

Experimental Analyses of the Surveillance Risks in Client-Side Content Scanning

Ashish Hooda

Andrey Labunets

Tadayoshi Kohno

Earlence Fernandes



UC San Diego

W
UNIVERSITY of WASHINGTON

Child Sexual Abuse Material (CSAM) is a Growing Problem

Circulation of child sexual abuse material rampant on Telegram

Sunitha Krishnan lodge complaint with DGP Anjani Kumar

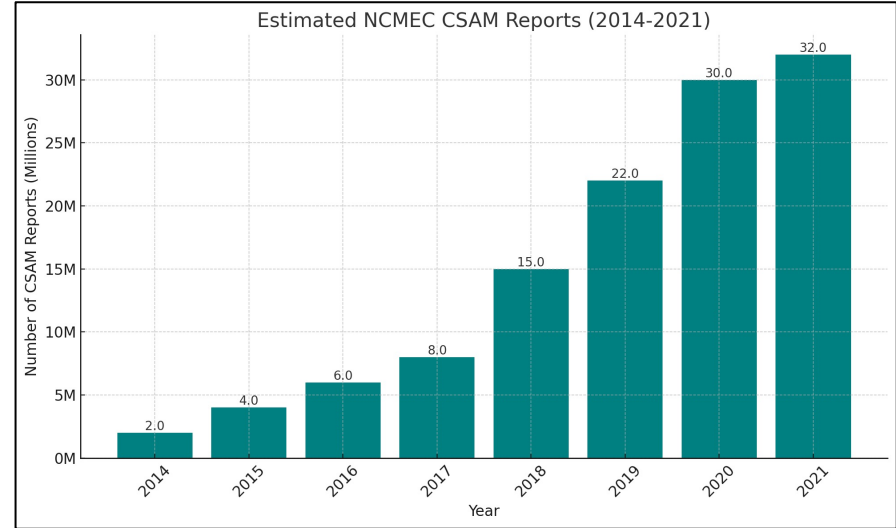
November 16, 2023 09:06 am | Updated 05:43 pm IST - HYDERABAD



December 19, 2022

FBI and Partners Issue National Public Safety Alert on Financial Sextortion Schemes

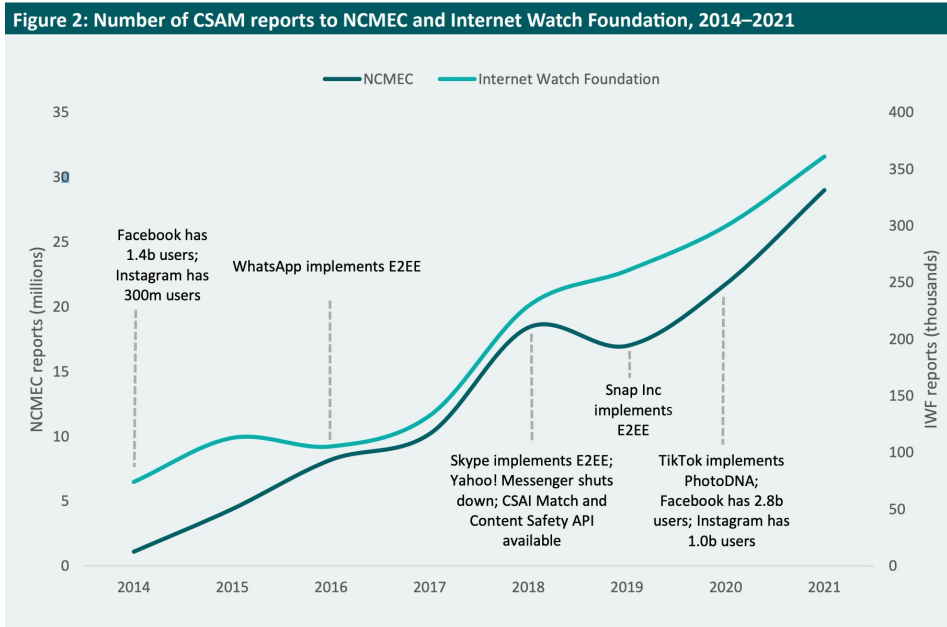
Over 3,000 minor victims targeted in the past year across the United States



Security vs. Privacy Tradeoff for CSAM Detection

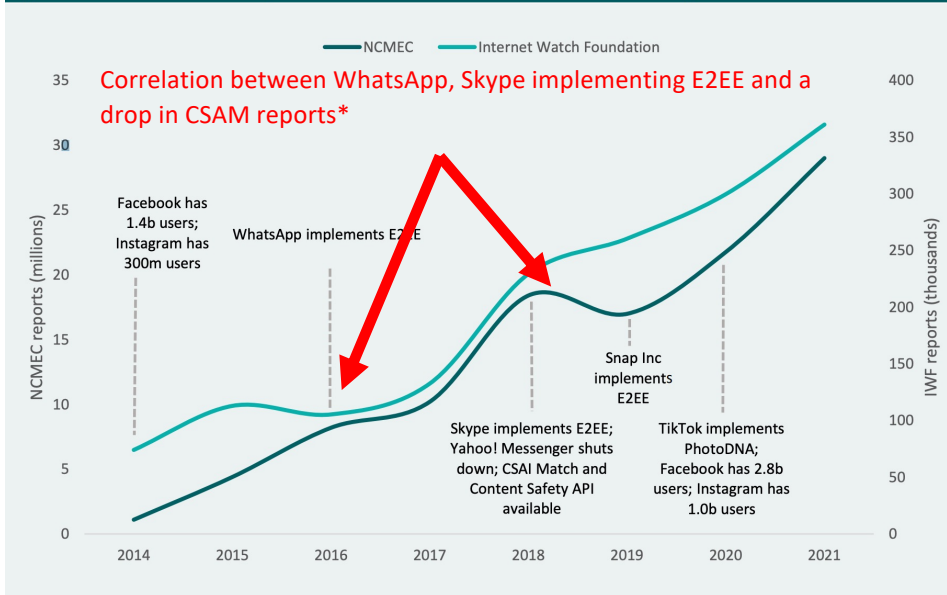
Security vs. Privacy Tradeoff for CSAM Detection

Figure 2: Number of CSAM reports to NCMEC and Internet Watch Foundation, 2014–2021



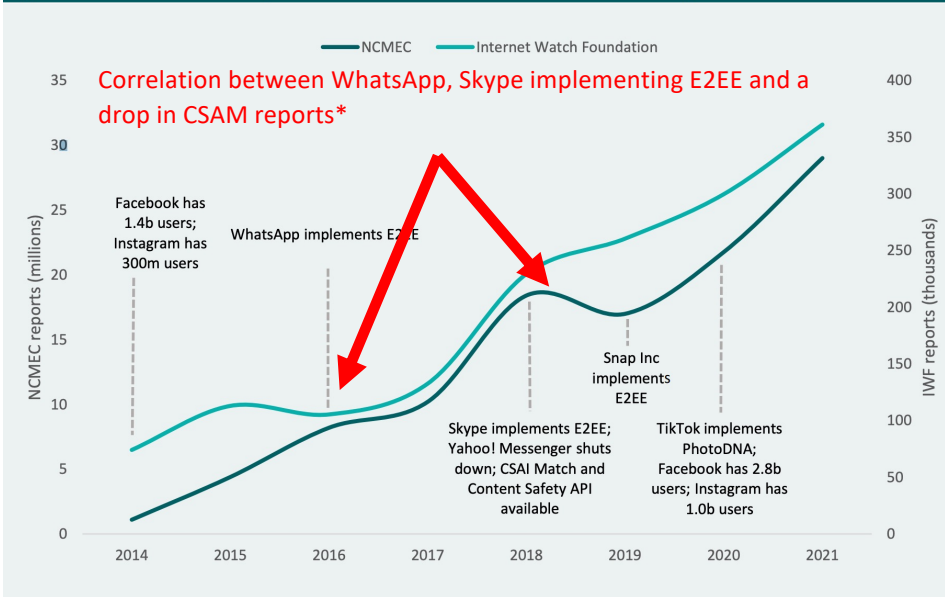
Security vs. Privacy Tradeoff for CSAM Detection

Figure 2: Number of CSAM reports to NCMEC and Internet Watch Foundation, 2014–2021



Security vs. Privacy Tradeoff for CSAM Detection

Figure 2: Number of CSAM reports to NCMEC and Internet Watch Foundation, 2014–2021



SILY HAY NEWMAN MORGAN HEAVER MATT BURDESS SECURITY MAY 22, 2023 3:23 PM

Leaked Government Document Shows Spain Wants to Ban End-to-End Encryption

In response to an EU proposal to scan private messages for illegal material, the country's officials said it is "imperative that we have access to the data."

