Scalable Privacy-Preserving Data Integration

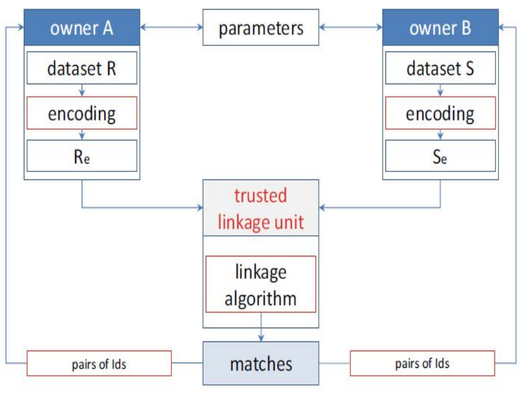
Ninth European Big Data Management & Analytics Summer School (eBISS 2019) Berlin, Germany
Ainhoa Zapirain, René Gómez

Privacy-Preserving Record Linkage

Goal: Identify and **link** records that refer to the **same entity**/individual **in different data sources** without compromising **privacy** and **confidentiality** of these entities/individuals

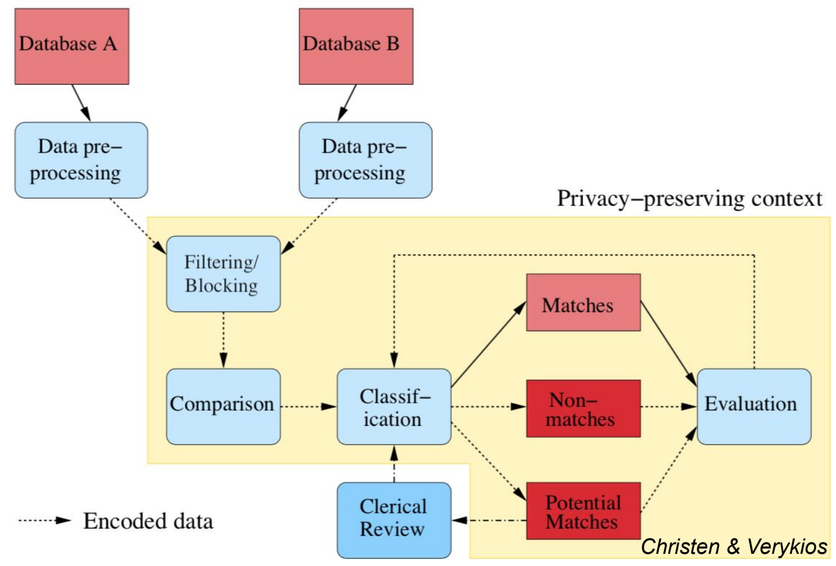
- ➔ **Anonymization:** removing, generalizing or changing quasi-identifiers **reduces** data utility
- ➔ **Pseudonymization:** Encode quasi-identifiers to **preserve privacy** & **data utility**.

