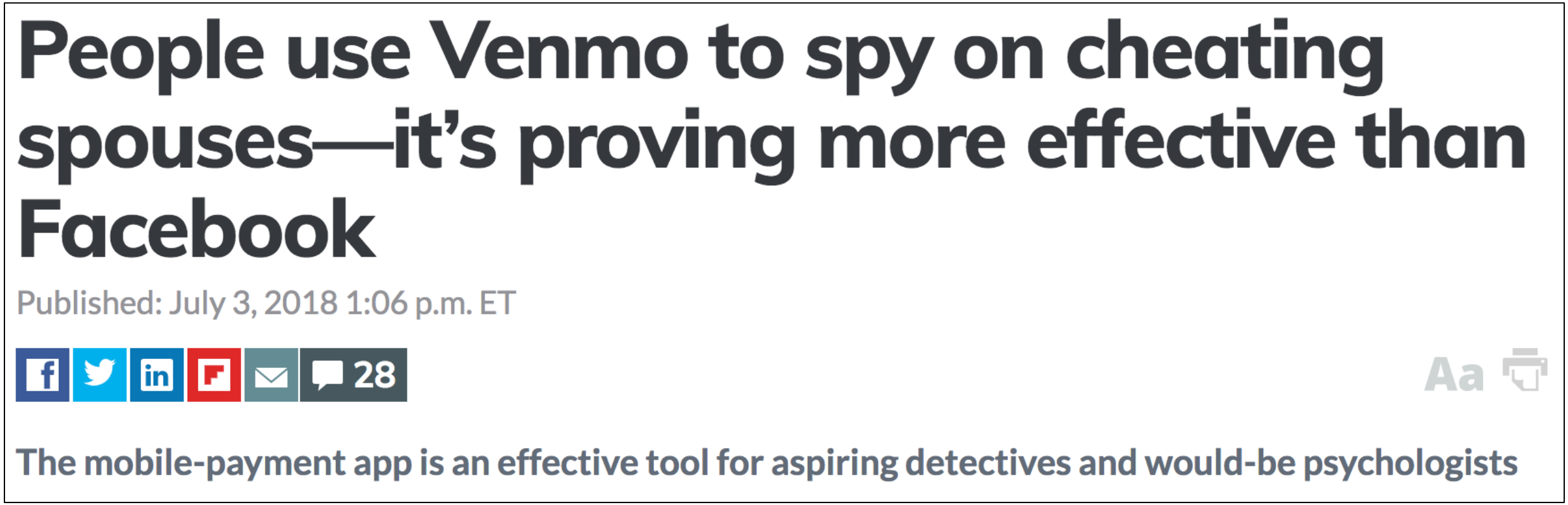
> *"It was WhatsApp. The survivor didn't know how to turn off the notification to the other party about whether it's read or not . . . [the abuser] would go on a rant like, "I know you read this, I know you saw it" . . . which prompted more harassment."*
>
> - Case manager



https://www.facebook.com/help/messenger-app/iphone/search/?query=disabling%20read%20receipts%20in%20messenger

Blocking & privacy defaults subtleties

Nuances significant in abuse settings



People use Venmo to spy on cheating spouses—it's proving more effective than Facebook

Published: July 3, 2018 1:06 p.m. ET

28

Aa

The mobile-payment app is an effective tool for aspiring detectives and would-be psychologists

Fake accounts

Abusers setup fake accounts avoid blocks, avoid orders of protection

> *"Something that I see often is survivors being harassed by abusers through social media when there is an order of protection . . .* ***a lot of fake Facebook profiles very obviously belong to the abuser but [the survivors] have no way to prove it****. Often it's because that person writes them a message, which you can do when you're not friends with someone on Facebook*." - Counselor

Fake account detection

Focus is on ***bot detection.*** Bots used by commercially-motivated attackers

Example:
[Stringhini, Kruegal, Vigna 2010] used honeyaccounts to identify spam bot accounts
The features they use to identify spam accounts:

- Friend requests / friends ratio
- URL ratio
- Message similarity
- # of distinct first names in friends
- # of messages sent (< messages is likely to be spam bot)
- # of friends