Meta targeted for fresh UK gov't warning against E2E encryption for Messenger, Instagram

Home Secretary demands unspecified 'safety measures' -- warning Ofcom has extensive new powers under Online Safety Bill

[Home](#) / [News](#) / [2023](#) / [May](#) / European Commission: "the content is the crime," so let's break encryption

European Commission: "the content is the crime," so let's break encryption

Middle Ground: Client-Side Scanning?

Middle Ground: Client-Side Scanning?

Apple proposed Client-Side Scanning for CSAM detection (Aug. 2021)

[Home](#) / [Tech](#) / [Security](#)

Apple is bringing client-side scanning mainstream and the genie is out of the bottle

Middle Ground: Client-Side Scanning?

Apple proposed Client-Side Scanning for CSAM detection (Aug. 2021)

Home / Tech / Security

Apple is bringing client-side scanning mainstream and the genie is out of the bottle

Program scraped by Apple due to potential privacy risks (Dec. 2022)

ANDY GREENBERG SECURITY AUG 5, 2021 5:03 PM

Apple Walks a Privacy Tightrope to Spot Child Abuse in iCloud

With a new capability to search for illegal material not just in the cloud but on user devices, the company may have opened up a new front in the encryption wars.

LILY HAY NEWMAN SECURITY DEC 7, 2022 1:11 PM

Apple Kills Its Plan to Scan Your Photos for CSAM. Here's What's Next

The company plans to expand its Communication Safety features, which aim to disrupt the sharing of child sexual abuse material at the source.

Middle Ground: Client-Side Scanning?

Apple proposed Client-Side Scanning for CSAM detection (Aug. 2021)

Home / Tech / Security

Apple is bringing client-side scanning mainstream and the genie is out of the bottle

Program scrapped by Apple due to potential privacy risks (Dec. 2022)

ANDY GREENBERG SECURITY AUG 5, 2021 5:03 PM

Apple Walks a Privacy Tightrope to Spot Child Abuse in iCloud

With a new capability to search for illegal material not just in the cloud but on user devices, the company may have opened up a new front in the encryption wars.

LILY HAY NEWMAN SECURITY DEC 7, 2022 1:11 PM

Apple Kills Its Plan to Scan Your Photos for CSAM. Here's What's Next

The company plans to expand its Communication Safety features, which aim to disrupt the sharing of child sexual abuse material at the source.

Renewed Interest (Oct. 2023)

Undermining Democracy: The European Commission's Controversial Push for Digital Surveillance

Danny Mekić · October 13, 2023

UK amends encrypted message scanning plans

19 July 2023

Meta targeted for fresh UK gov't warning against E2E encryption for Messenger, Instagram

Home Secretary demands unspecified 'safety measures' -- warning Ofcom has extensive new powers under Online Safety Bill

Natasha Lomas @riptari / 1:39 PM CDT · September 20, 2023

 Comment

Physical Surveillance Attack Model

We contribute to the understanding of surveillance risks of Client-Side Scanning

- Introduce a new type of attack for physical surveillance
- Perform extensive evaluation using real-world experiments

Threat Model

- Attacker is a Government entity that wants to surveil a physical location
 - A room in a hotel, a different country's embassy, etc.
 - Doesn't want to/Cannot place a camera

Threat Model

- Attacker is a Government entity that wants to surveil a physical location
 - A room in a hotel, a different country's embassy, etc.
 - Doesn't want to/Cannot place a camera
- Attacker has access to a small number of photos of the target location either through physical access or the internet

Threat Model

- Attacker is a Government entity that wants to surveil a physical location
 - A room in a hotel, a different country's embassy, etc.
 - Doesn't want to/Cannot place a camera
- Attacker has access to a small number of photos of the target location either through physical access or the internet
- Attacker wants access to photos/selfies taken by users at this location

