SECURE RECORD LINKAGE

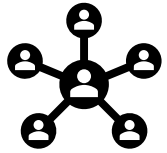
Constance Beguier

Nicolas Reyland

SOMMAIRE

- Pourquoi faire du SRL ?
- État de l'art
- Comment modifier l'état de l'art pour faire du SRL ?
- Notre implémentation, nos résultats
- Conclusion

POURQUOI FAIRE DU SRL ?



Mise en correspondance de données de différentes DB

Préserve la confidentialité des données

Uniquement les données correspondantes doivent être révélées !

QUELLE UTILITÉ ?

Analyse médicale

- Efficacité des traitements
- Identification des facteurs de risques pour certaines maladies

Sécurité nationale et internationale

- Échange d'informations sur des personnes considérées comme dangereuses

Prevenir la fraude

- Blanchiment d'argent
- Fraude à l'assurance

DATASET IDASH

- 2 fichiers CSV
- 2 * 500'000 patients
- 9 données possibles par patient (nom, date de naissance, SSN, etc.)
- 50'000 patients en commun
- 8671 correspondances parfaites

Données incompletes et malformées

Hopital 1

Nom	Prénom	Sexe	Poids	Taille	Age	Adresse	SSN	Grp S
Garen	Crownguard	M	100kg	1,95m	28		4506908	O+

Hopital 2

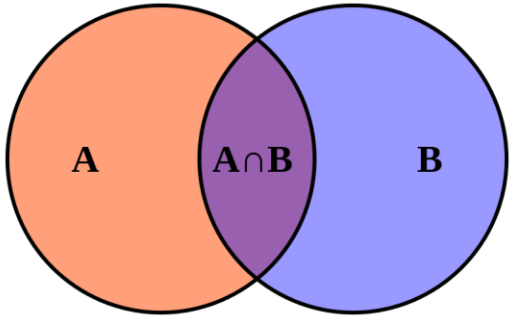
Nom	Prénom	Sexe	Poids	Taille	Age	Adresse	SSN	Grp S
Karen	Crownguard	M	99kg		28	Demacia, TGC		O+

PRIVATE SET INTERSECTION

Technique de SMC (Secure Multiparty Computation)

