

Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning

Shubham Jain*, **Ana-Maria Crețu***, Yves-Alexandre de Montjoye

Conference on Applied Machine Learning for Information Security (CAMLIS), Nov 4 2021

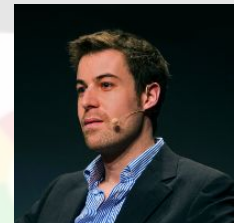
This talk is based on our USENIX 2022 [paper](#).



@shubhamjain0594



@anamariacretu5



@yvesalexandre

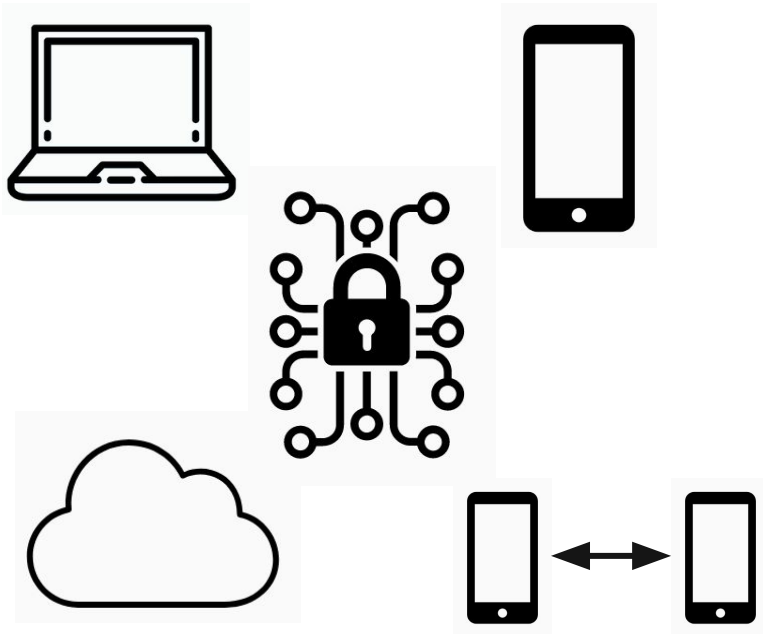
Imperial College
London



COMPUTATIONAL
PRIVACY
GROUP



Encryption is great for privacy and security...



... but allows illegal content to go undetected



Client-side scanning using perceptual hashing as a solution?



Identifying Harmful Media in End-to-End Encrypted Communication: Efficient Private Membership Computation

Anunay Kulshrestha and Jonathan Mayer, Princeton University
<https://www.usenix.org/conference/usenixsecurity21/presentation/kulshrestha>

Encryption and Combating Child Exploitation Imagery

By Nicholas Weaver Wednesday, October 23, 2019, 9:00 AM



This document has not been adopted or endorsed by the European Commission and is intended as a basis for discussion. It may not be shared further without permission of the European Commission services.

Technical solutions to detect child sexual abuse in end-to-end encrypted communications

Encryption, Privacy and Children's Right to Protection from Harm

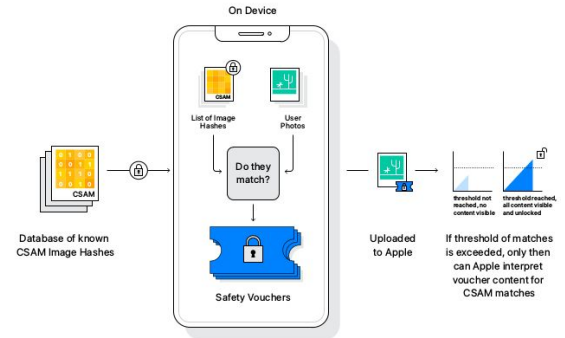
WhatsApp Monitor: A Fact-Checking System for WhatsApp

Philippe Melo,¹ Johnnatan Messias,² Gustavo Resende,¹
Kiran Garimella,³ Jussara Almeida,¹ Fabrício Benevenuto¹
¹Universidade Federal de Minas Gerais, ²MPI-SWS, ³MIT,
philipe@dcc.ufmg.br, johnme@mpi-sws.org, gustavo.jota@dcc.ufmg.br,
garimell@mit.edu, jussara@dcc.ufmg.br, fabricio@dcc.ufmg.br

Update as of September 3, 2021: Previously we announced plans for features intended to help protect children from predators who use communication tools to recruit and exploit them and to help limit the spread of Child Sexual Abuse Material. Based on feedback from customers, advocacy groups, researchers, and others, we have decided to take additional time over the coming months to collect input and make improvements before releasing these critically important child safety features.

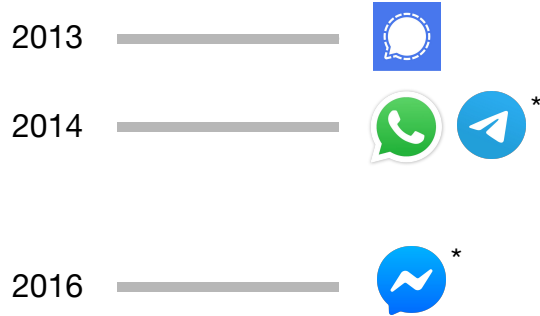
Expanded Protections for Children

At Apple, our goal is to create technology that empowers people and enriches their lives — while helping them stay safe. We want to help protect children from predators who use communication tools to recruit and exploit them, and limit the



Sources: https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf,
<https://ojs.aai.org/index.php/ICWSM/article/view/3271>, <https://www.lawfareblog.com/encryption-and-combating-child-exploitation-imagery>,
https://www.unicef-irc.org/publications/pdf/Encryption_privacy_and_children%E2%80%99s_right_to_protection_from_harm.pdf,
https://www.politico.eu/wp-content/uploads/2020/09/SKM_C45820090717470-1_new.pdf

The issue will not go away



POLITICO

OPINION

The last refuge of the criminal: Encrypted smartphones

Encryption unregulated is justice denied.