Client-Side Content Scanning



User



Cloud Service
Provider



Illegal
Hash DB

Curated by
NCMEC / IWF

Client-Side Content Scanning



User





Cloud Service
Provider





Curated by
NCMEC / IWF



Illegal
Hash DB

Hash() not in 

Hash() in 

Client-Side
Scanning

Client-Side Content Scanning

Curated by
NCMEC / IWF



Illegal
Hash DB




User



Cloud Service
Provider




Hash(🌄) not in 

Upload Image



Cannot be
decrypted

Hash(😬) in 

Upload Image



Can be
decrypted

Client-Side
Scanning

Client-Side Content Scanning

Curated by
NCMEC / IWF



Illegal
Hash DB





Cloud Service
Provider



User





Hash() not in 

Upload Image



Cannot be
decrypted

Hash() in 

Upload Image



Can be
decrypted



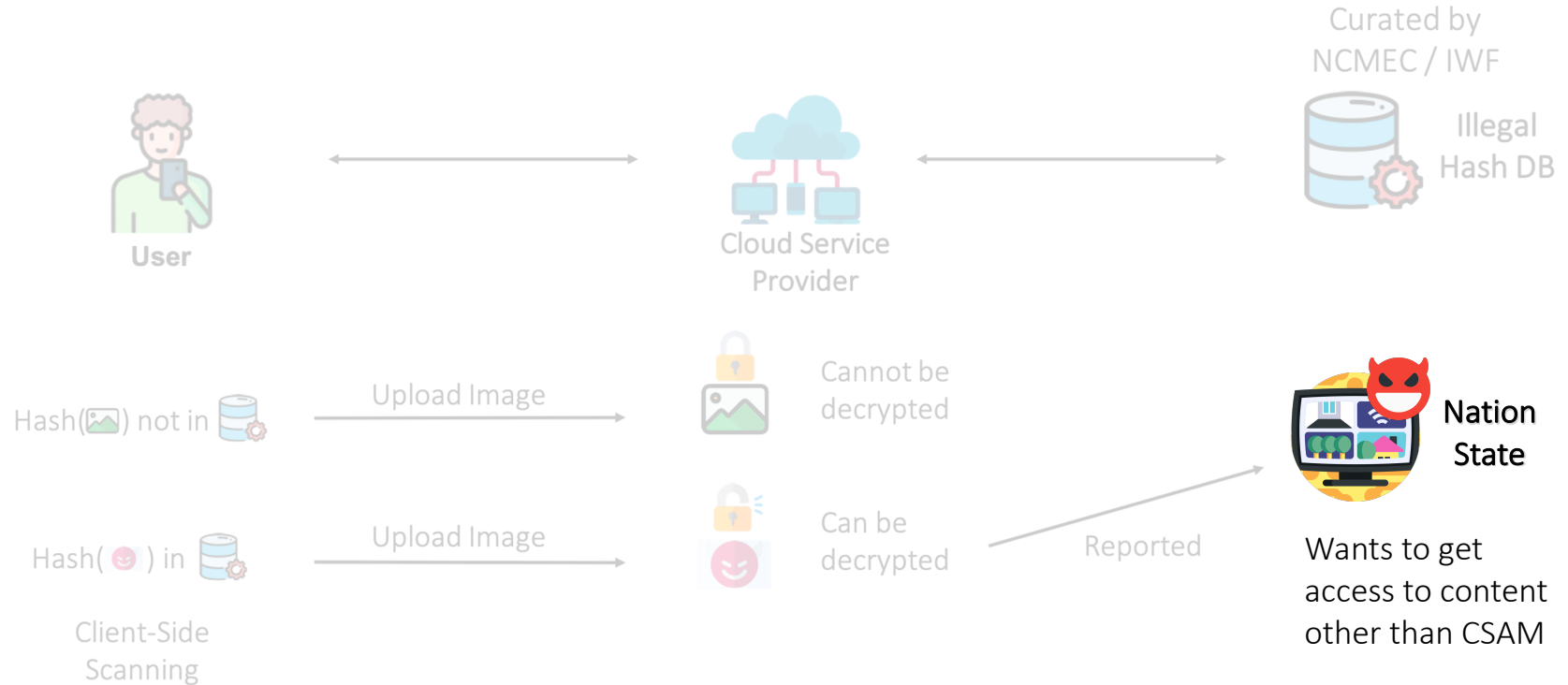
Reported



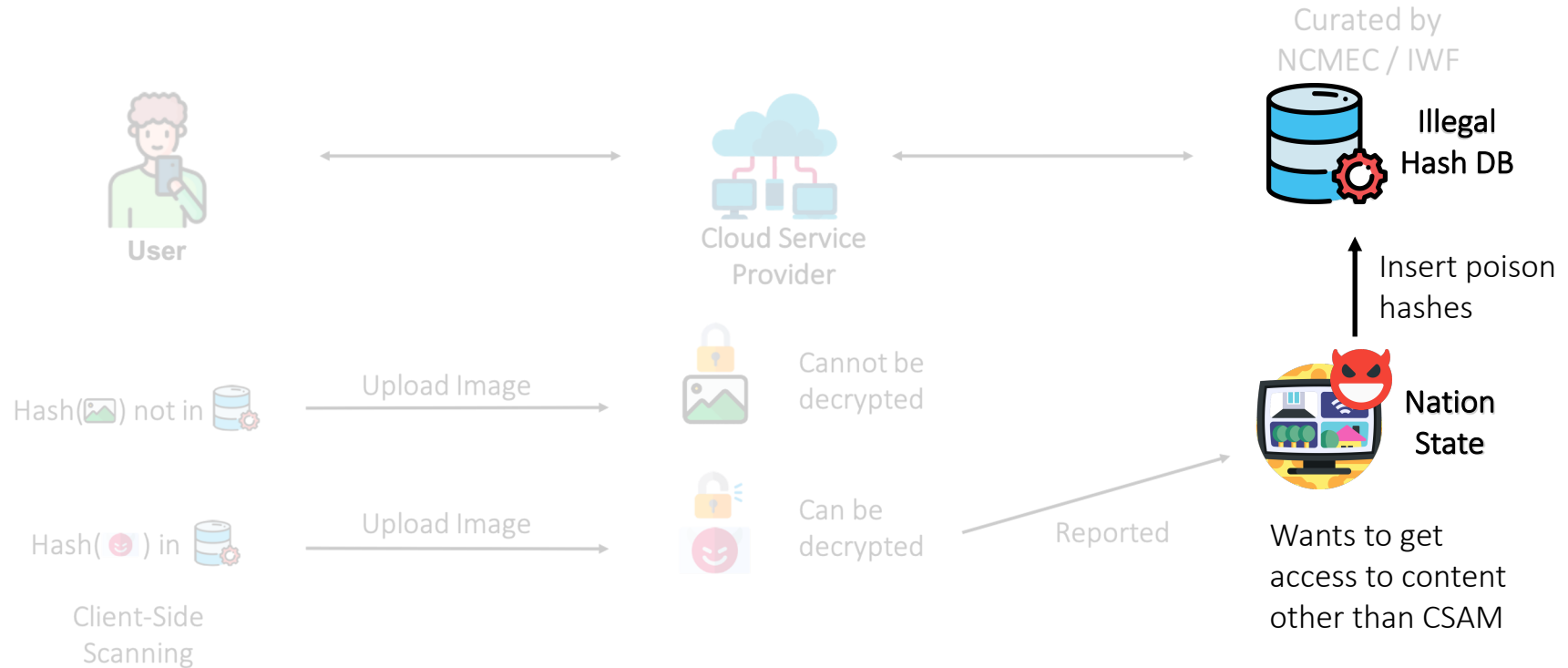
Nation
State

Client-Side
Scanning