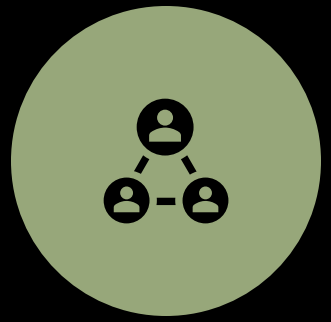
Intersection de deux ensembles

Les éléments qui ne sont pas dans l'intersection ne sont pas révélés



ÉTAT DE L'ART DU PSI

TYPES D'«ADVERSAIRES»
- SEMI-HONNÊTE
- MALICIEUX

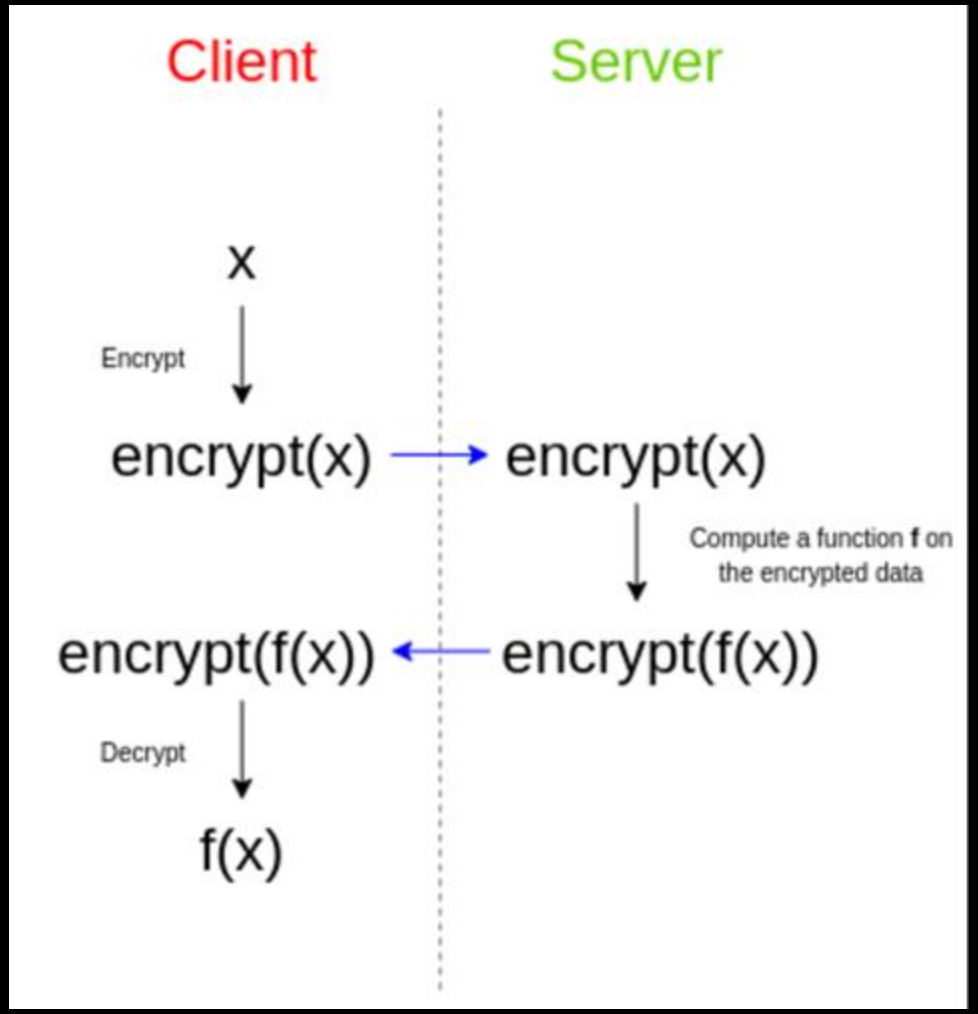


TROIS PRINCIPALES APPROCHES :
1. DIFFIE-HELLMAN
2. FHE (FULLY HOMOMORPHIC ENCYPTION)
3. OT (OBLIVIOUS TRANSFER)

ETAT DE L'ART: DIFFIE-HELLMAN

- C. Meadows. A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party. In IEEE S & P, 1986.

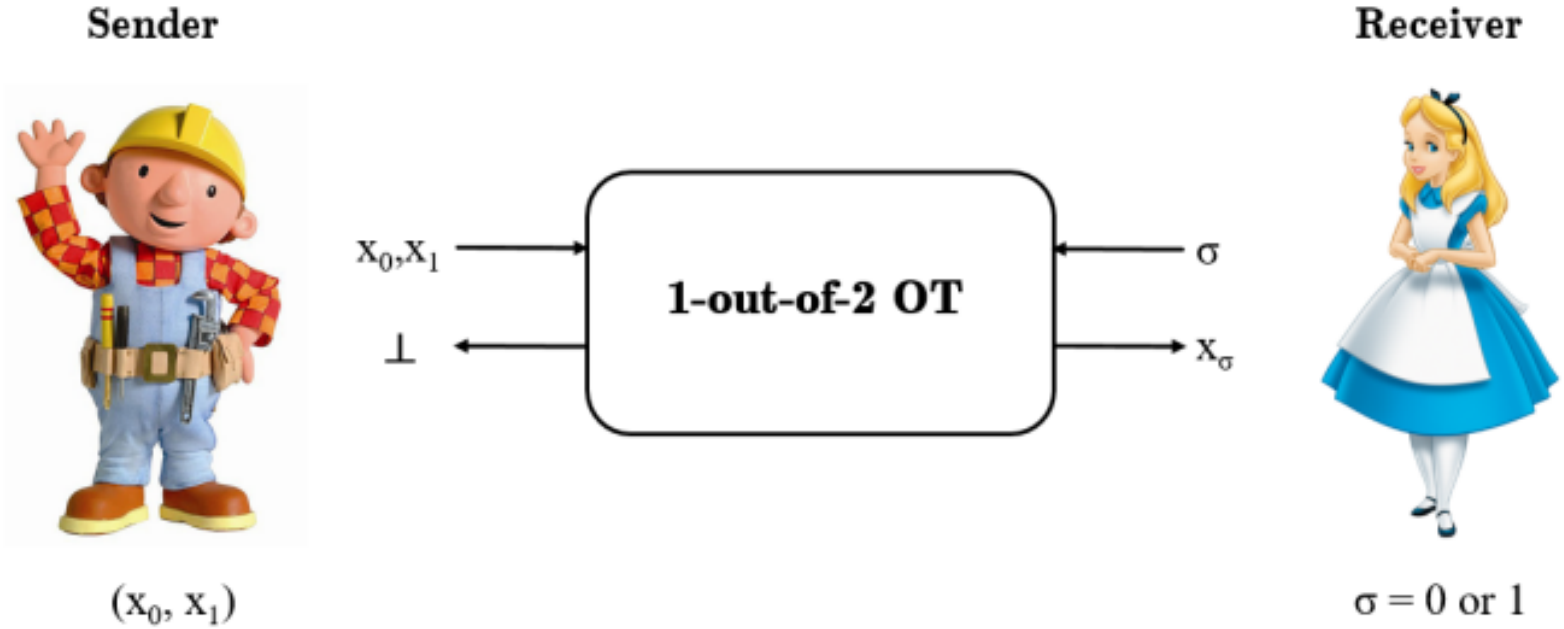
FHE ?