BY CATHERINE DE BOLLE AND CYRUS R. VANCE, JR.

July 26, 2021 | 4:01 am

> 2B users



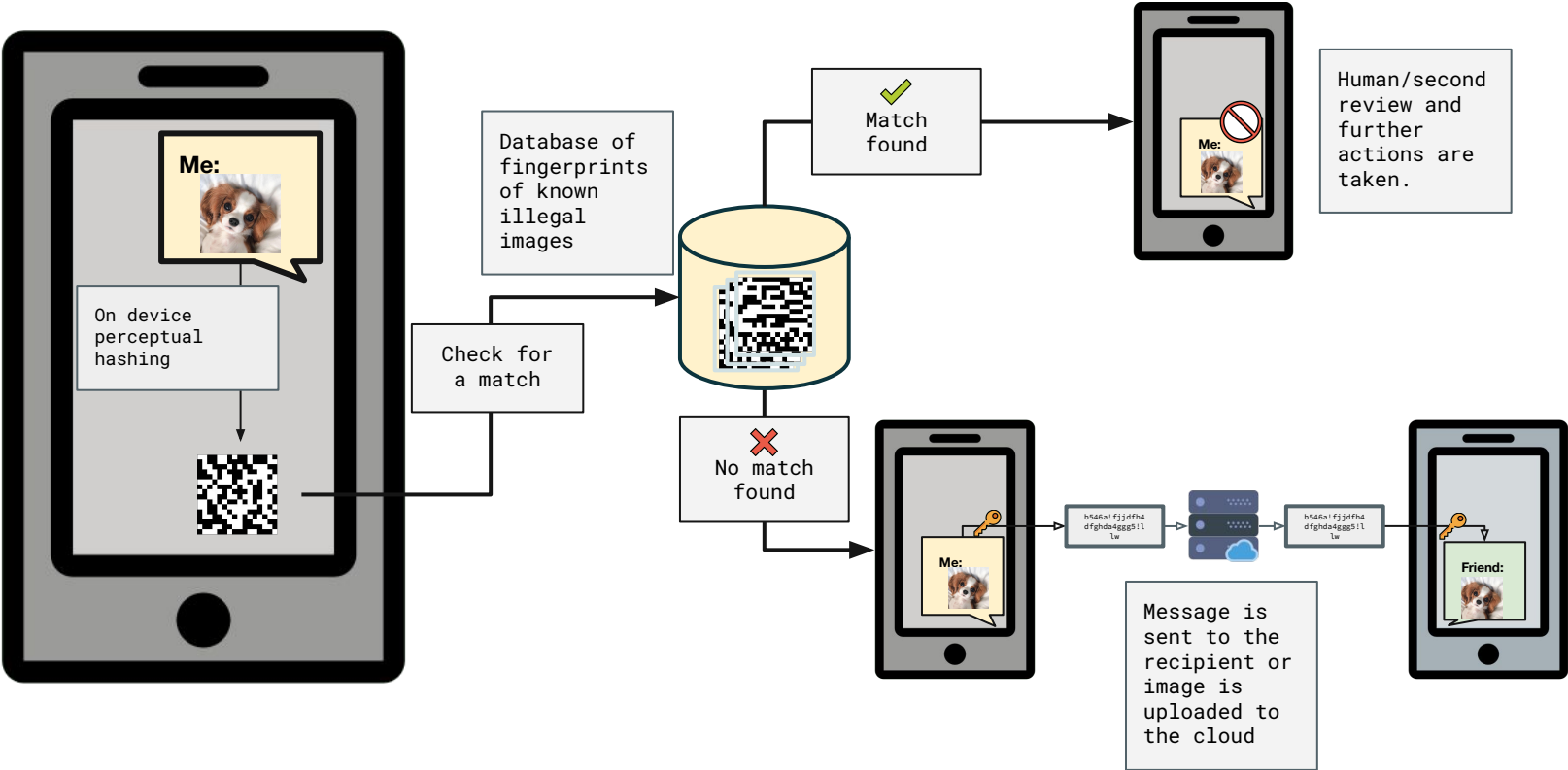
The Way Forward: Working Together to Tackle Cybercrime