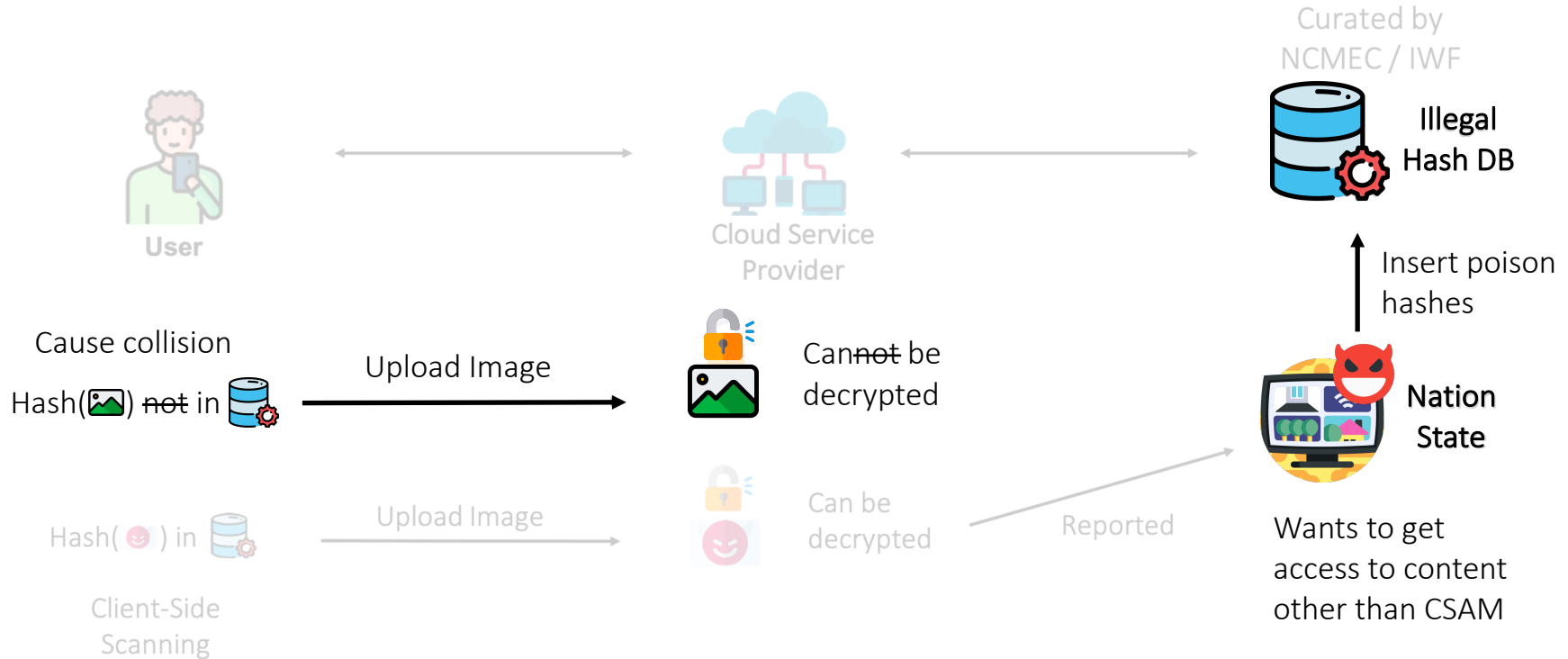
Attacking Client-Side Content Scanning



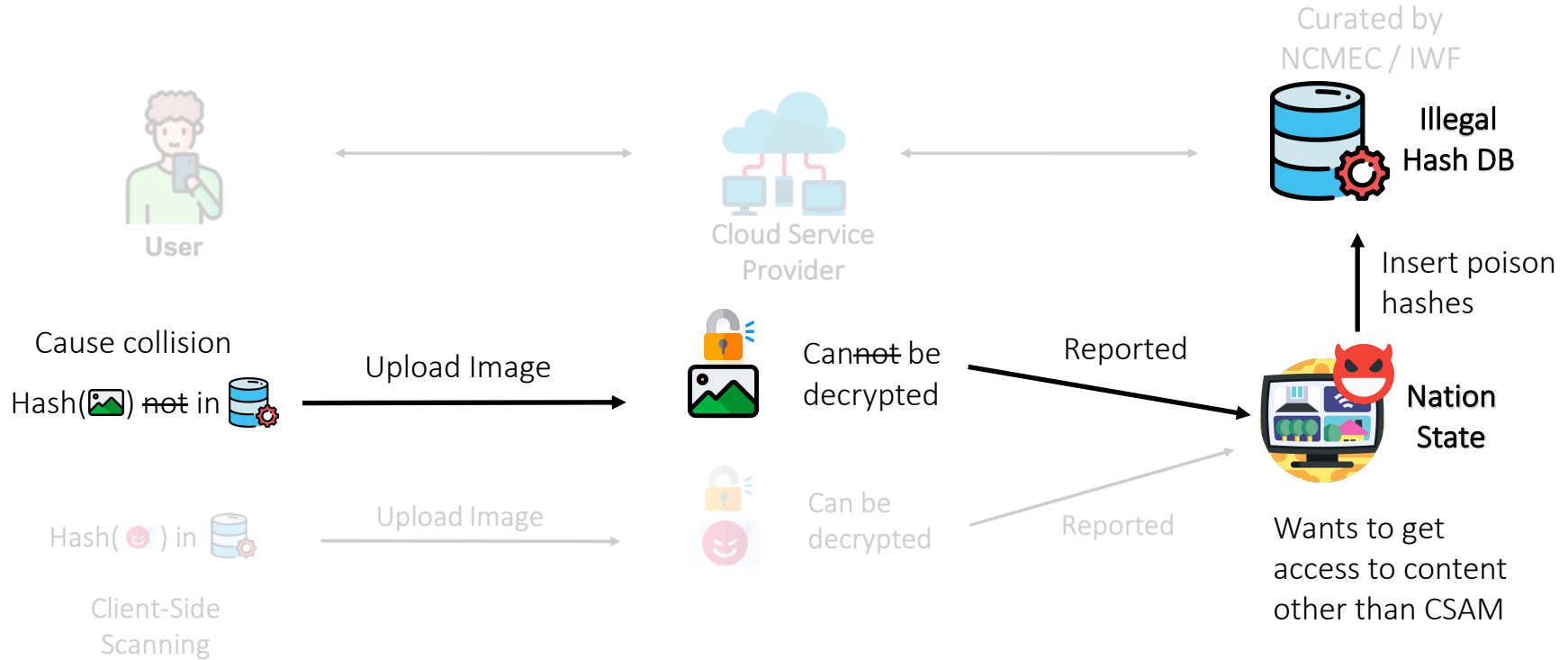
Attacking Client-Side Content Scanning



Attacking Client-Side Content Scanning



Attacking Client-Side Content Scanning



Attack Goals

Objective: Poison the CSAM database such that all user images from the target location get decrypted

Attack Goals

Objective: Poison the CSAM database such that all user images from the target location get decrypted

How to make the attack practical?

1. Minimize the number of poisons that need to be inserted
2. Insert poisons without being detected

Perceptual Hashing

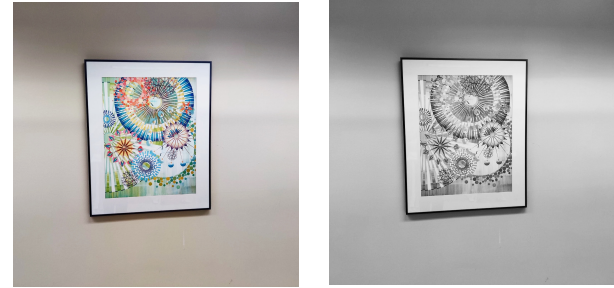
- Locality Sensitive Hashing
- Preserves Image Semantics

Different Perspective



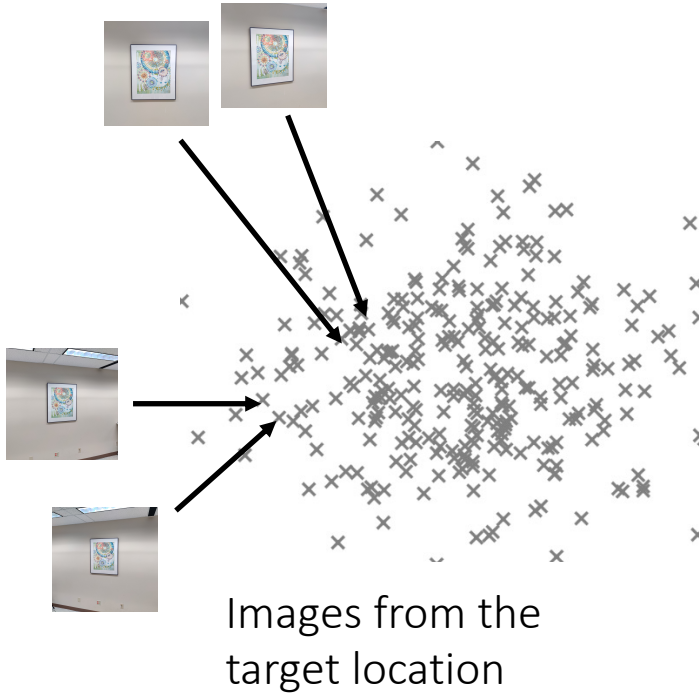
0027908355ce273bdbc48e34
Same Perceptual Hash