Client
Server
x
Encrypt
encrypt(x)
encrypt(x)
Compute a function f on the encrypted data
encrypt(f(x))
encrypt(f(x))
Decrypt
f(x)

ETAT DE L'ART: FHE (FULLY HOMOMORPHIC ENCYPTION)

- Hao Chen, Kim Laine, and Peter Rindal. Fast private set intersection from homomorphic encryption. In CCS, 2017.
- Yan Huang, David Evans, and Jonathan Katz. Private set intersection: Are garbled circuits better than custom protocols? In NDSS, 2012.
- Benny Pinkas, Thomas Schneider, Gil Segev, and Michael Zohner. Phasing: Private set intersection using permutation-based hashing. In USENIX, 2015.

OT ?


Sender
Receiver
x_0, x_1
$\perp$
1-out-of-2 OT
σ
x_σ
(x_0, x_1)
$\sigma = 0$ or 1

ETAT DE L'ART: OT (OBLIVIOUS TRANSFER)

- Vladimir Kolesnikov, Ranjit Kumaresan, Mike Rosulek, and Ni Trieu. Efficient batched oblivious PRF with applications to private set intersection. In CCS, 2016.
- Benny Pinkas, Thomas Schneider, Christian Weinert, and Udi Wieder. Efficient circuitbased PSI via cuckoo hashing. In EUROCRYPT, 2018.
- Benny Pinkas, Thomas Schneider, Oleksandr Tkachenko, and Avishay Yanai. Efficient circuit-based PSI with linear communication. In EUROCRYPT, 2019.
- Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. Spot-light: Lightweight private set intersection from sparse OT extension. In CRYPTO, 2019.
- Melissa Chase and Peihan Miao. Private set intersection in the internet setting from lightweight oblivious PRF. In CRYPTO, 2020.

DIFFÉRENCES ENTRE OT ET FHE

Avantages OT

- Efficacité des calculs
- OT Extension disponible

Désavantages OT

- Grande complexité de communication
  obligatoirement synchrone
  lourde en volume

Avantages FHE

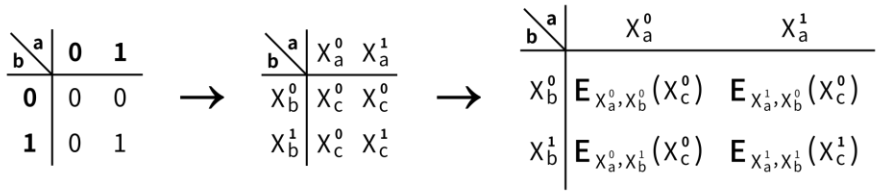
- Faible complexité de communication
  peut être asynchrone
  faible en volume

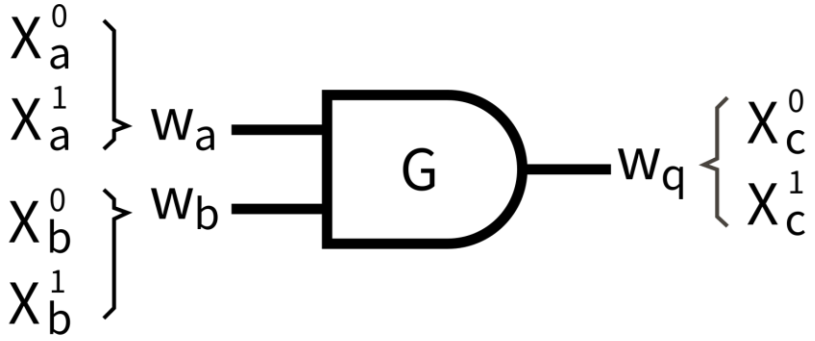
Désavantages FHE

- Grande complexité calculatoire
- Malléabilité

CIRCUITS « BROUILLÉS » (GARBLED CIRCUITS)