Christopher Wray

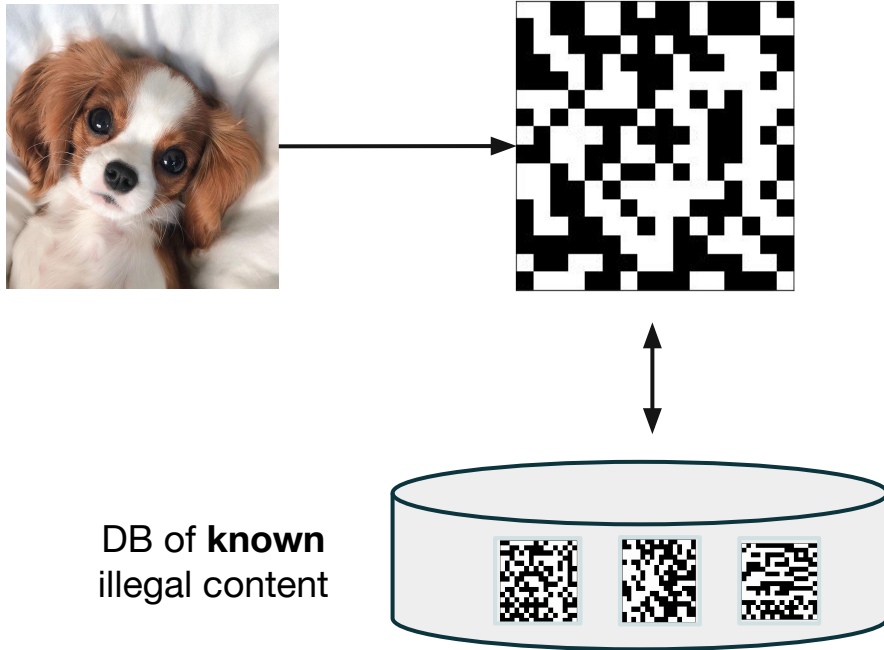
Director
Federal Bureau of Investigation

* E2EE chats not yet used as default option

Perceptual hashing-based client-side scanning (PH-CSS)



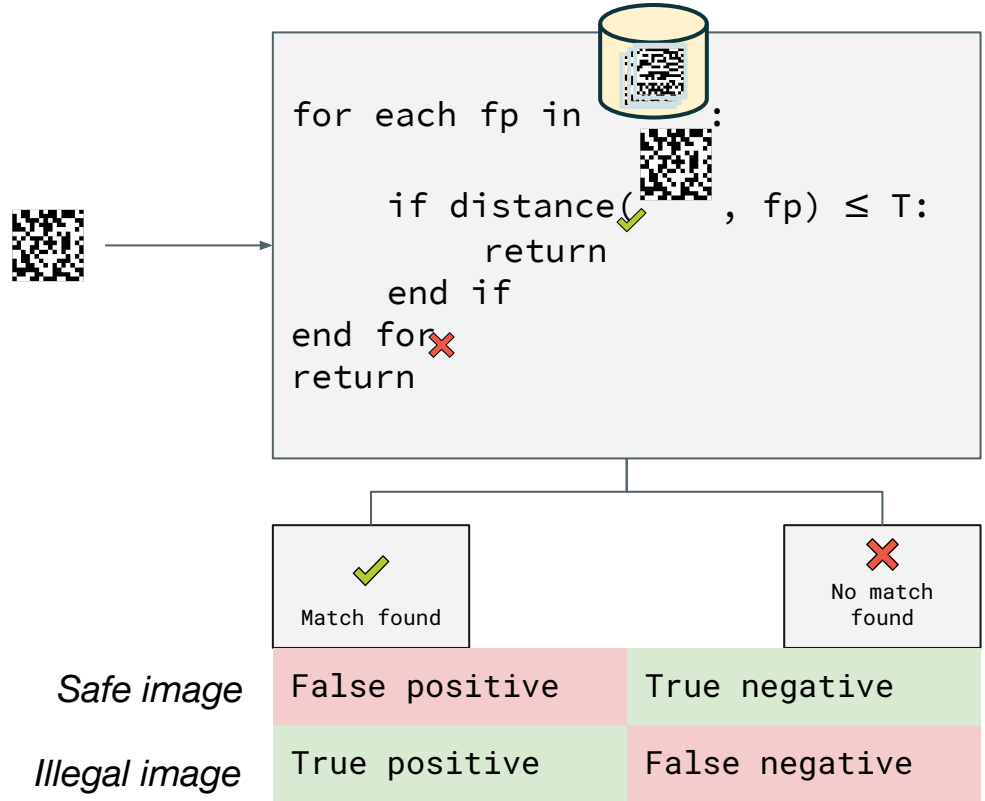
Overview of perceptual hashing



- Standard image fingerprinting technique
- Not a cryptographic hash! On the contrary, perceptual hashing is designed to detect “near-identical” images (resized, cropped, recolored, etc)
- Perceptual hashing algorithms can be manually designed (e.g. pHash but also Microsoft’s PhotoDNA or Facebook PDQ) or learned (e.g., Apple’s neuralmatch)
- They can be distance-based (similar image will be close to one another) or exact (similar images will have the same hash)