Different Colors

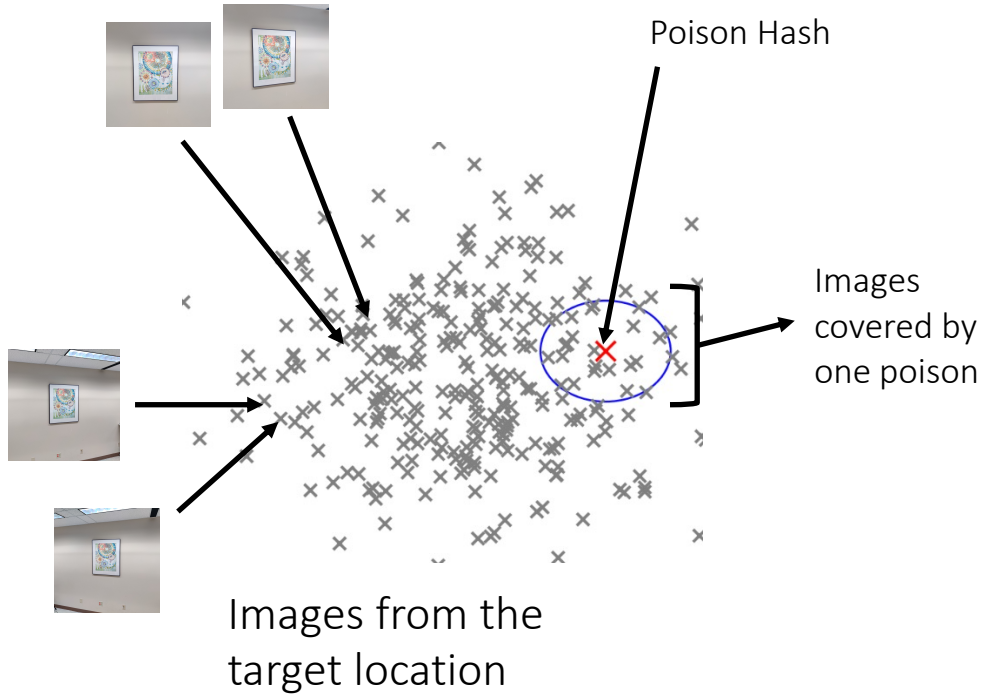


2ec11538306b80f345e128cd
Same Perceptual Hash

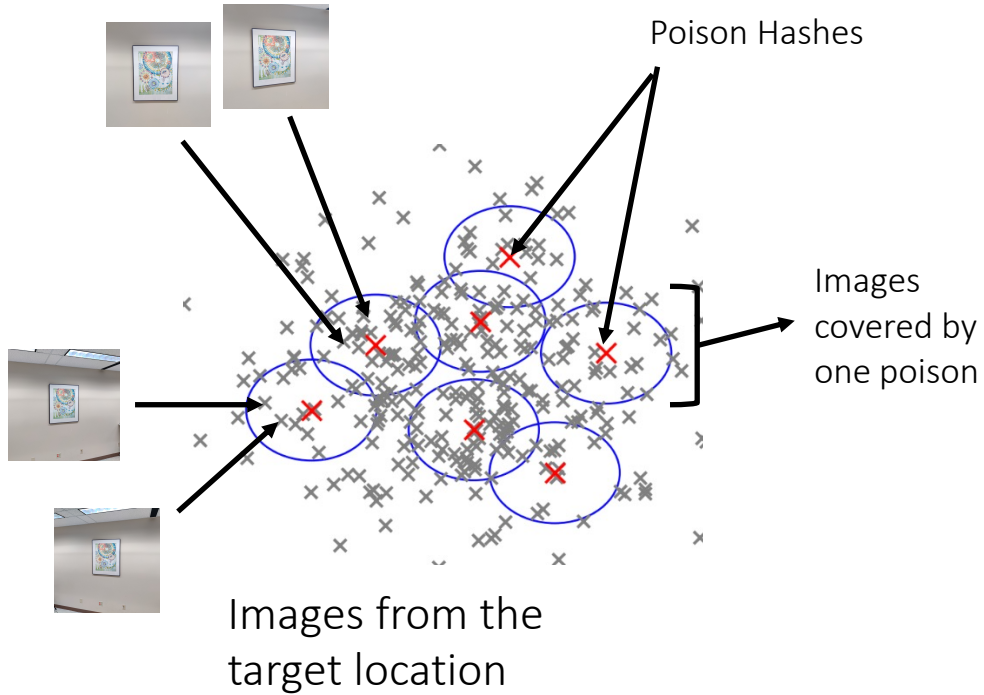
Computing Poison Hashes



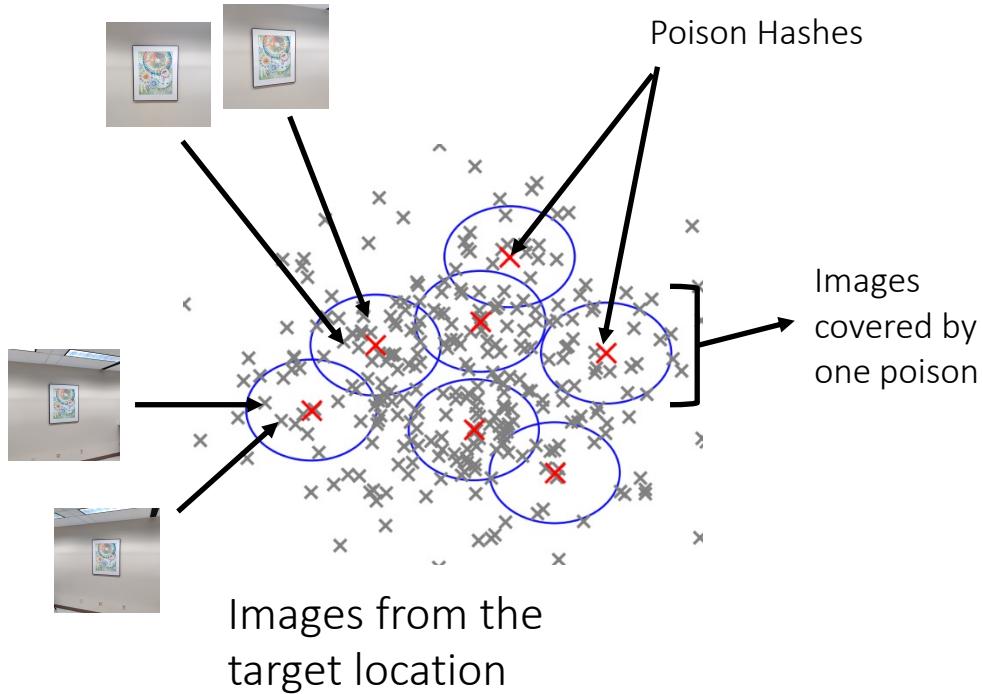
Computing Poison Hashes



Computing Poison Hashes

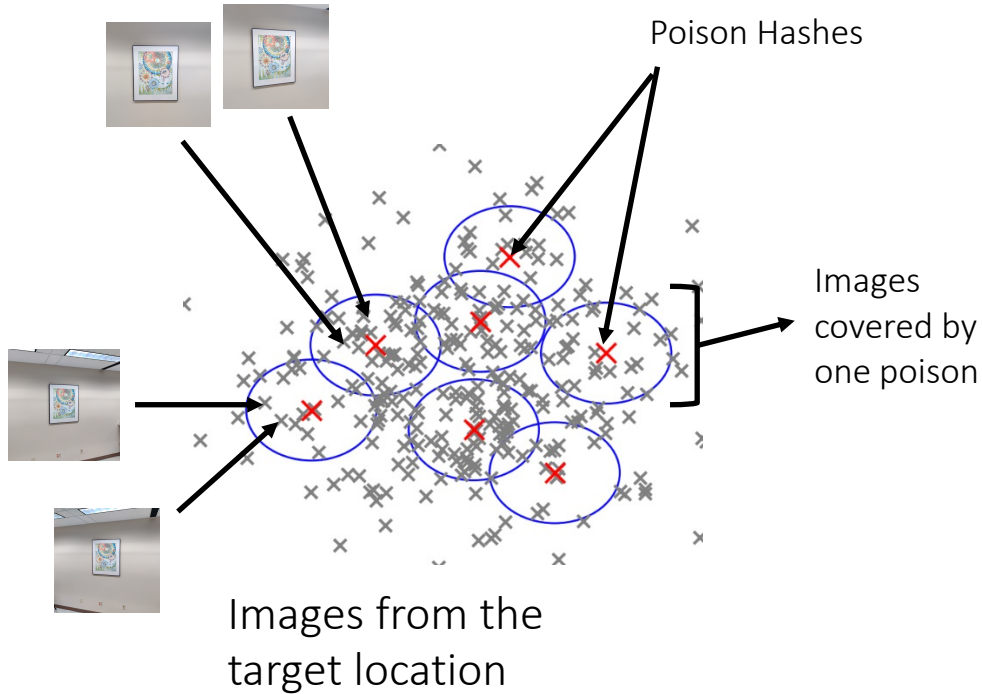


Computing Poison Hashes



Key Insight: Finding the minimum number of poison hashes for physical surveillance is a covering code problem

Computing Poison Hashes



Key Insight: Finding the minimum number of poison hashes for physical surveillance is a covering code problem

Compute an approximate solution of the covering code problem using Clustering, where the cluster centers are the poison hashes to be inserted

Inserting Poison Hashes

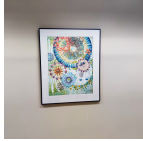


Nation
State



Database
Curator
(NCMEC or
IWF)

Inserting Poison Hashes



Submit



Not a CSAM
image ❌



Nation
State

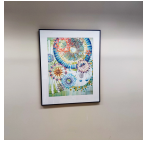


Database
Curator
(NCMEC or
IWF)

Inserting Poison Hashes



Nation
State



Submit



Not a CSAM
image ❌



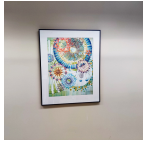
Database
Curator
(NCMEC or
IWF)

Key Insight: Perceptual hash
functions are vulnerable to
collision attacks

Inserting Poison Hashes



Nation
State



Submit

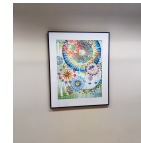


Not a CSAM
image ❌

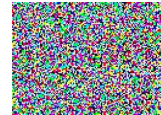


Database
Curator
(NCMEC or
IWF)