[Lee, Caverlee, Webb 2010] give even more detailed, similar analysis

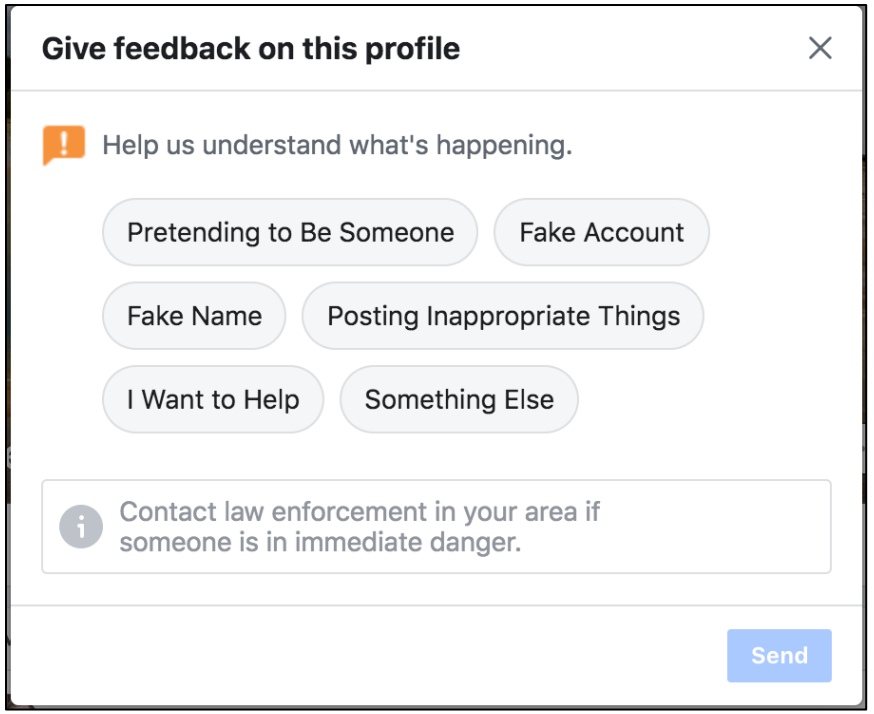
Abuse reporting challenges

Victims (and advocates) need to:

(1) flag abuser accounts
(2) ask for abusive content takedown

Professionals indicate some of the majors are doing relatively well (Facebook, Instagram mentioned), but long tail of uncooperative sites

Clients seemed generally frustrated, have trouble finding abuse reporting mechanisms and obtaining the results they wanted



"And even when you press something and say that you want to report this for offensiveness or abusiveness, they don't do anything about it."

– Client

Legal and policy issues

"You have to figure out how to tie it all together besides the Facebook because if you don't have an order of protection, he's not violating so he's just harassing. That's when I ask the other questions and then maybe we can tie it into stalking … she can get her order of protection." – Police officer

Abuse on Facebook not recognized as form of abuse that warrants an order of protection

Order of protection can legally restrict abusers from sending abusive Facebook messages

Challenges of managing abusive content

"Delete your Facebook completely. Delete everything so he doesn't have access to you this way. Just throw away your phone and get a new phone."

– Social worker

"Normally what the client does is block the person. But once they're blocked, they really don't know what's happening in the abuser's mindset. So they feel like he's going to show up. We say, "Okay, so what are some other ways he could contact you?" Well, we could do Facebook messages or WhatsApp . . . So that's the way they know how the abuser thinks or what might work. And at that point, if it's a stalker or someone that just wants communication, they'll take whatever you give them. They just don't want to get cut off."

- Case manager

Moving forward

- IPV feature review processes to assess security/privacy/safety implications for abuse victims
- Richer tools for managing abusive content
  - Blocking
  - Reporting
  - Archiving
- Legal improvements

Four categories of common attacks

Ownership-based

- Abuser owns device/account
- Shared account/device
- Buying children device

- Prevent use / destroy device
- Digitally control access
- Track location, monitor usage

Account/device compromise

- Physical access to unlocked device
- Force password / pin revelation
- Remotely “hack” via security questions / passwords

- Install spyware / “dual-use” app
- Track location, monitor victim
- Steal or delete info
- Lock victim out of account
- Impersonate victim

Harmful messages or posts

- Call/text/message victim (from spoofed account)
- Post harmful content (e.g., threaten violence)
- Harass victim’s friends/family
- Proxy harassment

Exposure of private information

- Blackmail by threat of exposure
- “Doxing” victim
- Non-consensual intimate images
- Fake profiles/advertisements of sexual services

Exposure of private information

“the stigma or what a community may think of [the victim] ... the perception of how they may be portrayed to their families and neighbors”