PPRL Protocols



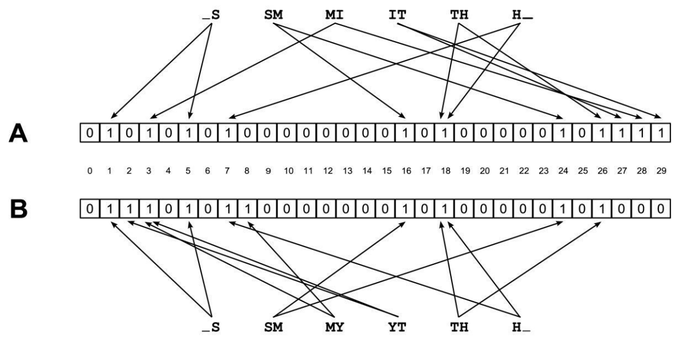
Leipzig Database Group

PPRL Process



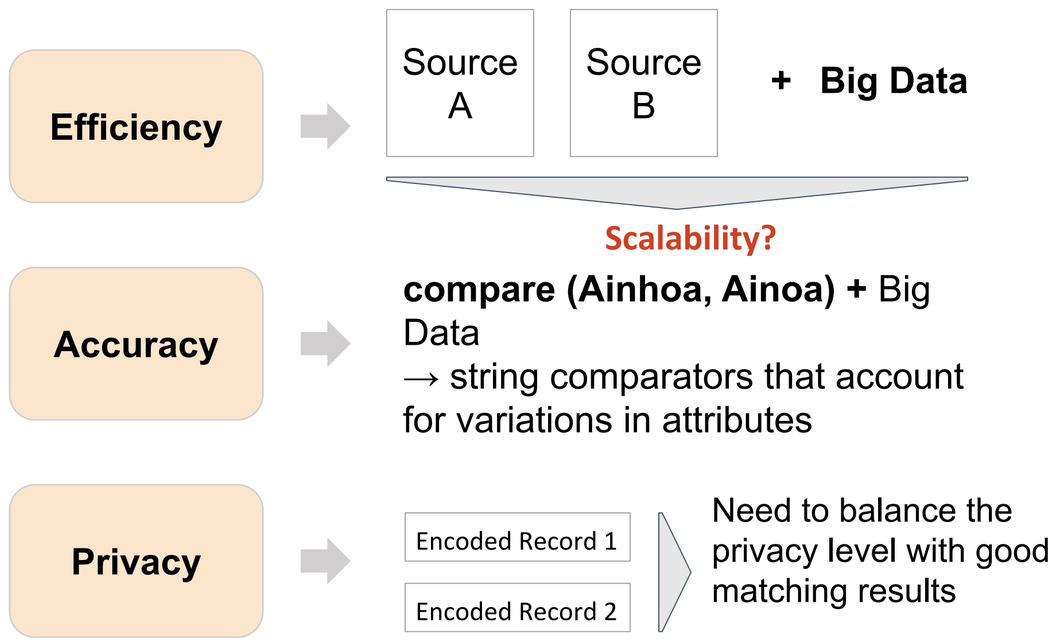
Data encoding

- **1-way** encoding preventing **re-identification** and **discrimination**
- Cryptographic-based, Perturbation-based (k-anonymity, **Bloom Filters** → scalable), Hybrid approaches.
- **Bloom Filters:** Splits each QID into q-grams and use different hash functions to create Bloom filter. The Dice-coefficient is applied to check the similarity. *Example checking similarity of SMITH and SMYTH*



Schnell, Rainer, Tobias Bachteler, and Jörg Reiher. 2009. "**Privacy-Preserving Record Linkage Using Bloom Filters.**" *BMC Medical Informatics and Decision Making*..

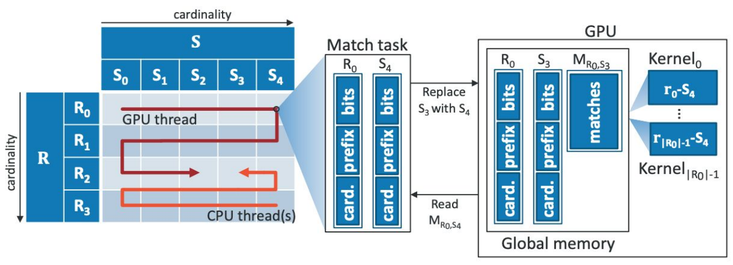
PPRL Challenges



Vatsalan, Dinusha, Ziad Sehili, Peter Christen, and Erhard Rahm. 2017. **"Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges."** In *Handbook of Big Data Technologies*, 851–95. Springer, Cham.

