Cryptography for Social Science

How Encryption and Hashing can be used to Anonymize, Limit Access, and Protect against Tampering

Thomas Pronk, pronkthomas@gmail.com, v1.0, 2019-03-20

Introduction

Social science in Europe is expected to meet The General Data Protection Regulation (GDPR), which sets a higher standard for privacy and security. It may be challenging to meet these standards while maintaining a practical and reproducible research workflow. Cryptography may help, but is often perceived as an obscure and complex topic. This document provides a basic introduction to cryptography and its uses for social scientists. Cryptography operates by transforming a text input via a cipher function into an obfuscated output, called a hash or ciphertext. Ciphers differ with regard to what kind of information one may obtain about the original input from the output. I describe the operation of two types of cryptography, being encryption and hashing. For each, I illustrate how they may be used to anonymize personal data, limit access to sensitive data, and protect against tampering. Additionally, I describe security consideration of each method from the perspective of an adversary that has access to the obfuscated data and tries to recover the input data.

Encryption

How it Works

An encryption function converts a given input into different *ciphertexts* each time the function is applied, but it does offer a method for decrypting the ciphertext back into the input text. *Symmetric encryption*, such as protecting a ZIP file with a password, uses a single *private key* for encrypting and decrypting. *Asymmetric encryption* does so by using a pair of two keys; a *public key* for encrypting and a private key for decrypting. Via encryption, a ciphertext (and public key) can be transferred or stored via relatively insecure methods, while limiting access to the original input to those that have the private key. For example, the workflow shown below may be used to have a participant submit General Practitioner (GP) information via a website, but limit access to this information to a confidant owning the private key:

1. Confidant generates public-private key pair on a secure system
2. Confidant uploads public key to website, but keeps the private key separate
3. Participant enters GP information on website, which is encrypted via the public key, and then stored in website database
4. Confidant downloads encrypted GP information
5. Confidant decrypts GP information with the private key on a secure system

Security Considerations

Encryption is a very active field of study, so be sure to apply the most recent recommendations. Additionally, security is limited by the degree to which the private key is kept secret, hence, special security measures for storing the private key may be useful.

Hashing

How it Works

A hashing cipher always converts a given input text into the same output *hash* (e.g. applying the MD5 cipher to “John Doe” will always yield the hash “4C2A904BAFBA06591225113AD17B5CEC”), and it does not offer any straightforward method for decrypting the hash back into the input text. Hashing can have at least three applications in social science:

1. It can be a powerful method of anonymization; a hashed name may offer a unique identifier on the basis of personal information that has a relatively low risk on being de-anonymized.
2. Publishing a hash of a dataset may protect against tampering of research data, because recalculating the hash of a given dataset should always give the same cipher.
3. Using hashes as participant identifiers may protect against tampering by participants. For example, it may be used to obfuscate participant IDs that are visible in web-browsers or to offer participants a unique code via which they can withdraw their data.

Security Considerations

Like encryption, hashing is a very active field of study, so be sure to apply the most recent recommendations[1]. Additionally, security is limited by the degree to which an adversary may try out input texts in order to find out whether any match. If this can reasonably be achieved, hashed data may not be considered anonymized, but pseudonymized. For example, and adversary may check whether a hash represents “John Doe” hashed via MD5, by applying the MD5 function to “John Doe” and comparing the output to the hash. More advanced attacks may employ *rainbow tables*, which contain hashes of a large range of commonly used passwords, so that these can be decoded more efficiently. In order to counter such *attacks*, a random string, called the *salt*, is appended to the input before hashing it. So long as the salt is kept secret, the hash is relatively secure. When the salt is not needed anymore, for instance when data collection has been completed, the salt can be deleted or stored securely.

Conclusion

I have briefly introduced cryptography via encryption and hashing, and illustrated how they may be used to anonymize personal data, limit access to sensitive data, and protect against tampering. When applying cryptography in your project, be sure to consult the latest insights as to which hashing or encryption functions may be most secure. The current tutorial was aimed to be a basic introduction for social scientists. More advanced topics may include *searchable encryption***,** which can anonymize data while still allowing some degree of data analysis[2].

Acknowledgments

I thank Jasper Wijnen for his valuable feedback on an a earlier version of this document.

[1] https://en.m.wikipedia.org/wiki/Hash_function_security_summary

[2] https://blog.cryptographyengineering.com/2019/02/11/attack-of-the-week-searchable-encryption-and-the-ever-expanding-leakage-function/