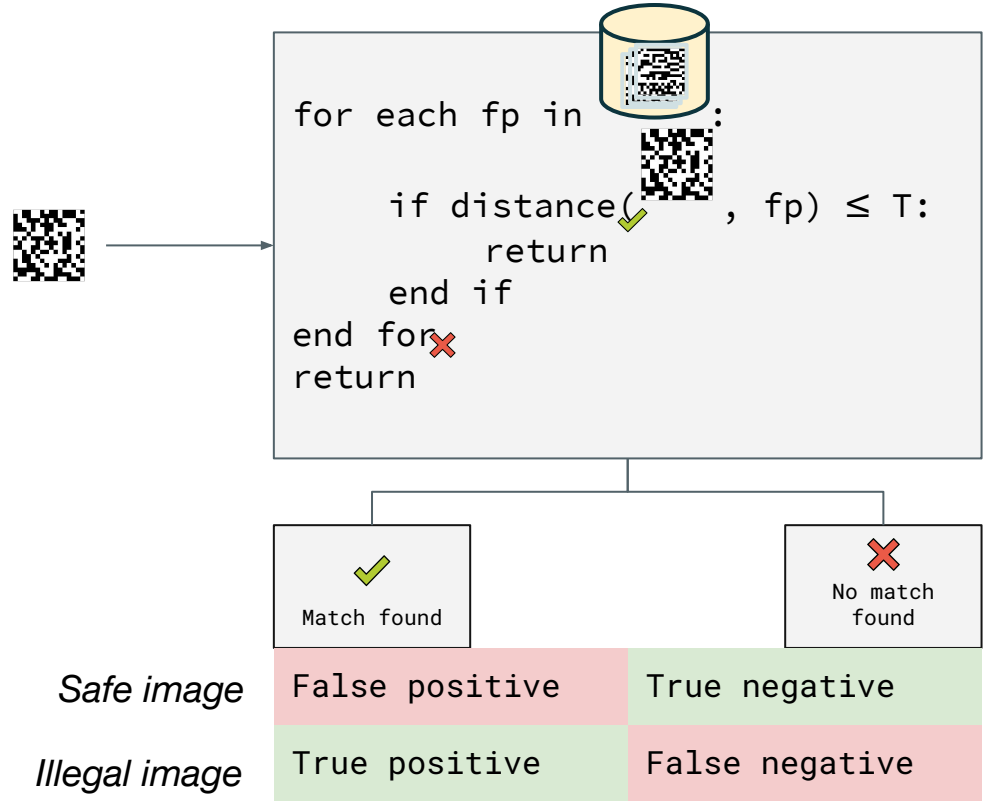
Distance-based matching of fingerprints


- Standard distance metrics used, e.g. Hamming distance
- Predefined threshold “T” is used to define a match



Distance-based matching of fingerprints

- Standard distance metrics used, e.g. Hamming distance
- Predefined threshold “T” is used to define a match
- “T” is chosen to balance trade-off between false positives and false negatives
- $T \uparrow$ leads to $FP \uparrow$ & $FN \downarrow$
- Facebook’s PDQ recommends $20 \leq T \leq 90$



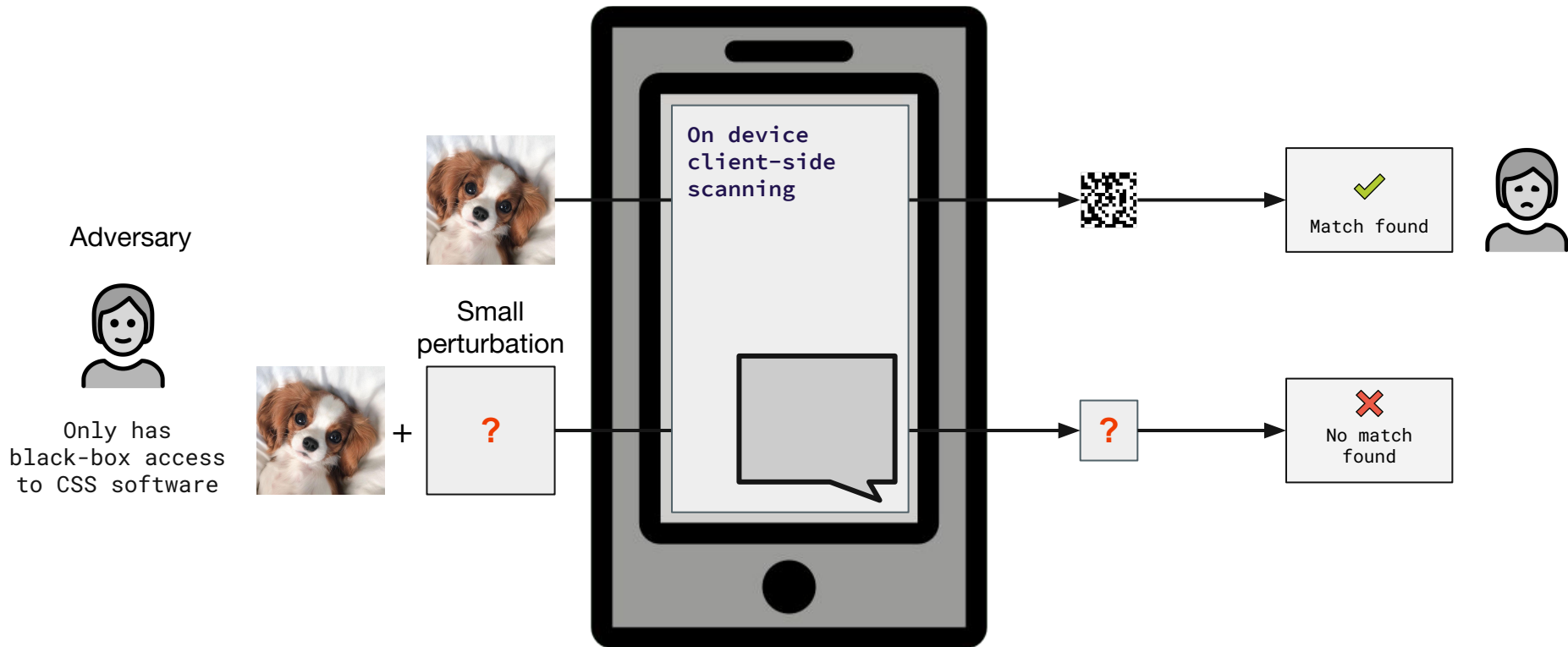


Is client-side scanning robust
solution to black-box adversarial
attacks?

