Client-Side Scanning

What It Is and Why It Threatens Trustworthy, Private Communications

August 2022

Encryption is a technology designed to help Internet users keep their information and communications private and secure. The process of encryption scrambles information so that it can only be read by someone with the "key" to unscramble the information. Encryption protects day-to-day activities like online banking and shopping. It also prevents data from being stolen in data breaches and ensures private messages stay private. Encryption is also crucial to protect the communications of law enforcement, military personnel, and increasingly, emergency responders.

End-to-End (E2E) encryption—where the keys needed to unscramble an encrypted communication reside only on the devices communicating—provides the strongest level of security and trust. By design, only the intended recipient holds the key to decrypt the message. E2E encryption is an essential tool to ensure secure and confidential communications. Adding message scanning, even if it is "client-side", breaks the E2E encryption model and fundamentally breaches the confidentiality that users expect.

What is Client-Side Scanning?

Client-side scanning (CSS) broadly refers to systems that scan message contents—i.e., text, images, videos, files—for matches or similarities to a database of objectionable content before the message is sent to the intended recipient. For example, your anti-virus software may do this to find and disable malware on your computer.

With major platform providers moving towards implementing more E2E encryption, and calls by some in law enforcement to facilitate access to message contents to help identify and prevent sharing of objectional content[1], client-side scanning could emerge as the preferred mechanism to address objectionable content shared on E2E encrypted services without breaking the cryptography.

However, client-side scanning would compromise the privacy and security that users both assume and rely on. By making the contents of messages no longer private between the sender and

[1] https://www.newamerica.org/oti/press-releases/open-letter-law-enforcement-us-uk-and-australia-weak-encryption- puts-billions-internet-users-risk/

receiver, client-side scanning breaks the E2E trust model. The complexity it adds could also limit the reliability of a communications system, and potentially stop legitimate messages from reaching their intended destinations.

Client-Side Scanning to Prevent the Sharing of Objectionable Content

When intended to prevent people from sharing known objectionable content, client-side scanning generally refers to a way for software on user devices (often referred to as "clients" and including smartphones, tablets, or computers) to create functionally unique2 digital "fingerprints" of user content (called "hashes"). It then compares them to a database of digital fingerprints of known objectionable content such as malicious software (malware), images, videos, or graphics.3 If a match is found, the software may prevent that file from being sent, and/or notify a third party about the attempt, often without the user being aware. Newer approaches to client-side scanning also look for new objectionable content using more sophisticated algorithms. This is difficult and makes the chance of false positives even more likely.

How Client-Side Scanning Works

There are two basic methods of client-side scanning for objectionable content on an E2E encrypted communications service. One performs the comparison of digital fingerprints on the user's device, and the other does the comparison on a remote server (the content stays on the device).

1. **Comparison performed on the user's device (local digital fingerprint matching)**
   The application on a user's device (phone, tablet, or computer) has an up-to-date full database of functionally unique digital fingerprints of known content of interest. The content that the user is about to encrypt and send in a message is converted to a digital fingerprint using the same techniques applied to digital fingerprints in the full database. If a match is found, or an algorithm classifies the content as likely objectionable, then the message may not be sent, and a designated third party (such as law enforcement authorities, national security agencies, or the provider of the filtering services) could be notified.
2. **Comparison performed on a remote server**

[2] A system could be developed where the digital fingerprints are less unique, resulting in more pieces of content using the same fingerprint. However, where false positives could result in the use of serious resources (such as a police raid) designers of client-side scanning systems are incentivized to make the digital fingerprints as unique as possible.

[3] Client-side scanning is just one of the ways proposed for law enforcement or security agencies to gain access to encrypted user communications. For more information see: https://www.internetsociety.org/resources/doc/2018/encryption-brief/

There can be significant challenges with maintaining a full database and sophisticated algorithms that perform real-time analysis on a user's device. The alternative is to transmit the digital fingerprints of a user's content to a server where a comparison with a central database is performed.

Problems with Client-Side Scanning for Objectionable Content

When the comparison of digital fingerprints is done on a remote server, it could allow the service provider, and anyone else with whom they choose to share the information, to monitor and filter content a user wants to send. When the comparison takes place on the user's device, if third parties are notified of any objectionable content found, the same considerations apply. This fundamentally defeats the purpose of E2E encryption. Private and secure E2E encrypted communications between two parties, or among a group, are meant to stay private. If people suspect their content is being scanned, they may self-censor, switch to another service without client-side scanning, or use another means of communication.