- Case manager

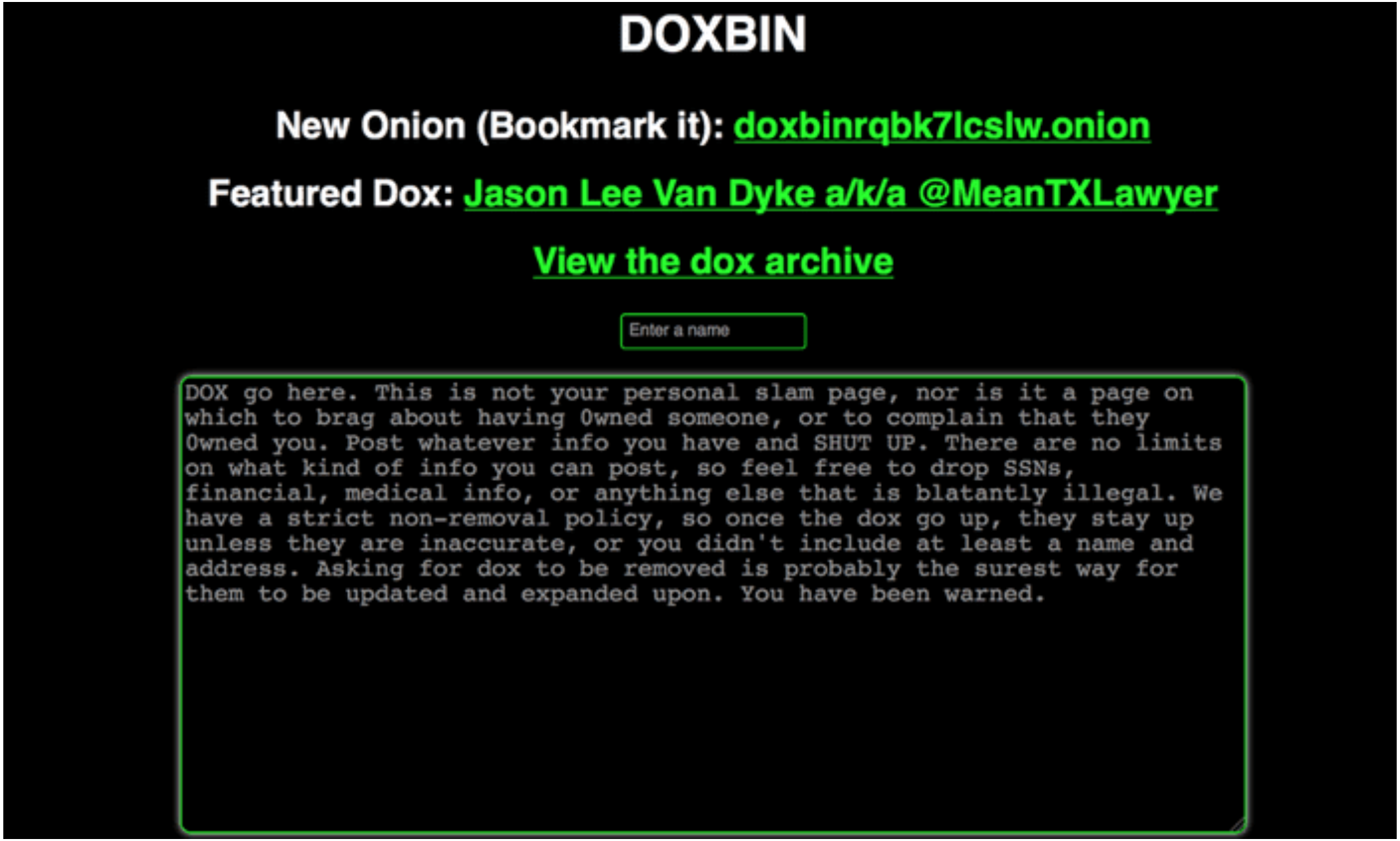
We have heard about lots of exposure-based harms:

- Blackmail by threat of exposure
- “Doxing” victim
- Fake profiles/advertisements of sexual services
- Non-consensual intimate images

Doxing

Finding sensitive information about a person and disclosing it online
Dox is an abbreviation of “documents”

Originally stems from hacker community: outing rival hacking groups



https://www.theguardian.com/technology/2014/dec/09/nachash-doxbin-dox-online

Doxing

Finding sensitive information about a person and disclosing it online
Dox is an abbreviation of “documents”

Originally stems from hacker community: outing rival hacking groups

Used as threat or (if performed) punishment of victim in IPV settings:

“Say, for example ... I’m HIV positive. My partner knows, but my family doesn’t know. It’s like, “I’m going to go on Facebook and post that you’re HIV positive . . . if you don’t do what I tell you” . . . and not only Facebook, there’s Instagram, Snapchat, Twitter.”