Key Insight: Perceptual hash
functions are vulnerable to
collision attacks

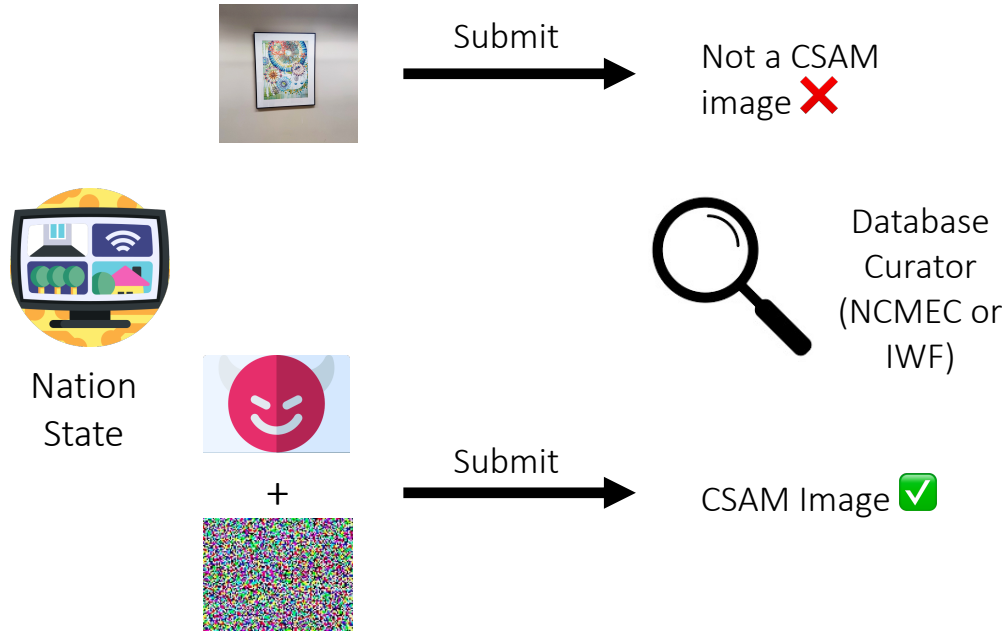


+

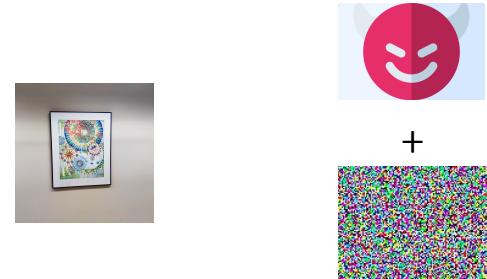


0027908355ce273bdbc48e34
Same Perceptual Hash

Inserting Poison Hashes

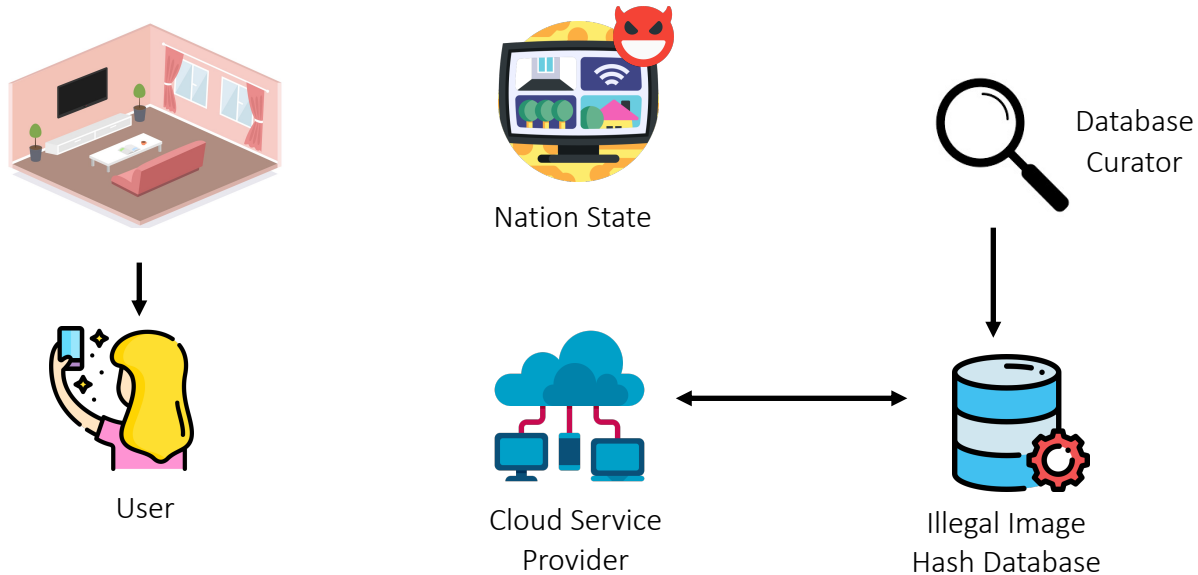


Key Insight: Perceptual hash functions are vulnerable to collision attacks

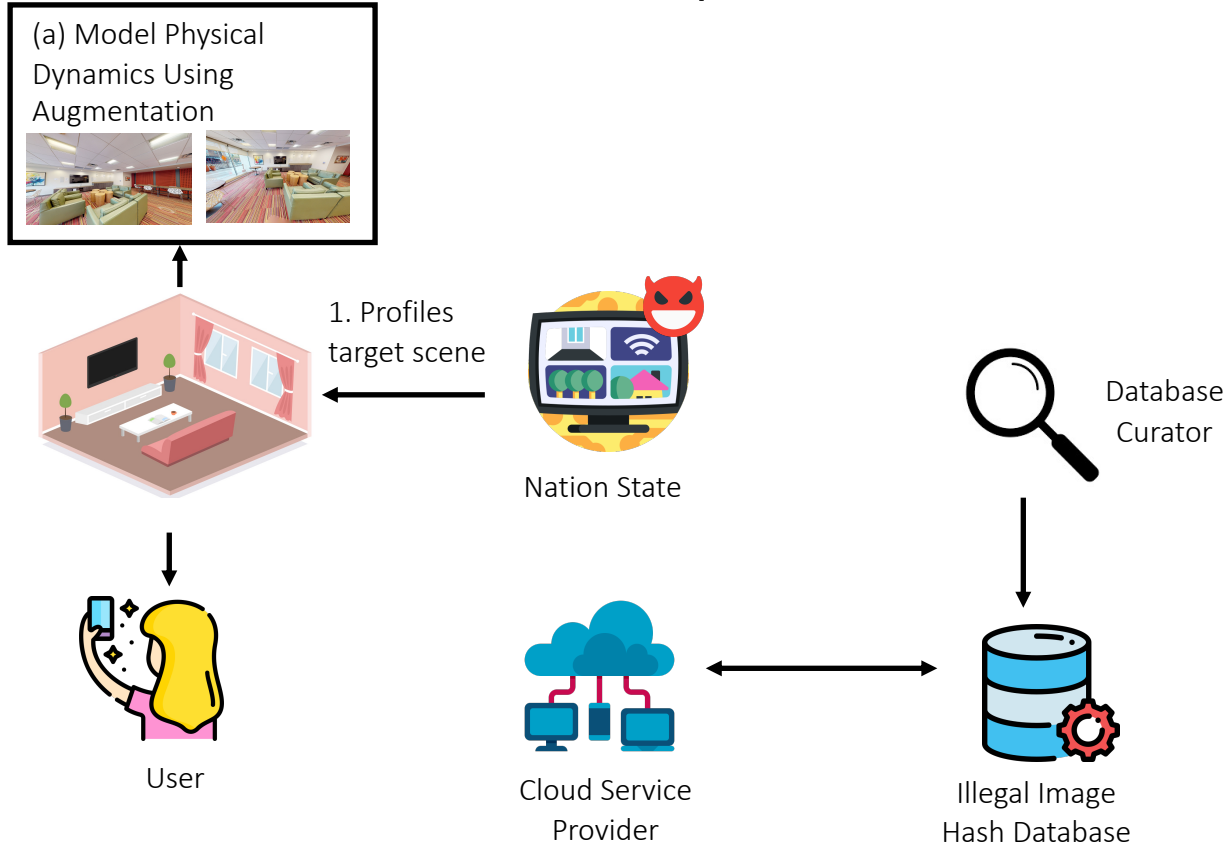


