Privacy-preserving SARS-CoV-2 Contact Research

Authors:

Ferenc Vágujhelyi
Chairman, Scientific Association for Infocommunications
(https://www.hte.hu/web/en/organization)

Gábor Magyar PhD
Associate Professor, Budapest University of Technology and Economics
(https://www.tmit.bme.hu/magyar.gabor?language=en)

Virág Réti
Chief Executive Officer, Xtendr Zrt.
(https://xtendr.io/, https://www.linkedin.com/in/vreti/)

Contact: papers@xtendr.io

Abstract
Contact research is an essential method to combat the pandemic until an effective vaccine is developed and made available. Students and teachers are essentially the bridges between two communities: the school and their households. If a positive test is detected the people at risk of infection must immediately be identified. Existing principles relating to "processing of personal data" prohibit the mass sharing of data from natural persons who do not currently have a visible connection to the epidemic. This article shows a method where schools collect purely encrypted data from the teachers and parents about the members of their household. The epidemiological authority sends an encrypted identifier of the person who tested positive to schools to look for contacts by the school without the ability to identify any natural persons. The school sends only ciphers back to the authority who can decrypt the result. The cryptosystem is based on homomorphic mappings and multiparty computing to withstand attacks.

The issue

The coronavirus pandemic has significantly restricted the possibility of attending schools in the last school year. Restoring the order of public education is an essential social interest. An effective tool in combating the mass spread of the epidemic is contact research of people detected with SARS-CoV-2 virus infection. Contact research is based on the information whether a virus carrier has met or may have encountered other people when the pathogen was believed to have been in their body.

The task

The schools are attended by students and teachers. Each of them belongs to two sets of people: the school community and the family. If the health care system detects a SARS-CoV-2 positive person, we want to identify their related communities in school and their household(s) and notify the epidemiological body about their contacts.

The "direct" approach

Parents and teachers report the Social Security Number (SSN) of each household member. If the epidemiological body detects a person as infected, it broadcasts their SSN to all schools participating in the proposed system. The school can find the infected person if they are in the database with all his direct school and household contacts. The recovered contacts are to be reported to the epidemiological body by the school.

We have one more problem yet to be answered. The identification of the person who is eligible to report. The school can send a secret value to the parent that she can use to identify herself.

Why does this "direct" approach solution fail?
Because it violates the principles relating to "processing of personal data". The epidemiological body will not be allowed to broadcast the identifier of an infected person to all schools, especially those who have no relation to the education system. The purpose of the processing of personal data in the school is to identify contacts, not persons. Such a data collection would not be purposeful neither at school nor at the health authority because they are not eligible to identify people as household or school mates if they are not a contact of an infected person.

The solution
We have to fit data processing to the purpose: the schools have to identify "relationships between persons", but not "the persons". SSNs must be encrypted before sending them to schools. If parents and the health authority use a common encryption key, then it has to be distributed thus can be considered as public information giving way to "rainbow table" based attacks. To give an example, if a cryptographic hash function is used with some "salt" then any parent is able to calculate the rainbow table. (He can reveal the "salt" from the developer window of his web browser.) We have to use unique encryption keys for each household report by parents and teachers. But unique keys mean unique ciphers for the same SSNs where relationships can hardly be recovered.

Unique ciphers have to be mapped to pseudonyms to ensure that the same SSNs are referred to as the same value. More cryptographic mappings performed with more keys and by more parties is a "multiparty computing" cryptographic scheme. The school needs special crypto-keys to do the mapping from ciphers to pseudonyms. The result is a database with SSN pseudonyms linked to the reported ciphers.

When the epidemiological body identifies a newly infected person, it encrypts their SSN the same way parents do. The schools can map the cipher to their pseudonym and look for a match in their database. In case of a hit, the original ciphers will be reported to the epidemiological body, with the serial number of the mapping key. The epidemiological body requests the decryption key with those serial numbers and decrypts the ciphers revealing the SSNs and the related class identifiers. The health authority can then contact the school or the concerned persons directly. But where are the encryption, the pseudonym mapping, and the decryption keys from?

The crypto-system
We have three cryptographic keys:
(1) encrypt key: the parent, the teacher or the epidemiological body encrypts the SSNs to ciphers before sending the data to the school,
(2) pseudonym mapping key: the school applies pseudonym mapping on the different ciphers resulting in the same pseudonym if the original SSNs were the same,
(3) decrypt key: the epidemiological body decrypts the ciphers reported by the schools to open data.

The key triplets are to be calculated in a single process by a "key service". The problem is that if the key service provider knows both the encryption and pseudonym mapping keys then it could map open data to pseudonyms building a rainbow table. Decryption keys should also be protected; however, key service has no access to ciphers to decrypt — unless a school cooperates with it fraudulently. Therefore, pseudonym mapping keys must be encrypted, only to be used by the belonging school, and decryption keys must be encrypted as well, only to be used by the epidemiological body.

We have to protect the system from fraudulent cooperation between a parent and the school to prevent them from using an encryption-pseudonymization keypair in more than one report session. The crypto-system guards these issues while the key service detects if a parent-school pair requests an unreasonable number of keys.

Conclusion

Contact tracking is possible despite heavy privacy concerns if proper technology is applied. Faster notifications to the epidemiological body are shown to be effective in slowing down the SARS-CoV-2 epidemic. The aforementioned solution reduces notifications by streamlining contact research in a potentially superspreader community: education participants; without violating privacy.

Figures

Figure 1: https://xtendr.io/wp-content/uploads/2020/09/Fig_1.pdf
Figure 2: https://xtendr.io/wp-content/uploads/2020/09/Fig_2.pdf
Figure 3: https://xtendr.io/wp-content/uploads/2020/09/Fig_3.pdf

References:

(1) ENISA, Recommendations on shaping technology according to GDPR provisions
Published: January 28, 2019
https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions

(2) US Patent Pending: US20190213356A1 - Data management method and registration method for an anonymous data sharing system, Inventor: F. Vagujhelyi, G. Magyar, Current Assignee: Xtendr Zrt.
https://patents.google.com/patent/US20190213356A1/en