- Case worker

Information disclosure exploited for further harassment

Harassment by proxy
Catfishing: pretending to be someone else online

"The abuser made a fake Tinder account and put some really horrible explicit things and she was having tons, like upwards of like 25 people a night like ringing their doorbell. The abuser was corresponding with these people that wanted to hook up as if it was my client."

– Case manager

Can be difficult to get "*come rape me*" style ads removed because they are "*legally paid for*"

I want to report an impersonation or fake profile.

What do I do if someone is impersonating me?

If someone has created a Tinder profile using your photos or other personal information, please write to us and include the following information:

- The reason for the report
- The exact name, age, bio, and photos that appear on the profile that you are reporting (screenshots are best)

Other information that could help includes the user's location, phone number, email address, and/or a link to their Facebook account.

The more details you can provide, the quicker we can identify and investigate the profile or user in question.

https://www.help.tinder.com/hc/en-us/articles/115004950423-I-want-to-report-an-impersonation-or-fake-profile-

Hi there!

Since the only way to sign into Bumble is through Facebook, Bumble is a password-free app! You can try changing your password on Facebook, or we can delete your profile. Would you like us to delete your account for you?

https://www.vice.com/en_us/article/8x4jbg/when-harassers-use-tinder-and-bumble-to-dox-and-women

How to build pipelines for detecting fake accounts?

Non-consensual intimate imagery (NCII)

"Revenge-porn" is problematic term, but widely used

> *". . . he shared naked pictures of me . . . he also sent them to [public media]. . . He took my phone and he sent them through private messages to friends, but he also sent them through my email and my [social media] because he had the password. … he threatened to send them to [my work] . . . "*
>
> - Client

Websites condone NCII disclosures

Other technologies also used for sharing NCII

Revenge Porn Moves to Slack

In its continuing move from websites such as Anon-IB, revenge porn has shifted not only onto Discord, but now Slack as well.

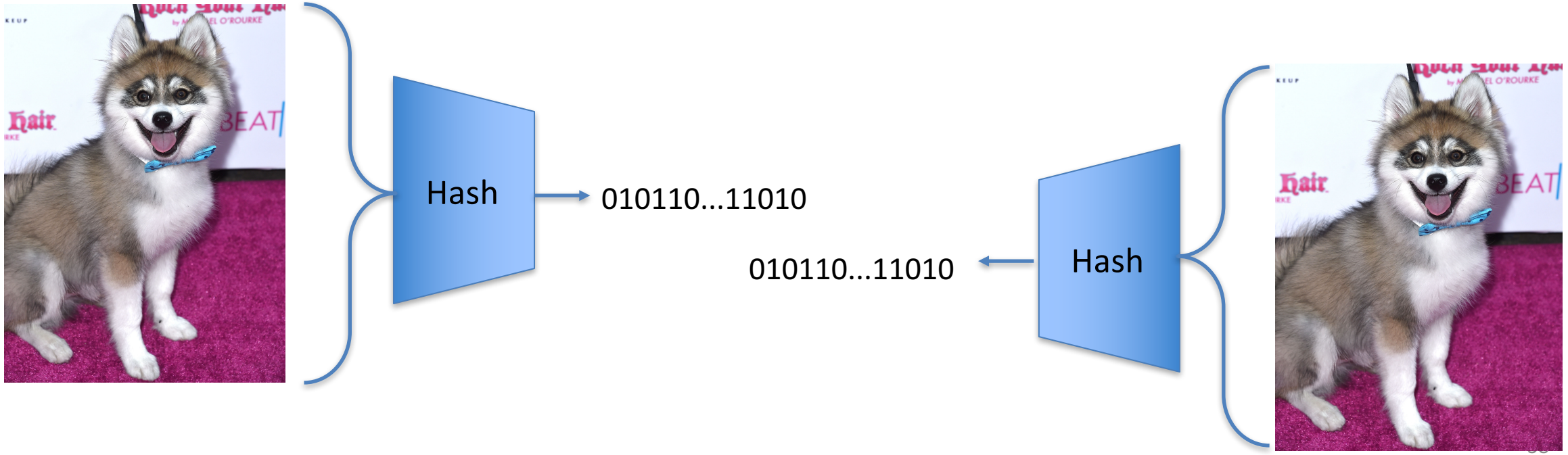
https://motherboard.vice.com/en_us/article/vbpaj8/revenge-porn-moves-to-slack

