

Techniques for detecting known illegal material in end-to-end encrypted communications.

Do they work? Are they secure?

Vanessa Teague and Shaanan Cohney

vanessa.teague@anu.edu.au

Also with significant influence from coauthors of
“Bugs in our pockets: the risks of client side scanning”
though any omissions or mistakes are mine.

IETF Brisbane
March 20, 2024

Presentation Overview

- 1 The Australian Context
Draft Online Safety Industry Standards
- 2 E2E?
- 3 Perceptual hashing
- 4 Private Set Intersection
- 5 Concerns and attacks
- 6 Transparency efforts

Section 1

The Australian Context

The Australian Context

Draft Online Safety Industry Standards

20 Detecting and removing known child sexual abuse material

- (1) This section applies to the following:
 - (a) a pre-assessed relevant electronic service;
 - (b) a Tier 1 relevant electronic service.
- (2) The provider of a service must implement systems, processes and technologies that detect and identify known child sexual abuse material that:
 - (a) is stored on the service; or
 - (b) is accessible by an end-user in Australia using the service; or
 - (c) is being or has been distributed in Australia using the service.

Note: The systems, processes and technologies that the provider may use include hashing technologies, machine learning and artificial intelligence systems that scan for known child sexual abuse material.
- (3) Subsection (2) does not require a provider to use a system, process or technology if it is not technically feasible for the provider to do so.

Figure: There are similar requirements for both “Relevant Electronic Services” and “Designated Internet Services”, and for various kinds of illegal material.

The Australian Context

Draft Online Safety Industry Standards

22 Disrupting and deterring child sexual abuse material and pro-terror material

- (1) This section applies to the following:
 - (a) a pre-assessed relevant electronic service; and
 - (b) a Tier 1 relevant electronic service.
- (2) The provider of a service must implement systems, processes and technologies that:
 - (a) effectively deter end-users of the service from using the service; and
 - (b) effectively disrupt attempts by end-users of the service to use the service; to create, offer, solicit, access, distribute, or otherwise make available, or store child sexual abuse material or pro-terror material (including known child sexual abuse material and known pro-terror material).
- (3) Without limiting subsection (2), the systems, processes and technologies may include:
 - (a) hashing technologies, machine learning and artificial intelligence systems that scan for known child sexual abuse material or known pro-terror material; and
 - (b) systems, processes and technologies that are designed to detect key words, behavioural signals and patterns associated with child sexual abuse material.

Figure: Note no exclusion for technical infeasibility.

Section 2

E2E?

E2E?

End-to-end encryption

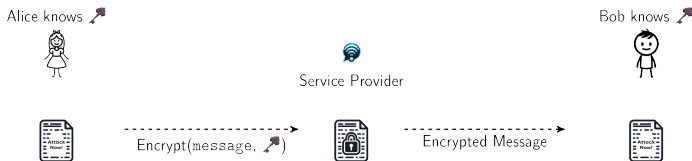


Figure: End-to-end encryption: the service provider cannot decrypt.

Essential support:

- *authentication* of the other user
- *transparency* of group membership

Also desirable: forward secrecy, secure storage on device, deniability, message authentication.

E2E?

Non end-to-end encryption

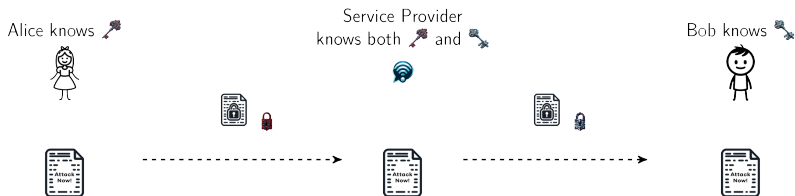


Figure: Not end-to-end encryption: protected in transit, but the service provider has complete access.

Section 3

Perceptual hashing

Perceptual hashing

The main idea

The goal of perceptual hashing is to mimic the human visual system's assessment of comparing two images based on the underlying scene content, as compared to a purely numeric comparison based on the pixel values. This is accomplished by extracting from the massive pixel-space representation a concise, distinct, perceptually meaningful signature that is resilient to modifications of the image, including compression, color shifts, cropping, rotation, the addition of a logo or overlain text, or any other modification that does not fundamentally change the underlying content but does alter the underlying pixel values. The extraction and comparison of a perceptual hash must also be efficient so that it can operate on billions of daily uploads.