- Évaluation sécurisée d'un circuit booléen
- Un brouilleur, celui qui a le circuit
- Un évaluateur, celui qui évalué le circuit avec les entrées des deux parties



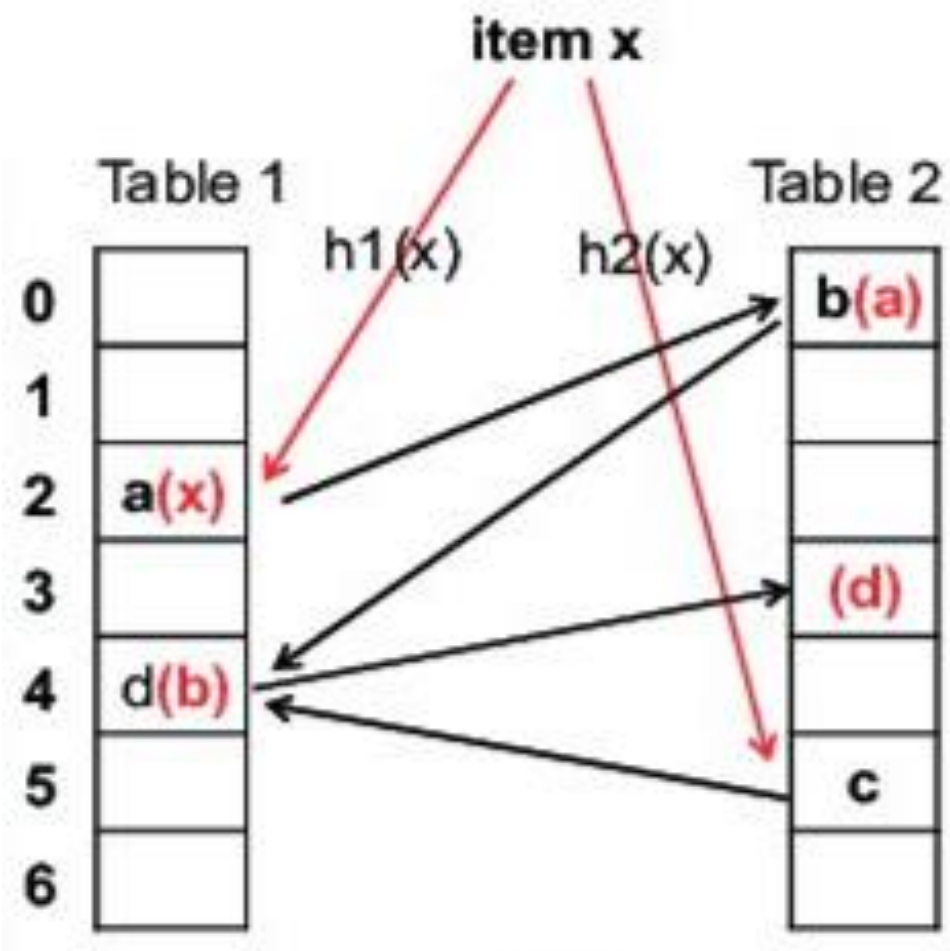


PAPIERS IMPORTANTS: CUCKOO

Benny Pinkas, Thomas Schneider, Christian Weinert, and Udi Wieder. "Efficient Circuit-based PSI via Cuckoo Hashing"

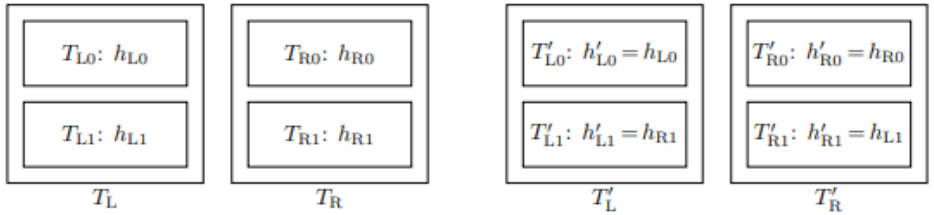
- Support pour plusieurs colonnes non intégré
- Très efficace (encore plus)
- Volume de communications faible
- Expérience avec le cuckoo hashing

