Technical Disclosure Commons

Defensive Publications Series

July 2020

Hashing and Encrypting Public Unsecured Datasets for Secure Dataset Fingerprinting

Anonymous

Hashing and Encrypting Public Unsecured Datasets for Secure Dataset Fingerprinting

ABSTRACT

There are a number of public unsecured datasets (PUDs) that include data from users of social media service providers or other entities. Such service providers and other organizations are often called upon to retrieve and secure PUDs. There are certain difficulties that arise, e.g., verifying that the PUD includes data from that organization and not any other; handling the data so that personnel from the organization or its vendors cannot access or use it for unauthorized purposes; efficiently analyzing the PUD; etc.

This disclosure describes secure procedures for handling PUD data using keyed hashing-and-encryption functions such that an organization can efficiently determine whether it was (or was not) the origin of the PUD data without individuals affiliated to the organization, e.g., employees or vendors, gaining access to the plaintext PUD data.

KEYWORDS

- Public unsecured dataset (PUD)
- Social media
- User account
- Dataset verification
- Data privacy
- Personally identifiable information (PII)
- User identifiable information (UII)
- Keyed-hash message authentication code (HMAC)

BACKGROUND

There are a number of public unsecured datasets (PUDs) that include data from users of social media service providers or other entities. Such service providers and other organizations are often called upon to retrieve and secure PUDs. There are certain difficulties that arise, e.g., verifying that the PUD includes data from that organization and not any other; handling the data so that personnel from the organization or its vendors cannot access or use it for unauthorized purposes; efficiently analyzing the PUD; etc.

There are a number of public unsecured datasets (PUDs) that include data from users of social media service providers or other entities. Although data within a user's account at such services is generally secure, public accessibility of portions of this data outside of the service provider poses a risk to both consumers and to the service provider. Such service providers and other organizations are often called upon to retrieve and secure PUDs, for which they may employ third-party vendors.

There are certain difficulties that arise during this process, as described below:

Mixed datasets: When a service provider is notified of PUDs that may contain unknown records of personally identifiable information (PII) or user identifiable information (UII), it is not possible to a priori determine conclusively whether the PUD includes PII/UII data specifically from that provider or from other organizations. An organization typically verifies that the PUD did originate from itself, prior to ingesting, uploading, or analyzing such PUD.

Data Security: During the data-handling procedures used to secure PUDs, potential risks arise as follows:

- Reading of unencrypted PII/UII data by a vendor or anyone else in the chain of PUD acquisition and transmittal to the organization.

- Indefinite storage of PUD data by vendors.
- Usage of PUD data acquired on behalf of the organization for unauthorized purposes.
- Disclosure of PII/UII through potential vendor breaches.
- Access of PUD by employees without business need or authorization.
- Use of PUD data for purposes other than protecting privacy.
- Storage of PUD data originating from one organization by another organization.
- Storage of data from deleted accounts.

Efficient analysis of datasets: Upon identifying a large PUD, it is efficient to determine whether the dataset is a recurrence of a previous one or is a novel dataset with new records. This also enables the determination of possible causes for the dataset exposure, such as whether the data originated from a third-party breach, the extent to which user privacy may have been compromised, etc.