0027908355ce273bdbc48e34
Same Perceptual Hash

Attack Pipeline




Attack Pipeline

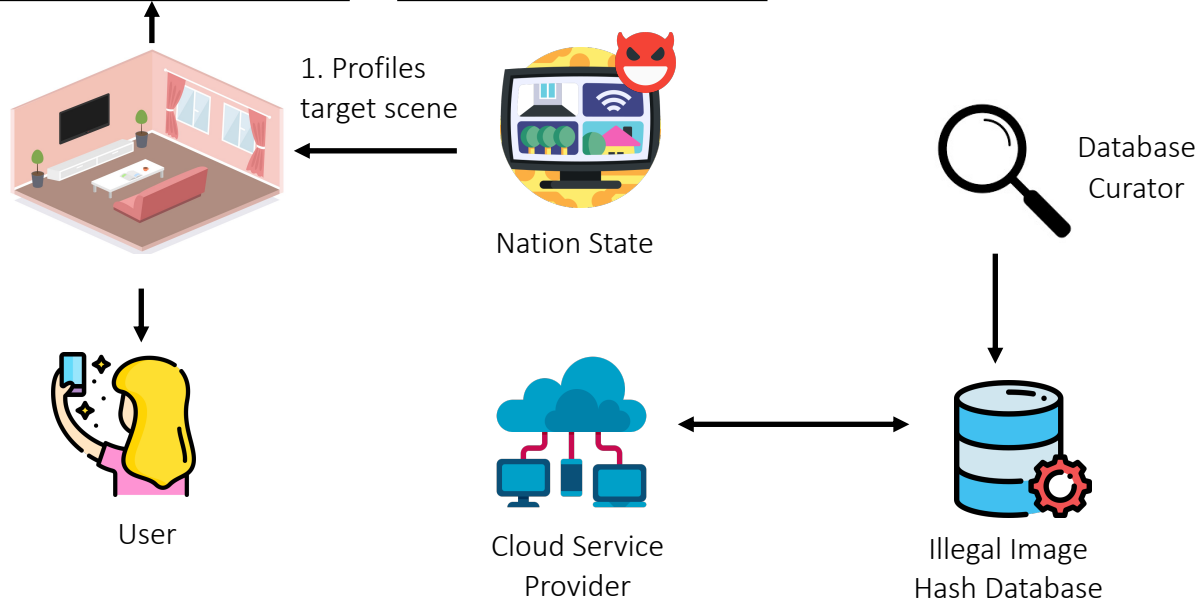
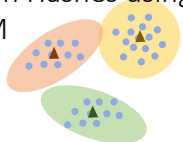


Attack Pipeline

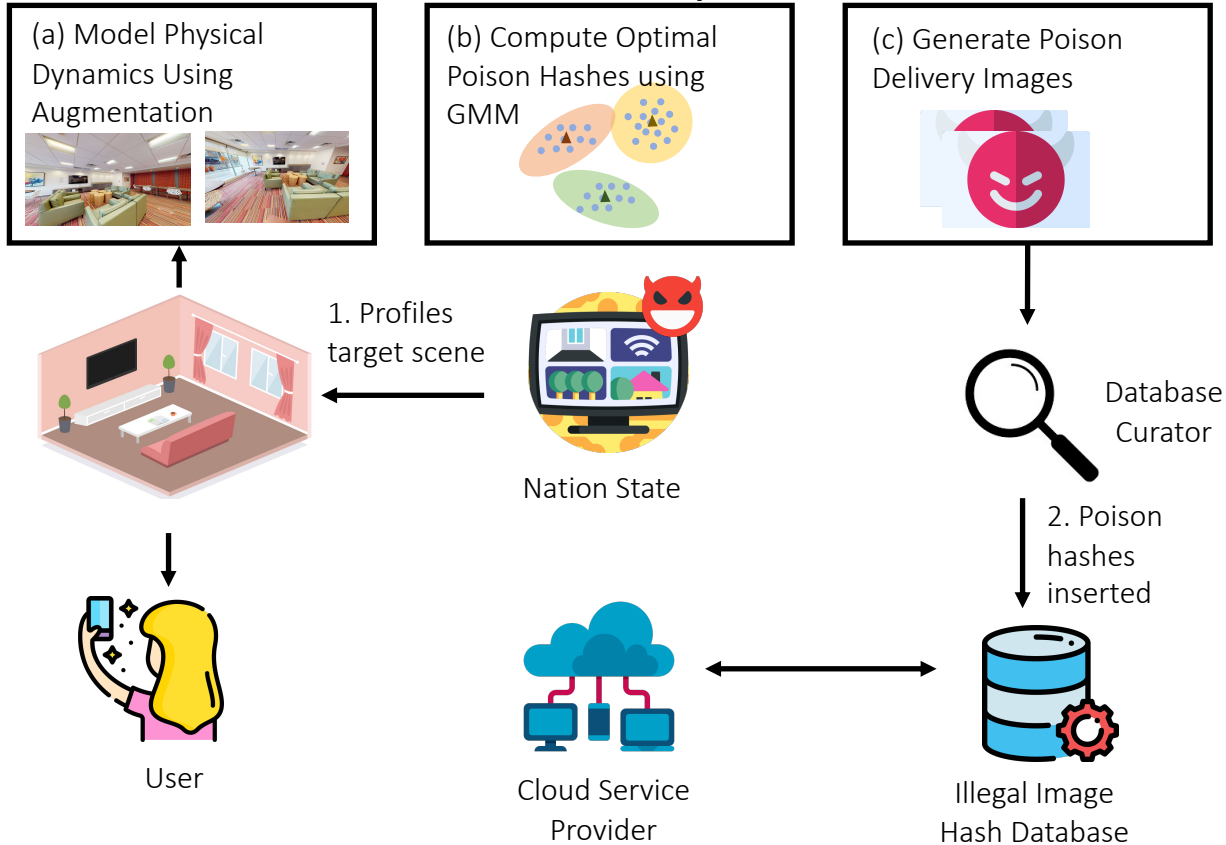
(a) Model Physical Dynamics Using Augmentation



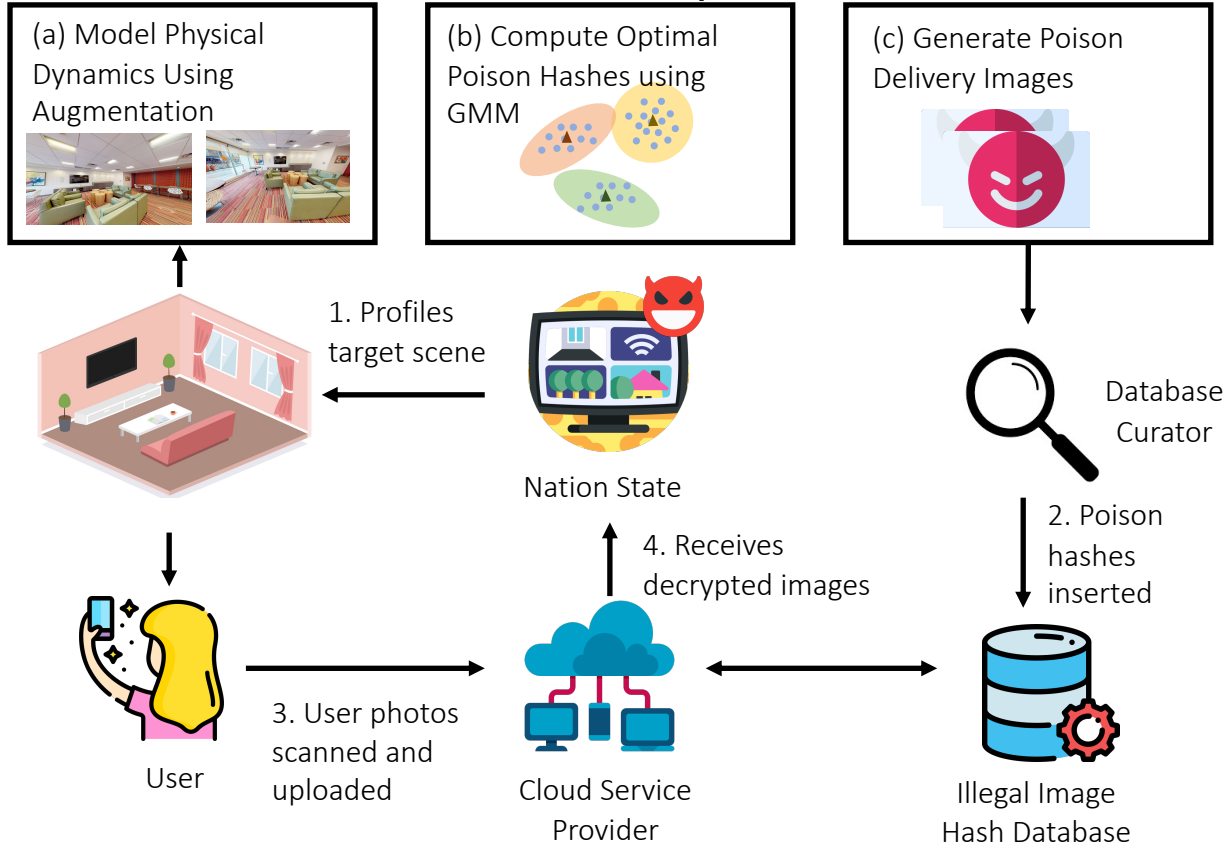
(b) Compute Optimal Poison Hashes using GMM