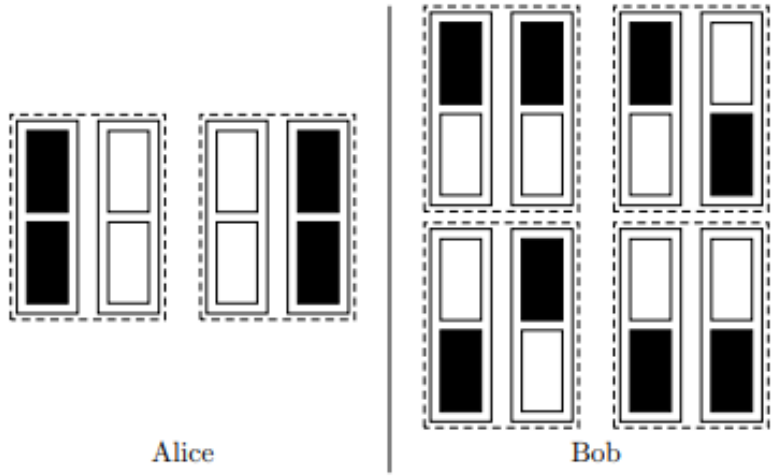
CUCKOO HASHING



MODIFICATION DU CUCKOO HASHING

- Modifier le circuit « brouillé » pour faire l'intersection sur plusieurs colonnes
- Transformer nos données pour qu'elles soient compatibles avec le circuit (re-hasher les données)





COMMENT MODIFIER L'ÉTAT DE L'ART POUR FAIRE DU SRL ?

- Récupération des index lors d'associations dans le PSI
- Comment établir l’égalité entre deux éléments ?
- Recherche d'une formule

UNE PREMIÈRE FORMULE

$$\text{Match(x)} = (1.0 * \text{SSN} + 0.2 * \text{First Name} + 0.4 * \text{Last Name} + 0.1 * \text{Birth Date} + 0.7 * \text{Phone} + 0.4 * \text{Address} + 0.4 * \text{Email}) \geq 1.0$$

Associations justes : 86.9%
Associations faussées : 0.0035%
Associations manquées : 13.1%

LA FORMULE FINALE

$$\begin{aligned}\text{Match(x)} = \text{SSN} \quad &\| \quad (FirstName \,\&\&\, LastName \,\&\&\, (Address \,\|\, BirthDate \,\|\, Email)) \\ \| \quad &(Phone \,\&\&\, (FirstName \,\|\, LastName \,\|\, Address \,\|\, Email \,\|\, BirthDate)) \\ \| \quad &(Email \,\&\&\, Address \,\&\&\, (FirstName \,\|\, LastName)))\end{aligned}$$

NOS RÉSULTATS

- Intersection entre 10% d'un premier ensemble de patients (50'000) et 100% d'un second ensemble de patients (500'000)
- Associations justes : 94.7%
- Associations faussées : 0.17%
- Associations manquées : 5.3%
- Un faible taux de fausses associations

NOTRE IMPLÉMENTATION

- idash: Bibliothèque python, qui peut être utilisée comme un outil sur l'invite de commande
- libcuckoo: Adaptation de l'implémentation du papier Cuckoo pour supporter plusieurs colonnes (C++)
- pycuckoo: Bindings python pour pouvoir utiliser libcuckoo dans l'outil idash
- Un produit final qui permet d'obtenir des résultats intéressants

CONCLUSION

Etat de l'art

- Aucun papier sur le Secure Record Linkage : c'est un problème ouvert
- De plus en plus de papiers sur le PSI, avec des perfs de plus en plus intéressantes
- Le PSI ne sert qu'a récupérer l'intersection $X \cap Y$
- Circuit-PSI : quelques papiers intéressants $f(X \cap Y)$

Notre travail

- Récuperation des index dans le PSI
- Application d'une formule sur les index de différents PSI
- Recherche d'une formule pour l'égalité entre patients

BIBLIO

- Rasmus Pagh, and Flemming Friche Rodler. "Cuckoo Hashing - BRICS"
- Benny Pinkas, Thomas Schneider, and Michael Zohner. "Scalable Private Set Intersection Based on OT Extension"
- Michele Ciampi, and Claudio Orlandi. "Combining Private Set-Intersection with Secure Two-Party Computation"
- Jason H. M. Ying, Shuwei Cao, Geong Sen Poh, Jia Xu, and Hoon Wei Lim. "PSI-Stats: Private Set Intersection Protocols Supporting Secure Statistical Functions"
- Peter Rindal and Phillipp Schoppmann. "VOLE-PSI: Fast OPRF and Circuit-PSI from Vector-OLE"
- Peter Rindal and Phillipp Schoppmann. "VOLE-PSI: Fast OPRF and Circuit-PSI from Vector-OLE"