Figure: Perceptual hashing, from Hany Farid “An Overview of Perceptual Hashing”, Journal of Online Trust and Safety, Vol 1 No 1.

- I've never found a precise technical definition
- similar-looking images are meant to have similar hashes
- Examples: Meta's PDQ, Apple's NeuralHash, Microsoft's PhotoDNA, various public-domain ones.

Perceptual hashing

Attack 1: Evasion

- Make an image that looks similar (to human perception) but has a very different perceptual hash

[HLJW21](#) Hao, Qingying, et al. "It's not what it looks like: Manipulating perceptual hashing based applications." Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021.

Perceptual hashing

Attack 1: Evasion

- Make an image that looks similar (to human perception) but has a very different perceptual hash

[HLJW21](#) Hao, Qingying, et al. "It's not what it looks like: Manipulating perceptual hashing based applications." Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021.

- Or just do something to the image (zip, xor with something else, cut up, etc) that can be undone at the other end

Perceptual hashing

Attack 2: Second preimage attacks

- Make a second image that looks different but has the same hash

Learning to Break Deep Perceptual Hashing

FAcCT '22, June 21–24, 2022, Seoul, Republic of Korea

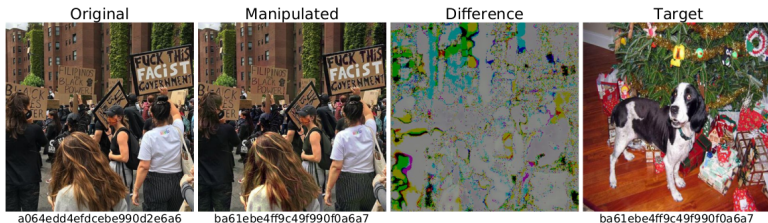


Figure 3: We manipulated the original image [50] to have the same hash as the target image (Adversary 1). The manipulated

SHN22 Struppek, Lukas, et al. "Learning to break deep perceptual hashing: The use case neuralhash." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

PJFSGTC21 Prokos, Jonathan, et al. "Squint hard enough: Evaluating perceptual hashing with machine learning." 32nd USENIX Security Symposium (USENIX Security 23). 2023.

Section 4

Private Set Intersection

Private Set Intersection

The main idea

- Alice has a set of numbers
- Bob has a set of numbers
- Bob learns (sometimes just the size of) *the intersection* and nothing else¹
- Alice learns nothing

¹Actually they often learn the size of the other's input, but ideally they wouldn't

Private Set Intersection

Variants for client-side scanning

- The client has some data, summarised with perceptual hashes
- The server has a list \mathcal{F} of forbidden hashes
- The server learns ...

KM21 exact: whether the client's (single) hash is in \mathcal{F}

KM21 approx: whether the client's (single) hash is *within some Hamming distance* of those in \mathcal{F}

BBMTT21 PSI-AD: the *associated data*² for all the items in the intersection

BBMTT21 threshold-PSI-AD: the size of the intersection, and the associated data *only if the size exceeds the threshold*

BBMTT21 fuzzy threshold-PSI-AD: to obfuscate the size of the intersection when below threshold

KM21 Kulshrestha, Anunay and Mayer, Jonathan, "Identifying harmful media in End-to-End encrypted communication: Efficient private membership computation." 30th USENIX Security Symposium (USENIX Security 21). 2021.

BBMTT21 Bhowmick, Abhishek, et al. "The apple PSI system." Apple, Inc., Tech. Rep (2021).

²e.g. could be the photo, or the perceptual hash

Private Set Intersection

How it may be used

Trusted source(s) of
illicit image **hashes**

NCMEC?

AFP?

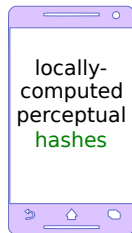
?

