Study of Fuzzy Hash Schemes

Sébastien Canard (Télécom Paris) Olivier Levillain (Télécom SudParis)

In various domains of computer science (and of cybersecurity in particular), it is useful to have a primitive to compute a fingerprint for a given file. Usually we tend to use cryptographic hash functions such as SHA-1 or SHA-256 to do so. Such functions have the following property: two similar files, differing only by one bit, will produce completely different hashes.

This is usually desirable, e.g. when we use hash functions in signature scheme. But in some use cases, we might want to have similar files produce similar hashes. For example, it may be interesting to detect malware variants by hashing a binary, and observing that the hash is similar to the hash of a known malware.

In this project, we will focus on a forensics use case of fuzzy hashes, which is inspired from previous work in a collaborative research project called PRESTO. Let us assume we have a mail server MS exporting encrypted logs, which include the fuzzy hash of the attachments contained in the messages. The logs are then stored by a service provider SP, which can be shared by various servers. When a new threat is discovered, an authority A provides IoCs (Indicators of Compromise) in the form of the fuzzy hash of a dangerous file. Our goal is to study cryptographic schemes allowing to compare the IoC to fuzzy hashes from the messages while keeping the latter (and possibly the former) encrypted.

To this aim, we can explore cryptographic technologies such as FHE (Fully Homomorphic Encryption), but we can also build ad-hoc solutions relying on the specific function used to compare two fuzzy hashes. Indeed, if comparing two hashes is just computing the Hamming distance between fixed-size binary strings, we could imagine simpler solutions such as Function Encryption or other pairing-based schemes.

Based on that, the goal of this project is to study existing fuzzy hash schemes and explore encryption schemes to find cases where comparing hashes can be done on the encrypted data, while ideally having decent performance.

The project will consist of the following tasks:

- Enumerate a set of existing fuzzy hash primitives available (ssdeep, sdhash, TLSH, etc.);
- Establish a set of relevant characteristics for these primitives (digest length, minimum length, assumptions on the documents to hash), and describe the primitives found with these characteristics;
- Study encryption schemes such as FHE, FE or other pairing-based schemes;
- Propose a protocol (relying on a fuzzy hash scheme and an encryption scheme) where it is possible to encrypt the fuzzy hashes in a way it is possible to compare it with a target hash, as presented in the use case.

Obviously, you might propose several protocols, with different degrees of practicality and efficiency. If time permits, you may want to work on extensions, e.g. implementing a protocol, or proving the security properties thereof.

Prerequisites

- Solid Cryptographic Background on so-called “boring” cryptography.
- Notions about fancier cryptography such as FHE or Functional Encryption.

Practical Information

Regular meetings will be scheduled with the advisors, either via visioconference or in Palaiseau. If you are interested, send an email to `sebastien.canard@telecom-paris.fr` and `olivier.levillain@telecom-sudparis.eu`.