A Review of the Cryptography Behind the Apple PSI System

Benny Pinkas

Dept. of Computer Science
Bar-Ilan University

July 9, 2021

Writer bio: My research in cryptography has spanned more than 25 years. I initiated the applied research on privacy preserving computation, an area of cryptography that makes it possible for multiple participants to run computations while concealing their private inputs. In particular, I pioneered research on private set intersection (PSI).

The Apple PSI system solves a very challenging problem of detecting photos with CSAM content while keeping the contents of all non-CSAM photos encrypted and private. Photos are only analyzed on users' devices. Each photo is accompanied by a safety voucher that includes information about the photo, protected by two layers of encryption. This information includes a NeuralHash and a visual derivative of the photo. Only if the Apple cloud identifies that a user is trying to upload a significant number of photos with CSAM content, the information associated with these specific photos can be opened by the cloud. If a user uploads less than a predefined threshold number of photos containing CSAM content then the information associated with all of photos of this user is kept encrypted, even if some of these photos contain CSAM content. It is important to note that no information about non- CSAM content can be revealed by the Apple PSI system.

As someone who has worked on cryptographic research for over 25 years, I can attest that the system uses a combination of well-established and well-tested cryptographic techniques, including encryption, threshold secret sharing, and private set intersection. The security of the system is based on cryptographic primitives such as elliptic-curve cryptography and AES encryption in GCM mode, which are also used by many other secure systems, including the TLS protocol which secures most internet traffic. The design is accompanied by security proofs that I have evaluated and confirmed.

Can there be a wrong match of a fingerprint? The system essentially compares a set of fingerprints of user photos with the set of fingerprints of the CSAM database. Cryptographic techniques are used to compare fingerprints while keeping the fingerprints themselves private. The probability that a fingerprint that does not exist in the database is falsely flagged as being in the database is crypto-level negligible. This probability is comparable to the extremely low likelihood of an attacker guessing the keys used by common encryption standards.

Do users learn the CSAM database? No user receives any CSAM photo, not even in encrypted form. Users receive a data structure of blinded fingerprints of photos in the CSAM database. Users cannot recover these fingerprints and therefore cannot use them to identify which photos are in the CSAM database.

Do users learn if their photos were positively matched against the CSAM database? Users get no direct feedback from the system and therefore cannot directly learn if any of their photos match the CSAM database.

What cryptographic tools are used in the implementation of the system? Every photo that is uploaded is accompanied by a safety voucher. The information about the photo in the safety voucher (NeuralHash and visual derivative) is encrypted by an external layer of encryption, which the server can only open for photos that are identified as CSAM. The information about the photo is also encrypted by an inner-layer encryption, keyed with a random key that is generated by the user's device uploading the photo. In addition, the external layer of encryption of the photo also encrypts a share of a threshold secret sharing encoding of the key used for the inner-layer encryptions of all photos of this user.

For CSAM photos, the server is able to decrypt the external layer of the safety voucher and learn the encrypted information and the share. If the number of CSAM identified photos reaches a certain threshold then the server can use the shares to recover the key which was used to encrypt the information of these photos and decrypt it.

In order to obscure the exact number of CSAM photos when this number is below the threshold, the user also uploads a random number of synthetic matches that initially look like encrypted CSAM data. The server at first only learns the combined number of initial matches, which includes both synthetic matches and actual CSAM photos. Only if the number of CSAM photos exceeds the threshold then the server becomes able to identify which of the initial matches are CSAM photos.

Each user receives a dataset of blinded fingerprints of the CSAM database. In order to be able to identify which photos are in this dataset, or which of his photos match the dataset, the user must break an elliptic-curve Diffie-Hellman cryptographic construction. This task is considered infeasible using current technology.

In conclusion, the security of the system is based on two main technologies: threshold secret sharing and PSI. The threshold secret sharing component provides unconditional security, and is safe even against attackers with unlimited computing power. The security of the PSI component relies on the AES encryption standard and on elliptic-curve cryptography. AES encryption and elliptic-curve cryptography are widely used and considered very secure. For example, a major part of the security infrastructure of the entire internet is based on AES and elliptic-curve cryptography.

Can you compare the privacy of this system to that of systems that scan photos for CSAM on the cloud? Another option for scanning photos for CSAM is to run the scanning

process on the cloud. Doing this at scale requires scanning the images in the clear.

In contrast, the Apple PSI system makes sure that only encrypted photos are uploaded. Whenever a new image is uploaded, it is locally processed on the user's device, and a safety voucher is uploaded with the photo. Only if a significant number of photos are marked as CSAM, can Apple fully decrypt their safety vouchers and recover the information of these photos. Users do not learn if any image is flagged as CSAM.

In conclusion, I believe that the Apple PSI system provides an excellent balance between privacy and utility, and will be extremely helpful in identifying CSAM content while maintaining a high level of user privacy and keeping false positives to a minimum.