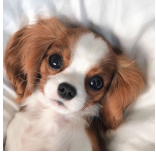
Plan

1. Attack model
2. Attack methodology
3. Results and robustness to countermeasures
4. A white-box algorithm against the Discrete Cosine Transform

Can an adversary evade detection by CSS?



Attack model



X := Image to be attacked
 h := Hashing algorithm (black-box access)
 d := Distance function
 T := Threshold

Find minimum δ such that:

$$d(h(X+\delta), h(X)) > T$$

Attack as an optimization problem

$$f(\delta) := d(h(X+\delta), h(X))$$

Find $\max_{\delta} f(\delta)$

Under constraints of:

1. Visual dissimilarity $\|\delta\|_2 \leq \epsilon$
2. Image should be valid $0 \leq X + \delta \leq 1$

...under black-box assumptions

- The attacker does not have direct access to the gradient
- Natural Evolution Strategies¹ provide a way and were shown to work in adversarial ML²
- Search distribution $p(x; \Theta)$ (we use a Gaussian $\Theta \sim N(\delta, \sigma I)$)
- Estimate the gradient w.r.t. δ of $E_p[f(\delta)] = \int f(\delta)p(\delta; \Theta)d\delta$

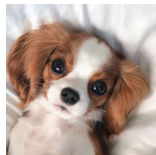
$$E_p[f(\delta)] = \int f(x)p(x; \Theta)dx$$

$$\nabla E_p[f(\delta)] \approx$$

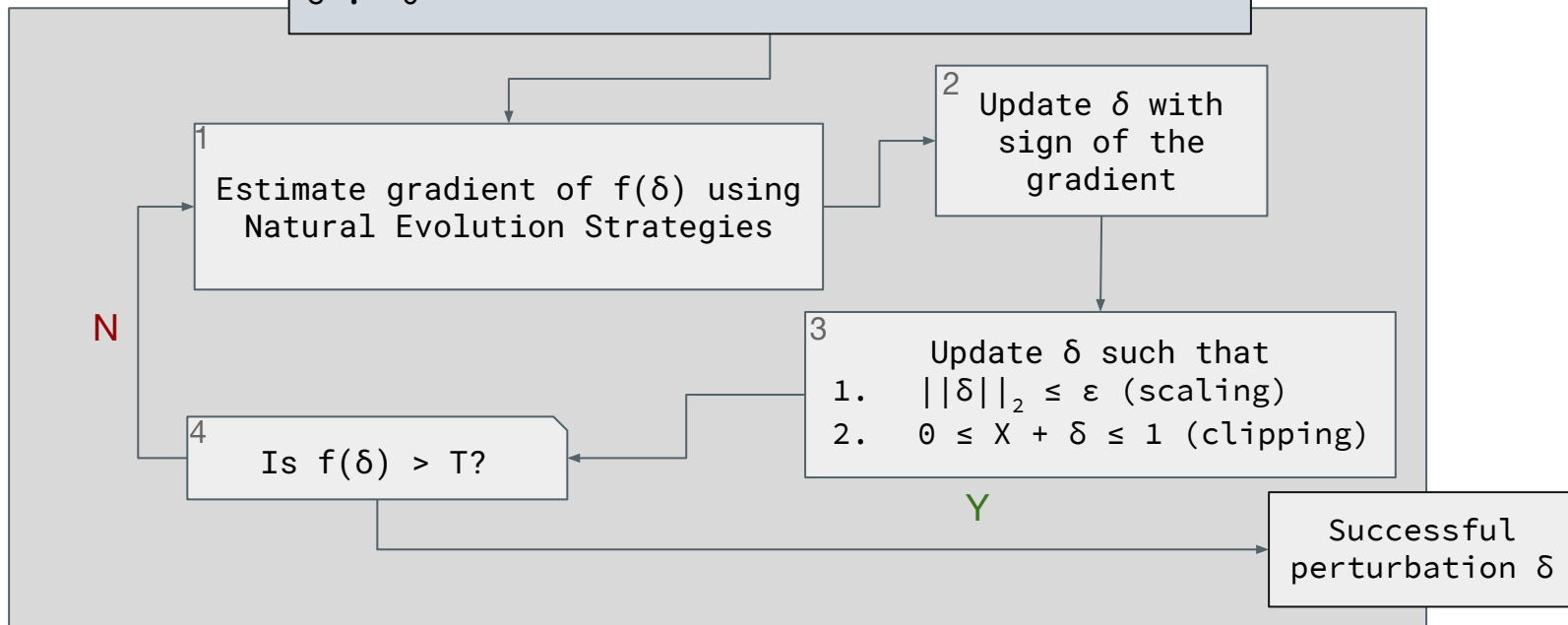
¹ Natural Evolution Strategies. Wierstra et al. JMLR (2014). <https://www.jmlr.org/papers/volume15/wierstra14a/wierstra14a.pdf>

² Black-box Adversarial Attacks with Limited Queries and Information. Ilyas et al. ICML (2018) <http://proceedings.mlr.press/v80/ilyas18a/ilyas18a.pdf>

How the algorithm works



X := Image to be attacked
 h := Hashing algorithm (black-box access)
 d := distance function
 T := Threshold
 δ := 0



