Privacy Preserving Set Intersection.

Giuseppe Bruno[1], Diana Nicoletti[1], Monica Scannapieco[2] and Diego Zardetto[2]

[1]Bank of Italy
[2]Italian National Statistical Office

Irving Fisher Committee Conference. BIS, Basel, August 30th 2018

Outline

Why do we want to link datasets

merging datasets

- Administrative records on firms and individuals have a huge potential for statistical studies.
- The law forbids the merging and processing of non-anonymized data, thus making it difficult to carry out studies requiring several sources of data.
- It would be helpful to take advantage of hashing and cryptographic techniques to carry out safe linkage between different datasets.

Why do we want to link datasets

merging datasets

- Administrative records on firms and individuals have a huge potential for statistical studies.
- The law forbids the merging and processing of non-anonymized data, thus making it difficult to carry out studies requiring several sources of data.
- It would be helpful to take advantage of hashing and cryptographic techniques to carry out safe linkage between different datasets.

Why do we want to link datasets

merging datasets

- Administrative records on firms and individuals have a huge potential for statistical studies.
- The law forbids the merging and processing of non-anonymized data, thus making it difficult to carry out studies requiring several sources of data.
- It would be helpful to take advantage of hashing and cryptographic techniques to carry out safe linkage between different datasets.

Envisaged social benefit

leveraging larger datasets

Possible social benefits from sharing otherwise private databases:

- Different hospitals could improve their medical analytics for better healthcare delivery.
- State tax authority would like to check banking relationships with suspect tax evader.
- National law enforcement bodies of different countries would like to compare their respective database of suspected terrorists.

Envisaged social benefit

leveraging larger datasets

Possible social benefits from sharing otherwise private databases:

- Different hospitals could improve their medical analytics for better healthcare delivery.
- State tax authority would like to check banking relationships with suspect tax evader.
- National law enforcement bodies of different countries would like to compare their respective database of suspected terrorists.

Envisaged social benefit

leveraging larger datasets

Possible social benefits from sharing otherwise private databases:

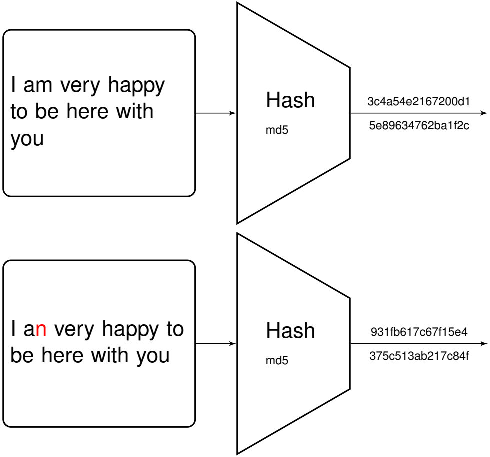
- Different hospitals could improve their medical analytics for better healthcare delivery.
- State tax authority would like to check banking relationships with suspect tax evader.
- National law enforcement bodies of different countries would like to compare their respective database of suspected terrorists.

Asymmetric encryption and digital signature

RSA asymmetric encryption guarantees a bilateral secure communication.

- RSA (for Rivest, Shamir & Adleman) was introduced in 1977 MIT;
- known as public-key scheme;
- based on modular exponentiation on an integer field;
- security is linked to the complexity of factoring huge numbers (300 digits);

What is a hash function?



Residual disclosure risk

Main assumption: Honest but curious behaviour. A unit is defined at risk when it can easily be singled out from other records. We distinguish three cases:

- quasi-identifiers are of *categorical* kind;
- quasi-identifiers are of *continuos* kind;
- quasi-identifiers are of mixed kind.

Our protocol doesn't protect against malicious behavior aiming at individual re-identification. Generalization and suppression techniques could be helpful.

Residual disclosure risk

Main assumption: Honest but curious behaviour. A unit is defined at risk when it can easily be singled out from other records. We distinguish three cases:

- quasi-identifiers are of *categorical* kind;
- quasi-identifiers are of *continuos* kind;
- quasi-identifiers are of mixed kind.

Our protocol doesn't protect against malicious behavior aiming at individual re-identification. Generalization and suppression techniques could be helpful.

Residual disclosure risk

Main assumption: Honest but curious behaviour. A unit is defined at risk when it can easily be singled out from other records. We distinguish three cases:

- quasi-identifiers are of *categorical* kind;
- quasi-identifiers are of *continuos* kind;
- quasi-identifiers are of mixed kind.

Our protocol doesn't protect against malicious behavior aiming at individual re-identification. Generalization and suppression techniques could be helpful.

Private Set Intersection flavours

Private Set Intersection: a cryptographic protocol involving two parties/institutions endowed with a private set. The two parties, a client and a server, want to jointly compute the intersection of their private input sets in a way that at the end the client learns the intersection and the server learns nothing.

- *Plain Private Set Intersection* (PSI)
- *Authorized Private Set Intersection* (APSI)

The difference between these two protocols is that in APSI each element in the client set must be authorized for sharing by some recognized and mutually trusted authority.