DESCRIPTION

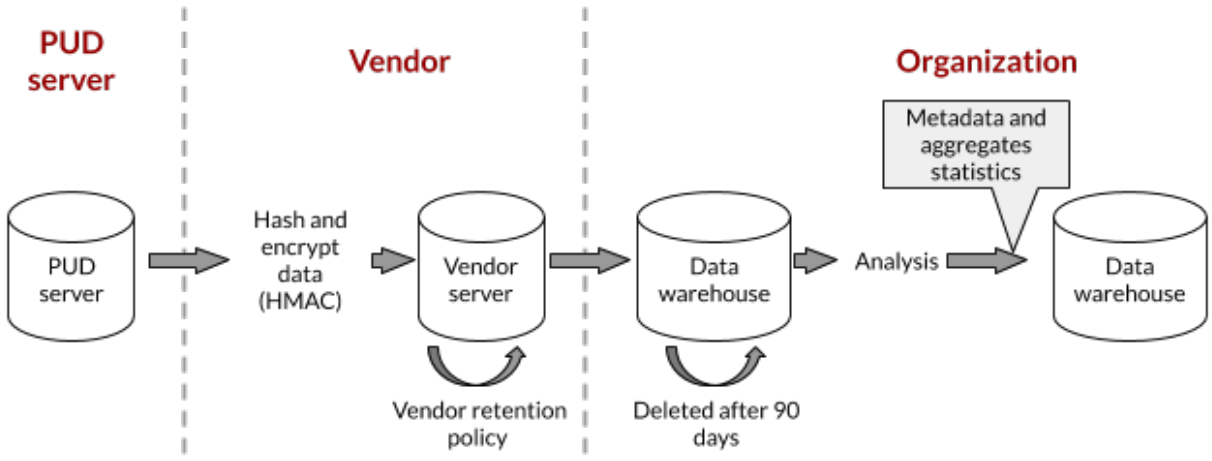


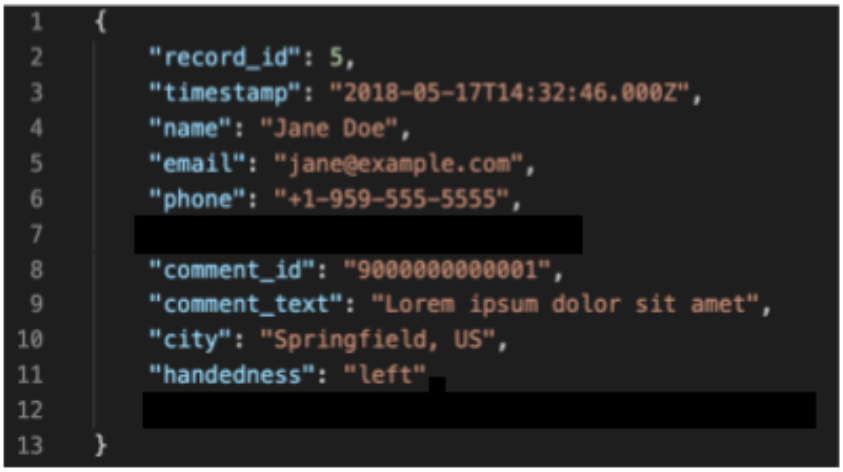
Fig. 1: Procedure for handling PUD datasets in a secure manner

Per the techniques of this disclosure, illustrated in Fig. 1, to improve privacy protection for user data and to minimize risks for an organization and for its third-party vendors when

handling PUDs, a specific procedure for handling any datasets ingested into the organization is utilized.

Vendor dataset acquisition: The vendor hashes and encrypts PUD data upon retrieving the data prior to storing it on disk. Each individual datapoint, e.g., each value in a PUD record, is hashed separately using a keyed hash function, e.g., HMAC.

(a)



(b)

Fig. 2: (a) A (synthetic) record of user data in plaintext; (b) The record after processing with an HMAC function. After HMAC processing, the record is essentially indecipherable.

Fig. 2, which illustrates HMAC operation on a record of user data, shows that processing by an HMAC function renders a user record in the PUD essentially indecipherable. The HMAC

function is deterministic, e.g., given the same input, it produces the same output. It is also irreversible, e.g., given the output of an HMAC function, it is not computationally feasible to recover its input. Although an HMAC function disables the viewing of plaintext data by humans, it enables the comparison of two datasets, e.g., to determine if two hashed-and-encrypted datasets are identical or not.

Dataset transfer to the organization: The organization performs intermediate storage of the hashed and encrypted data for a limited amount of time, e.g., ninety days, in a data warehouse. The limited data retention serves as an additional measure of data security.

Fingerprinting and dataset analysis: A fast, accurate fingerprint of the just-acquired dataset is computed. By comparing the fingerprint of the just-acquired PUD against a PUD-fingerprint archive, analysts can determine whether there are any data, e.g., user IDs, emails, names, usernames, phone numbers, etc., in the PUD that likely originate from the organization. Non-organizational data or data from deleted accounts are filtered out. Metadata and aggregate statistics, e.g., number of users in the PUD, distribution of the age of the accounts, distribution of the country of the accounts, etc., can be stored in a data warehouse with suitable access controls. Aggregate statistics enable the organization to accurately respond to data protection agencies.