>99.9%

Images¹ can be modified successfully using our attack

...for five popular hashing algorithms

...and a wide range of detection thresholds

¹ ImageNet dataset

The modified images are visually similar to the original



Original image



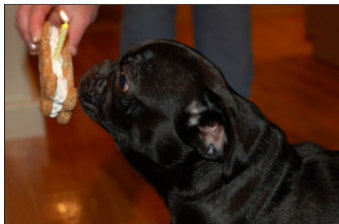
Modified image: PDQ, T=30



Original image



Modified image: PDQ, T=70



Original image



Modified: PDQ, T=30



Modified: PDQ, T=70



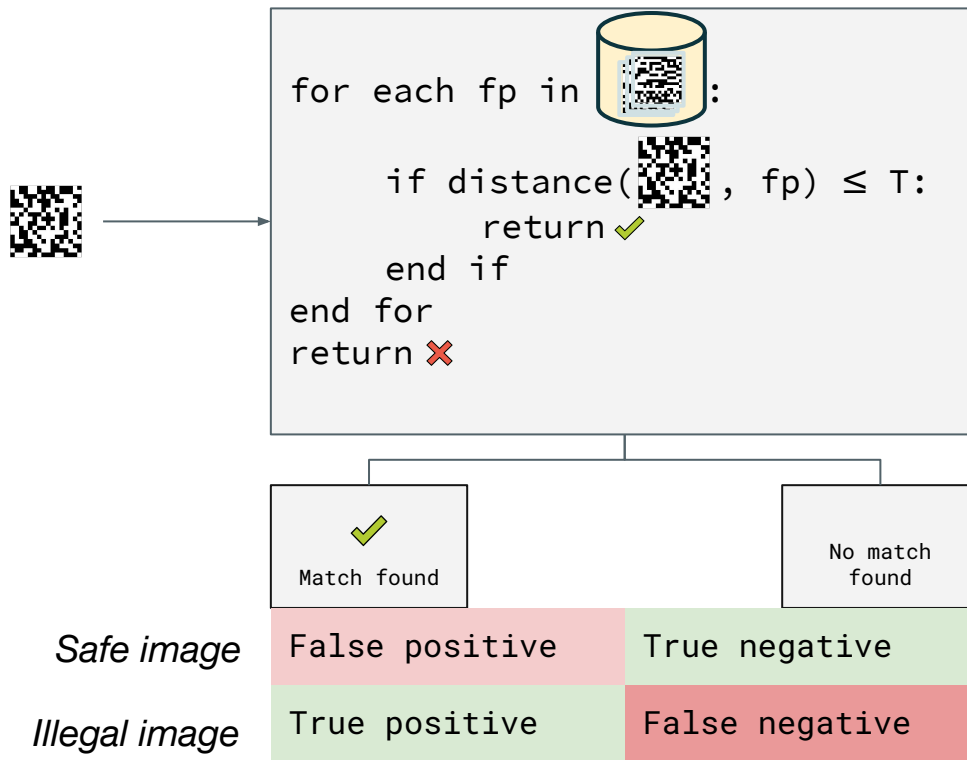
Modified: PDQ, T=85



Modified: PDQ, T=90

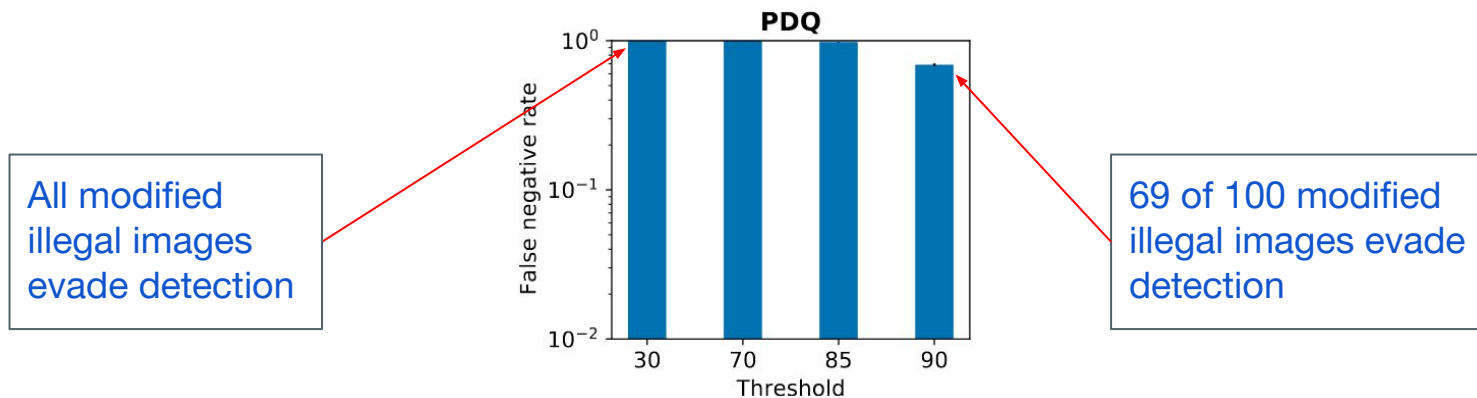
Is pushing the hash beyond T enough to evade detection?

- Even though $d(h(X+\delta), h(X)) > T$, the adversary might not be able to avoid detection
- Because the modified image $X+\delta$ could be close to some other image in the database



Is pushing the hash beyond T enough to evade detection?

False negative rate: Fraction of modified illegal image that evade detection.