Attack Pipeline



Attack Pipeline



Evaluation

How effective is Physical Surveillance in Client-Side Scanning Systems?

How does this attack inform on the Security-Privacy tradeoff of CSAM detection?

Evaluation

How effective is Physical Surveillance in Client-Side Scanning Systems?

Get access to around 30% of target location images

How does this attack inform on the Security-Privacy tradeoff of CSAM detection?

Evaluation

How effective is Physical Surveillance in Client-Side Scanning Systems?

Get access to around 30% of target location images

How does this attack inform on the Security-Privacy tradeoff of CSAM detection?

More robust CSAM detection \Rightarrow More severe Physical Surveillance

Evaluation Setup

- We demonstrate our attack on 6 different physical locations
 - Four tourist spots from the Public-Instagram dataset
 - Two room on the university campus

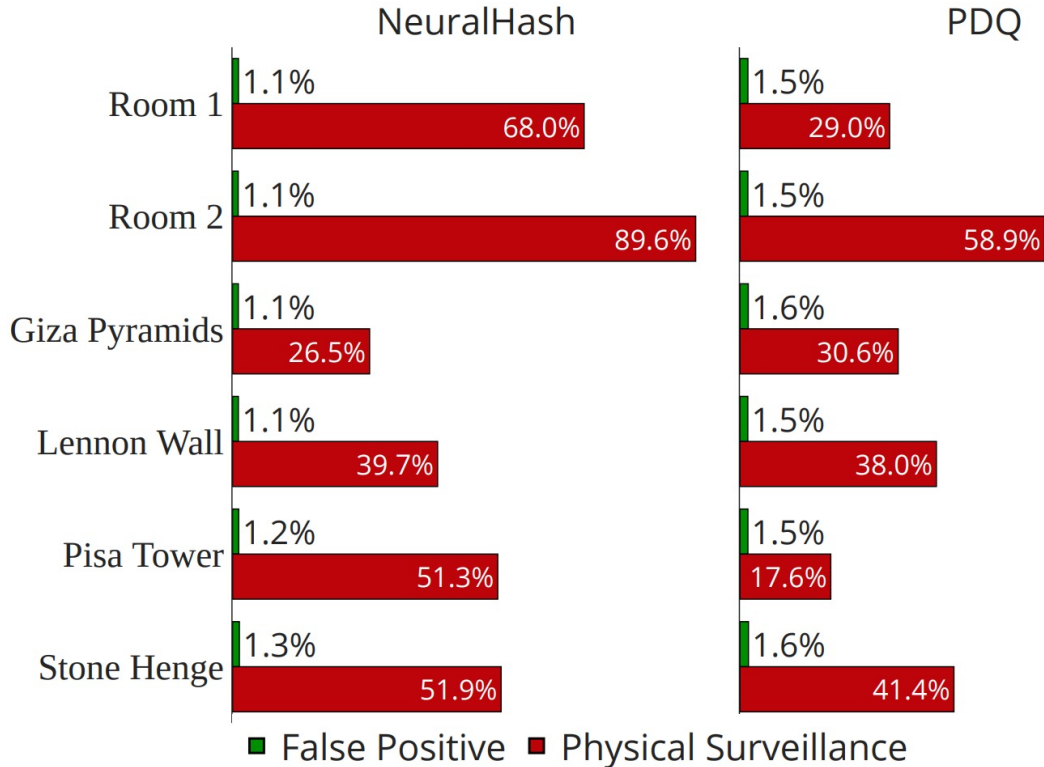
Evaluation Setup

- We demonstrate our attack on 6 different physical locations
 - Four tourist spots from the Public-Instagram dataset
 - Two room on the university campus
- We evaluate two perceptual hash functions
 - ML based : NeuralHash
 - Non ML based : PDQ

Evaluation Setup

- We demonstrate our attack on 6 different physical locations
 - Four tourist spots from the Public-Instagram dataset
 - Two room on the university campus
- We evaluate two perceptual hash functions
 - ML based : NeuralHash
 - Non ML based : PDQ
- Metrics
 - Surveillance Rate : % of target images decrypted
 - False Positive Rate : % of other images decrypted

Results



Takeaways

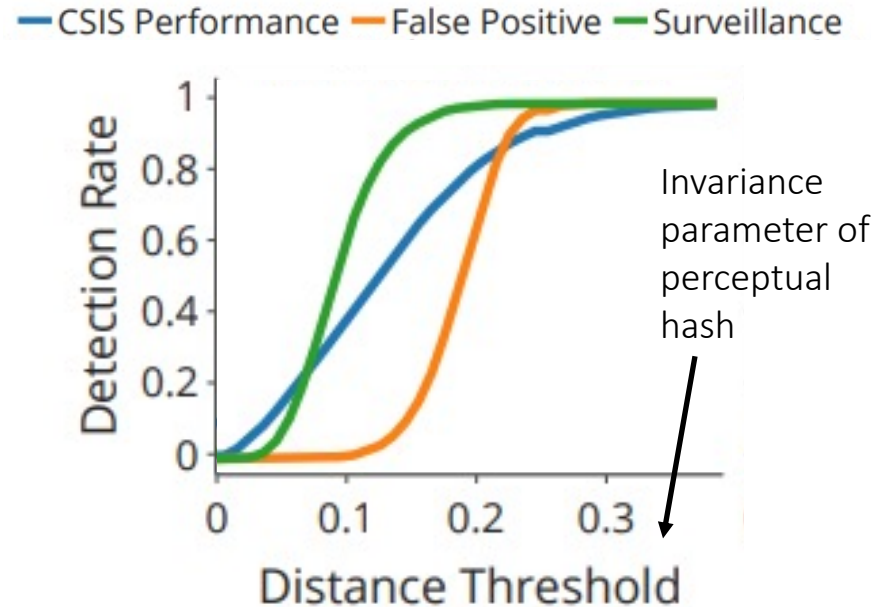
- Attack achieves nontrivial surveillance rates
- Overall, NeuralHash is more susceptible than PDQ
- Surveillance rates vary depending on the underlying scene

Security vs. Privacy Tradeoff for CSAM Detection

*To detect more CSAM ->
Perceptual hashing must
be robust to image
variations*

*If perceptual hashing is
more robust ->*

It is more vulnerable to
surveillance



Summary

- We provide experimental evidence for evaluating the surveillance risks of Client-Side Scanning Systems
- Achieve Physical surveillance rate of >30% by poisoning 0.2% of the hash database for a single location (on average)
- We characterize an undesirable trade-off: robust CSAM detection implies more robust surveillance

ahooda@wisc.edu
pages.cs.wisc.edu/~hooda/