Dealing with NCII: Robust hashing

Comparing two images by their bit representations insufficient:

- Changes to size or cropping
- Changes to pixel values
- …

Robust hashing: build hash function for which similar images have small-distance hashes

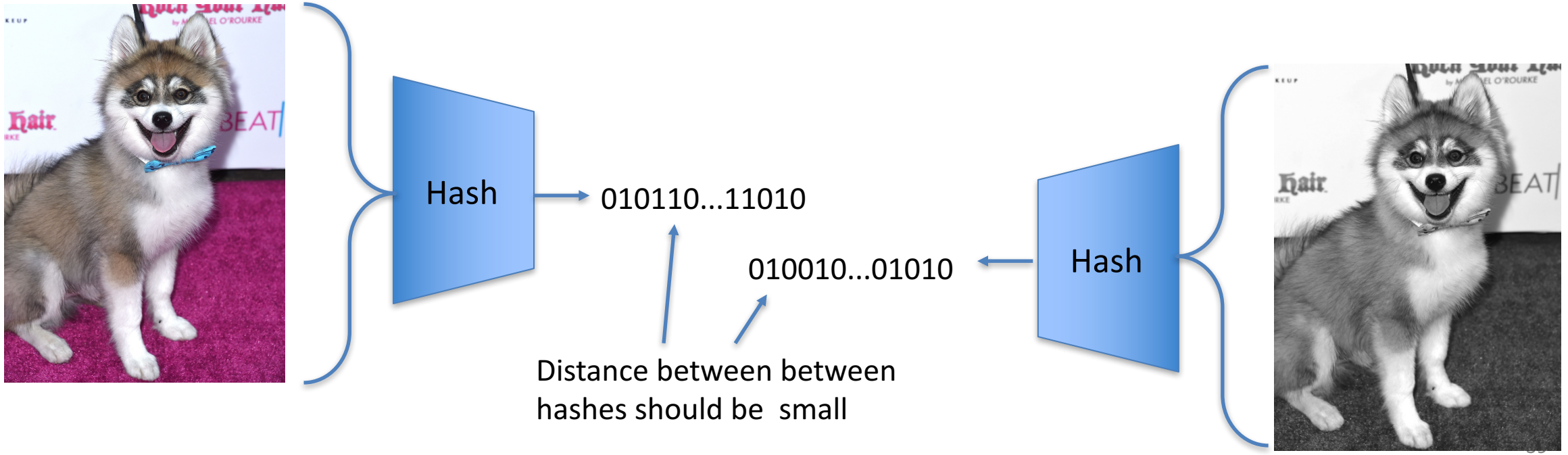


Dealing with NCII: Robust hashing

Comparing two images by their bit representations insufficient:

- Changes to size or cropping
- Changes to pixel values
- ...

Robust hashing: build hash function for which similar images have small-distance hashes



Dealing with NCII: Robust hashing

Apply image transformations to derive feature vectors more robust to changes

Split into grid, extract “statistical” properties

01010011

10110110

⋮

11110011

Result is feature vector that can be treated as bit string. Comparison of two hashes via Euclidean distance

How good is it? We don’t know: PhotoDNA a proprietary algorithm...

- Low matching rate for “distinct” images (false positives)
- Low non-matching rate for adversarial manipulation of image (false negatives)

Dealing with NCII: Facebook's recent feature

Can proactively report NCII

- Fill out report with a partner org (Australian e-Safety commisioner's office, Cyber Civil Rights Initiative & NNEDV in USA, UK Revenge Porn helpline, YWCA Canada)
- Facebook & agency email link to upload image
- Image reviewed by employee
- PhotoDNA hash of image generated, stored.
  - Image deleted (within one week)
- Hash used to flag uploading of image or slight variants

Facebook's plan to stop revenge porn may be even creepier than revenge porn

To halt revenge porn, potential victims may have to submit their own nudes to Facebook.

By Aja Romano | @ajaromano | Updated May 24, 2018, 2:15pm EDT

Dealing with disclosure of information

- Lots of hard challenges
  - Defining what is inappropriate content

    (See Facebook's leaked content policies)
  - Detecting accounts impersonating others
  - Ensuring abuse reporting / takedown mechanisms not themselves prone to abuse
    - Example: Google's commitment to transparently reporting takedown requests
- At least: think about it when building social media / communications platforms

http://www.trendmicro.co.uk/media/wp/exploring-the-deep-web-whitepaper-en.pdf (2015 report)