Parallel PPRL with GPUs

- Use **GPUs** to **speed up** the **computation** to link the data. **P4Join** is used as filtering technique in the GPU
- **Trade off** → GPUs have **smaller memory**, therefore the tasks need to be divided and executed multiple times
- **Solution** → Hybrid CPU/GPU

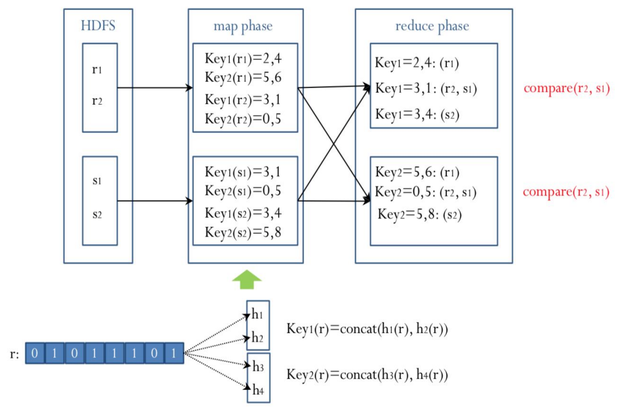


- **Results** → Using the **hybrid** CPU/GPU **approach**, there is an **improvement** of **10-20%** comparing with the approach of using the GPU only

Sehili, Ziad, Lars Kolb, Christian Borgs, Rainer Schnell, and Erhard Rahm. 2015. ***Privacy Preserving Record Linkage with PPJoin.*** Gesellschaft für Informatik e.V.

Parallel PPRL with Hadoop Ecosystem

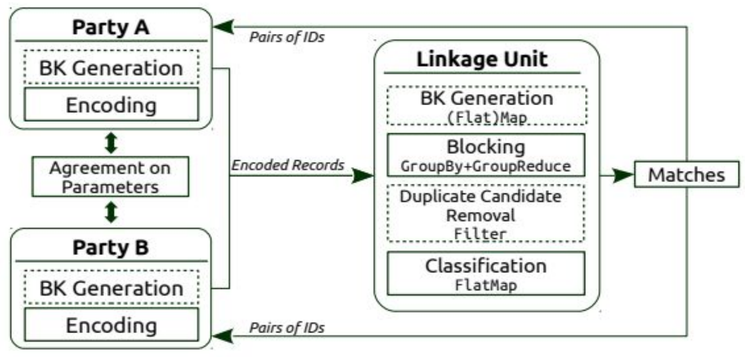
- **HDFS** → to store large datasets
- **MapReduce** → links records
- LSH blocking method with **MinHash**



- If **two values** are associated to **one specific key** → chain two MapReduce jobs
  - First MapReduce → Bloom filter **identifiers** will be **redistributed** and the data will be store in a RDBMS
  - Second MapReduce → link the records and **output** the pair records **identifiers**

Parallel PPRL with Apache Flink

- Apache Flink
- LSH blocking method with HLSH



- Data owners → encode their data
- Linkage unit → generates the linkage
  - If a record does not have a key → LU generates one key
  - **Stores** these **data** in **HDFS**, distributing and replicating among nodes
  - **Group** records by **block key**
  - The matching IDs are sent to their owners

Franke, Martin, Ziad Sehili, and Erhard Rahm. 2018. "**Parallel Privacy-Preserving Record Linkage Using LSH-Based Blocking.**" In *3rd International Conference on Internet of Things, Big Data and Security*, 195–203.

Applications

- **Healthcare** → *«statistical information would become more meaningful because it would be linked to other types of data» Halbert Dunn (1946)*
- **City Analytics** → ***Sensitive*** *data is* ***essential*** *to* ***bring data-driven intelligence*** *to cities.*
- **PIMS** → *PIMS allow individuals to manage their personal data in* ***secure,*** *local or online storage systems and* ***share them when and with whom they choose***

Conclusions

- **Challenges balancing conflicting requirements** → high privacy, link quality and scalability to large datasets
- Big Data **variety reduces** the **match quality** → difficult to manage with **multi-party** integration
- There are no **standard** methods to **evaluate** the **privacy level** of PPRL protocols

Vatsalan, D., P. Christen, and E. Rahm. 2017. "Scalable Privacy-Preserving Linking of Multiple Databases Using Counting Bloom Filters." In *ICDM Workshop on Privacy and Discrimination in Data Mining (PDDM)*. IEEE. https://ieeexplore.ieee.org/document/7836761.