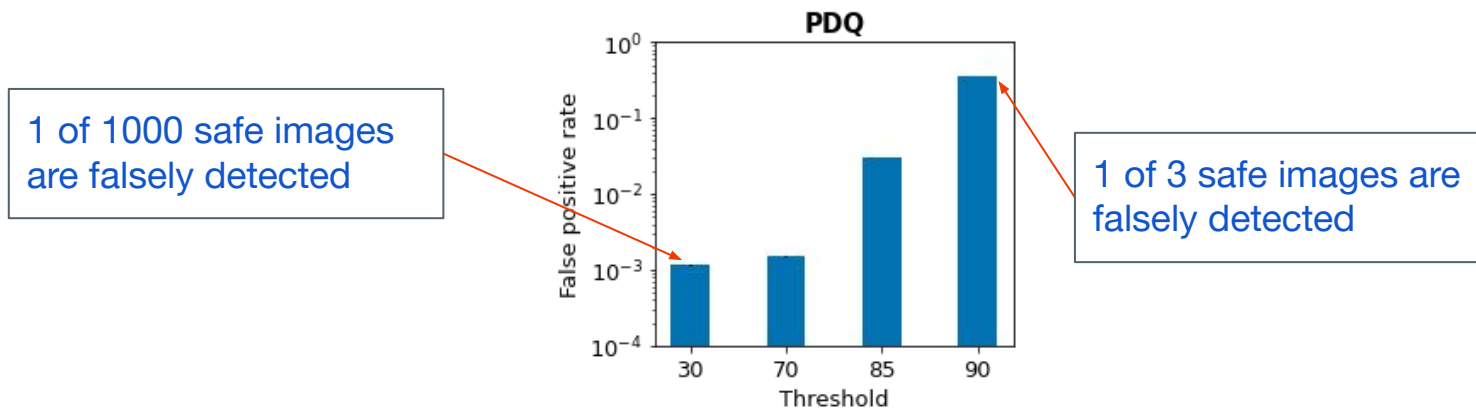
All modified
illegal images
evade detection

69 of 100 modified
illegal images evade
detection

- Experimental setup
- Dataset: ImageNet
 - Database size: 100,000

Is increasing the threshold an effective defense?

False positive rate: Fraction of safe images that are falsely detected.



- Experimental setup
- Dataset: ImageNet
 - Database size: 100,000

A variety of perturbations is possible

- A potential countermeasure against our attack could be to expand the database with modified images
- We adapt our attack to produce diverse perturbations

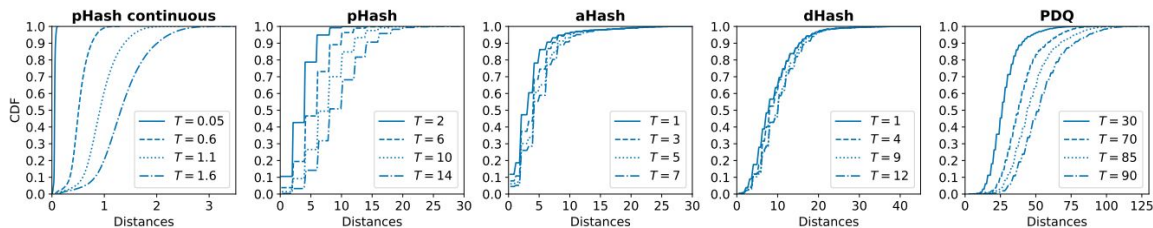
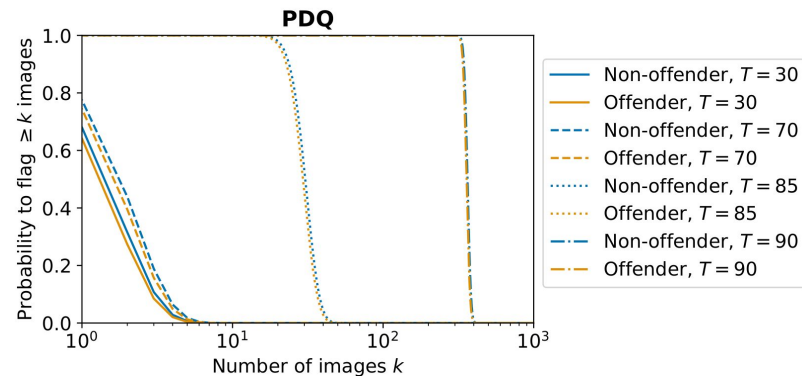


Figure 4: Pairwise distances between hashes of the multiple modified images generated for the same image, for different algorithms and thresholds. $D = 50$ modified images are generated for each ($N' = 100$) original image.

Flagging more images before deciding a match

- The CSS system could flag users only after the number of matches exceeds a predefined threshold k
- We model two types of users each sending 1000 images:
 - an offender sending 100 illegal images
 - a non-offender sending no illegal images
- Offenders and non-offenders are similarly likely to have at least k of their images flagged
- Flagging a user with at least k matches does not seem to be a trivial countermeasure against our attack

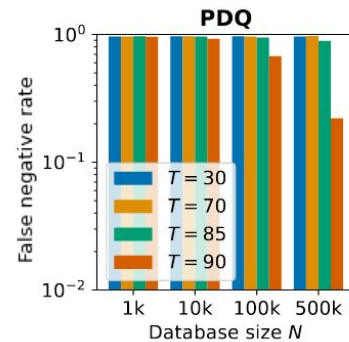
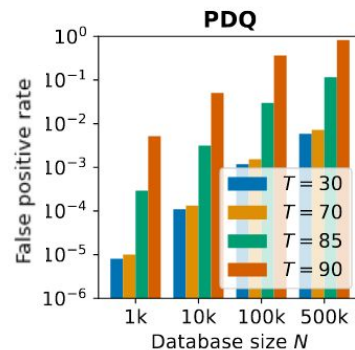


Experimental setup

- Dataset: ImageNet
- Database size: 100,000

Sensitivity to database sizes

- Results shown so far used a database size of 100,000
- How do FPR and FNR vary with the database size?