Private Set Intersection flavours

Private Set Intersection: a cryptographic protocol involving two parties/institutions endowed with a private set. The two parties, a client and a server, want to jointly compute the intersection of their private input sets in a way that at the end the client learns the intersection and the server learns nothing.

- *Plain Private Set Intersection* (PSI)
- *Authorized Private Set Intersection* (APSI)

The difference between these two protocols is that in APSI each element in the client set must be authorized for sharing by some recognized and mutually trusted authority.

Private Set Intersection flavours

Private Set Intersection: a cryptographic protocol involving two parties/institutions endowed with a private set. The two parties, a client and a server, want to jointly compute the intersection of their private input sets in a way that at the end the client learns the intersection and the server learns nothing.

- *Plain Private Set Intersection* (PSI)
- *Authorized Private Set Intersection* (APSI)

The difference between these two protocols is that in APSI each element in the client set must be authorized for sharing by some recognized and mutually trusted authority.

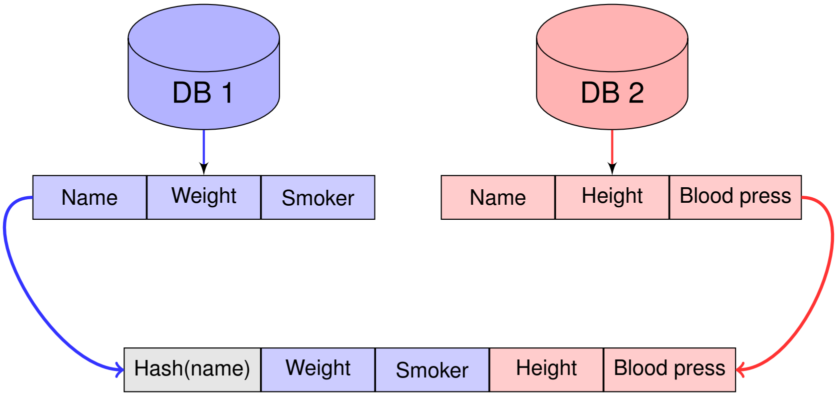
Private Set Intersection flavours

Private Set Intersection: a cryptographic protocol involving two parties/institutions endowed with a private set. The two parties, a client and a server, want to jointly compute the intersection of their private input sets in a way that at the end the client learns the intersection and the server learns nothing.

- *Plain Private Set Intersection* (PSI)
- *Authorized Private Set Intersection* (APSI)

The difference between these two protocols is that in APSI each element in the client set must be authorized for sharing by some recognized and mutually trusted authority.

The Private set intersection scheme



The protocol: offline section

Initial data:

- RSA public and private keys;
- Client's input: $\mathcal{C} = \{hc_1, \ldots, hc_v\}$ where $hc_i = hash(c_i)$;
- Server's input: $\mathcal{S} = \{hs_1, \ldots, hs_w\}$ where $hs_i = hash(s_i)$;

The protocol is broken down into two phases:

OFF-LINE:

1. Server: $\forall j : K_{s:j} = \left(hash(s_j)\right)^d \mod n; \quad t_j = H'(K_{s:j})$
2. Client: $\forall i : R_{c:i} \sim \mathcal{U}[0, Z_n^*]; \quad y_i = hash(c_i) \cdot (R_{c:i})^e \mod n$

The protocol: online section

ON-LINE:

1. Client: $\xrightarrow{y_1, y_2, \ldots, y_v}$ Server;
2. Server: $\forall i : y'_i = (hash(y_i))^d \mod n$
3. Server: $\xrightarrow{\{y'_1, \ldots, y'_v\} \quad \{t_1, \ldots, t_w\}}$ Client;
4. Client: $\forall i : K_{c:i} = y'_i / R_{c:i}$ and $t'_i = H'(K_{c:i})$

   Result: $\{t'_1, \ldots t'_v\} \cap \{t_1, \ldots, t_w\}$

Protocol characteristics

Our protocol satisfy the following conditions:

- **Correctness:** at the end of *Interaction*, Client outputs the exact intersection;
- **Server privacy:** The client learns no information about the server elements not belonging to the intersection ;
- **Client privacy:** The Server learns no information about the client elements except the upper bound on the client's set size ;
- **Client unlinkability:** a malicious server cannot tell if any two instances of *Interaction* are related, (executed on the same inputs);

Concluding Remarks

- suggested how to take advantage of cryptographic functions for sharing private data;
- shown how to implement a Private Set Intersection protocol giving a Client only the anonymized common records;
- provided a data sharing environment without a trusted third party;
- improving the security with some form of authentication;
- outlining possible avenues for computing scalability up to 10^9;

For Further Reading

E. De Cristofaro and G. Tsudik.
Practical Private Set Intersection Protocols with linear Computational and Bandwidth Complexity.
proc Financial Cryptography and data Security, 2010.

R. Agrawal, A. Evfimieski and R. Srikant.
Information Sharing across Databases.
Sigmod Conference, 2003.

M. Scannapieco, I. Figotin, E. Bertino and A. Elmagarmid.
Privacy Preserving Schema and Data Matching.
Sigmod Conference, 2007.

Thank you very much for your attention.

Vielen Dank für ihre Aufmerksamkeit.

Merci beaucoup pour votre attention.

Questions?