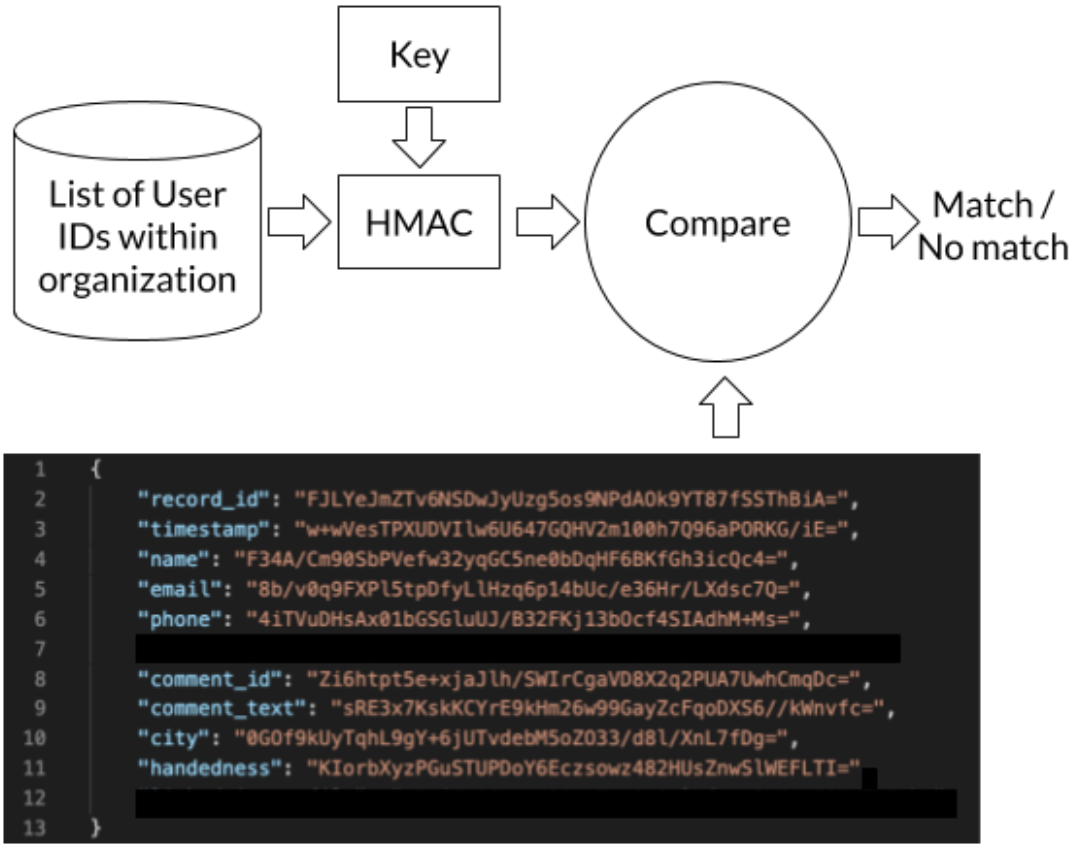


Fig. 3: Testing hashed-and-encrypted records for matches

Fig. 3 illustrates the testing of hashed-and-encrypted records for matches. User data held by the organization is effectively hashed and encrypted using an HMAC function with a key, and compared with the hashed-and-encrypted PUD data to detect a match. In this manner, PUD data that originated from the organization can be discovered without any human user having access to plaintext PUD data. The use of HMAC functions, per the techniques of this disclosure, protect data from being used for unauthorized purposes that may not preserve data privacy. Also, subjecting data to HMAC processing ensures that the data is protected in the event of a breach.

In this manner, the techniques of this disclosure enable any service provider or organization that holds user data to engage with third-party vendors to proactively identify, acquire, and analyze PUDs that may include organizational or user data. The techniques enable

organizations to build data-handling procedures and tools that protect user privacy, and fortify controls around the data during the time it may be handled by third-party vendors and ingested into the organization. The techniques also enable quick analysis of the data to determine if the PUD originated from the organization, and if so, the proper ways of handling it. For example, a service provider that discovers data pertaining to a user in a PUD can inform that user of the data breach and/or take other mitigating actions.

CONCLUSION

This disclosure describes secure procedures for handling PUD data using keyed hashing-and-encryption functions such that an organization can efficiently determine whether it was (or was not) the origin of the PUD data without individuals affiliated to the organization, e.g., employees or vendors, gaining access to the plaintext PUD data.

REFERENCES

1. https://en.wikipedia.org/wiki/HMAC