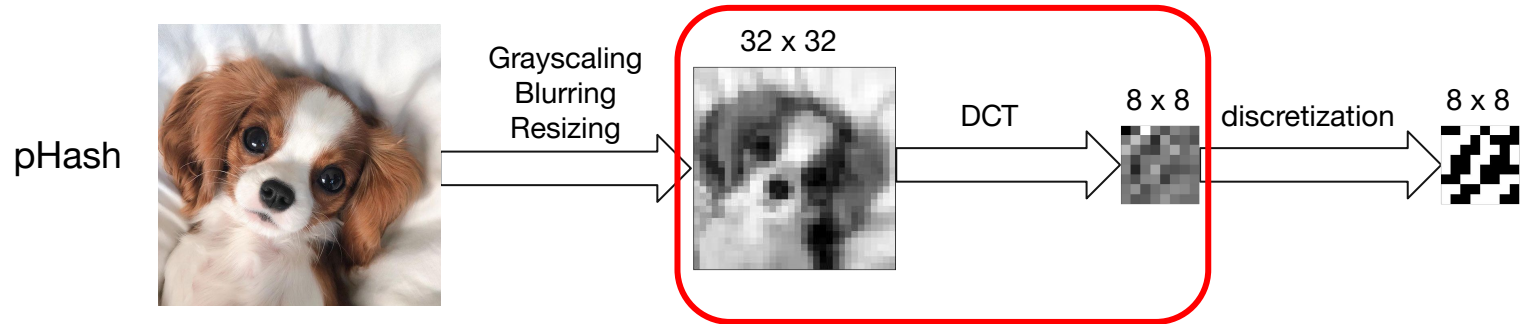


Why does the attack work?

- Perceptual hashes change gradually as the image changes
- A large number of hashes are T away from a given hash, potentially giving rise to multiple perturbations
- Mitigations like \uparrow database size $\Rightarrow \uparrow$ FPR

White-box attack against pHash

- The Discrete Cosine Transform (DCT)¹ is a popular image compression algorithm
- pHash² and Facebook's PDQ are popular DCT-based algorithms



¹ Ahmed, N. et al. Discrete Cosine Transform. IEEE Transactions on Computers (1974). https://www.ic.tu-berlin.de/fileadmin/fg121/Source-Coding_WS12/selected-readings/Ahmed_et_al._1974.pdf

² <https://hackerfactor.com/blog/index.php%3Farchives/432-Looks-Like-It.html>

Optimal perturbations for DCT hashes

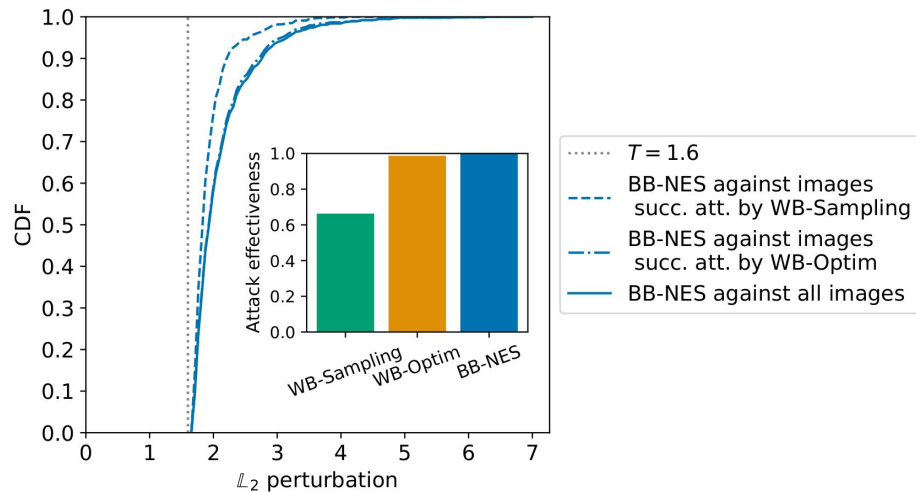
The DCT step can be rewritten as a linear transform $h: \mathbb{R}^{1024} \rightarrow \mathbb{R}^{64}$, $h(X) = AX$

$$\|h(X+\delta) - h(X)\|_2^2 = \|A \delta\|_2^2 \leq \|\delta\|_2^2$$

$T \leq$ output
perturbation
(8x8 image) input
perturbation
(32x32 image)

Optimality of black-box perturbations

- We ran the black-box and the white-box approaches against the DCT algorithm
- When they succeed, the two white-box approaches yield optimal perturbation (i.e., $\|X+\delta\|_2 = T$)
- The black-box approach is close to optimal and more flexible



Conclusion

- Perceptual Hashing-based Client-Side Scanning (PH-CSS) is proposed as a privacy-preserving solution to detect illegal content
- Apple recently announced such a mechanism to be deployed on iOS and MacOSx
- We show here that PH-CSS might not be a robust solution as an image can almost always be modified to avoid detection in black-box setup
- We also show how simple fixes such as increasing DB size (diversity), or increasing the threshold do not help



Thank you!