It creates vulnerabilities for criminals to exploit: Adding client-side scanning functionality increases the 'attack surface' by creating additional ways to interfere with communications by manipulating the database of objectionable content. Adversaries with the ability to add digital fingerprints to the database and receive notifications when matches to those fingerprints are found would have a way to monitor select user content before it is encrypted and sent. This would allow them to track to whom, when, and where certain content was communicated. These fingerprints could include commonly used passwords or other information to enable attacks such as social engineering, extortion, or blackmail. By leveraging a system's blocking features, criminals could even choose to block users from sending specific content. This could be targeted to impact legitimate uses, potentially impeding the communications of law enforcement, emergency response and national security personnel.

It creates new technical and process challenges: If comparisons are made on the user's device, maintaining an up-to-date version of the full reference database and algorithms on every device presents its own set of challenges. These include potential process constraints (e.g., the process to add or remove content fingerprints to the database, and who has control over or access to it), bandwidth needed to transmit updated versions of the database, and the processing power on devices required to perform the comparison in real-time. Other considerations include the potential exposure of the reference database by installing it on the client device, potentially providing criminals with information about the scanning system. If comparisons are made on a central server, the digital fingerprint of content the user is attempting to send will be available to whoever controls that central server—regardless of whether it qualifies as "objectionable" in the view of the surveilling party. This opens a new set of issues around the security and privacy of users, potentially exposing details of their activity to anyone with access to the server.

Function creep—it could be used for other things: The same methods implemented in the hope of combating the worst of the worst (e.g., child exploitation or terrorism content, the two most often cited purposes to justify their use) can also be turned to mass surveillance and repressive purposes. A 2021 paper on the risks of client-side scanning noted that a CSS system could be built in a way that gives an agency the ability to preemptively scan for any type of content on any device, for any purpose, without a warrant or suspicion. Likewise, the same techniques to prevent the distribution of child sexual abuse material (CSAM) can be used to enforce policies such as censorship and suppression of political dissent by preventing legitimate content from being shared or blocking communications between users, (such as political opponents). Restricting the database to solely include fingerprints of images, videos, or URLs related to illegal activity (as some propose) is difficult. By creating digital fingerprints of more content to compare with the digital fingerprints of user content or by broadening the scope of an algorithm to classify additional types of user content as objectionable, whoever controls the system can screen for any content of interest. A client-side scanning system could be extended to monitor the text content of messages being sent, with clear and devastating implications for freedom of speech.

Lack of effectiveness: E2E encrypted communications systems exist outside the jurisdiction of any one government. A truly determined criminal would be able to switch away from services known to be using client-side scanning to avoid getting caught. It is technically simple for criminals to make modifications to objectionable content, thus changing the digital fingerprint and avoiding detection by the client-side scanning system.

Conclusion

Stopping the spread of terrorist and child exploitation material is an important cause. However, it cannot be achieved by weakening the security of user communications to potentially monitor what people say to each other. Client-side scanning reduces overall security and privacy for law-abiding users while running the risk of failing to meet its stated law enforcement objective. E2E encryption guarantees billions of users around the world can communicate securely and confidentially.[4] Major platforms continue to move towards its adoption as a way to underpin trustworthiness in their platforms and services.[5] Client-side scanning in E2E encrypted

[4] https://telegram.org/blog/200-million and https://www.newsweek.com/whatsapp-facebook-passes-two-billion-users-pledges-encryption-support-1486993

[5] https://www.facebook.com/notes/2420600258234172/

communications services is not a solution for filtering objectionable content. Nor is any other method that weakens the core of the trusted and private communications upon which we all rely.

References

Internet Society, June 2018. Encryption Brief.
https://www.internetsociety.org/resources/doc/2018/encryption- brief/

Matthew Green, December 2019. Can end-to-end encrypted systems detect child sexual abuse imagery?

https://blog.cryptographyengineering.com/2019/12/08/on-client-side-media-scanning/

Electronic Frontier Foundation, November 2019. Why Adding Client-Side Scanning Breaks End-To-End Encryption. https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end- encryption

Centre for Democracy and Technology (CDT), 2021. Content Moderation in Encrypted Systems: https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/

Hal Abelson, Ross Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G. Neumann, Ronald L. Rivest, Jeffrey I. Schiller, Bruce Schneier, Vanessa Teague, Carmela Troncoso, October 2021. Bugs in our Pockets: The Risks of Client-Side Scanning. https://www.cs.columbia.edu/~smb/papers/bugs21.pdf