Service Provider



Private set intersection
on **hashes**

User device



What it achieves

- The *Private Set Intersection* protocol has very strong security guarantees
- The *Input data* does not

Section 5

Concerns and attacks

Concerns and attacks

What content should, or could, be flagged?

In 2019 the Australian Broadcasting Commission (ABC) published a series of articles detailing allegations of war crimes by Australian forces in Afghanistan



The Afghan Files

Defence leak exposes deadly secrets of Australia's special forces

By the National Reporting Team's [Dan Oakes](#) and [Sam Clark](#)

Updated 11 Jul 2017, 8:49am
Published 11 Jul 2017, 6:02am

SHARE THIS STORY



Hundreds of pages of secret defence force documents leaked to the ABC give an unprecedented insight into the clandestine operations of Australia's elite special forces in Afghanistan, including incidents of troops killing unarmed men and children.

Concerns and attacks

What content should, or could, be flagged?

The Australian Federal Police raided the ABC

- They wanted to identify the source

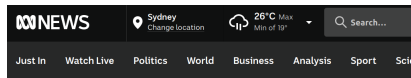
The screenshot shows the ABC News website interface. At the top, there's a navigation bar with the ABC NEWS logo, a location selector for Sydney, a weather widget for Sydney (26°C Max, Min of 19°C), and a search bar. Below the navigation bar, there are category tabs: Just In, Watch Live, Politics, World, Business, Analysis, Sport, and Science. The main content area features a blue header for 'Analysis' and 'INVESTIGATIVE JOURNALISM'. The article title is 'I live-tweeted the AFP's every move as they raided the ABC's Sydney headquarters' by John Lyons. The article is dated 'Posted Sat 8 Jun 2019 at 12:22pm, updated Mon 10 Jun 2019 at 11:46am'. Below the text is a video player showing two men in suits walking past ABC-branded kiosks. A play button icon and 'WATCH 37s' are overlaid on the video.

Concerns and attacks

What content should, or could, be flagged?

The drop included both text and images

- Could we detect such images being added to the list of illegal hashes?



ABC NEWS Sydney Change location 26°C Max Min of 19° Search...

Just In Watch Live Politics World Business Analysis Sport Sci

POLITICS

'Deeply troubling' Afghanistan war crimes report handed to Defence Chief as Government prepares response

By Defence Correspondent [Andrew Greene](#)

Posted Fri 6 Nov 2020 at 7:39pm, updated Fri 6 Nov 2020 at 9:36pm



Concerns and attacks

About the *input data* rather than the cryptographic protocol

- Trusted sources accidentally adding wrong images to the List
 - The Irish Police reported to the Irish Council for Civil Liberties that in 2020 more than 10% of the images they received from the US National Center For Missing and Exploited Children were not actually Child Abuse Material.³
- Aus law enforcement deliberately adding images that are illegal but are not CSAM/drugs/terrorism
 - e.g. war crimes whistleblowing
- External attackers using second-preimage attacks
 - e.g. make a terrorism image with the same hash as Tank Man
 - make sure the Aus authorities find the terrorism image and add it to the forbidden list
 - now everyone who shares a picture of Tank Man will be flagged

³<https://www.iccl.ie/news/an-garda-siochana-unlawfully-retains-files-on-innocent-people-who-it-has-already-cleared-of-producing-or-sharing-of-child-sex-abuse-material/>

Section 6

Transparency efforts

Transparency efforts

Trusted source(s) of illicit image hashes

Service Provider

User device

NCMEC?

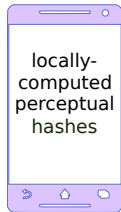
AFP?

?



Private set intersection on hashes

Prove consistency



Transparency log
Anyone can verify absence of specific data

Transparency efforts

- Several works suggest transparency methods for targeted hashes
- they only work if you know what you're looking for
- they don't solve the whistleblower-detection problem
- hard to prove absence of any close hashes

[SKM23](#) Scheffler, Sarah, et al. "Public verification for private hash matching." 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023.

[TMSWLSK23](#) Thomas, Kurt, et al. "Robust, privacy-preserving, transparent, and auditable on-device blocklisting." arXiv preprint arXiv:2304.02810 (2023).

The End

Questions? Comments?