IEEE BigData 2023, Sorrento, Italy

Special Session on Privacy and Security Big Data

Privacy-Preserving Data Integration [Vision Paper]

Domenico Beneventano , Sonia Bergamaschi, Lisa Trigiante

Università degli Studi di Modena e Reggio Emilia, Modena, Italy
{name.surname}@unimore.it
https://www.linkedin.com/company/dbgroup-unimore

Outline

- Data Integration
- Privacy-Preserving Data Integration
- Privacy-Preserving Schema Matching
- Privacy-Preserving Record Linkage
- Privacy-Preserving (Virtual) Data Fusion

Data Integration

Data Integration (DI) is the task of **creating an accurate and unified representation of data** that resides in multiple autonomous data sources

Schema Matching

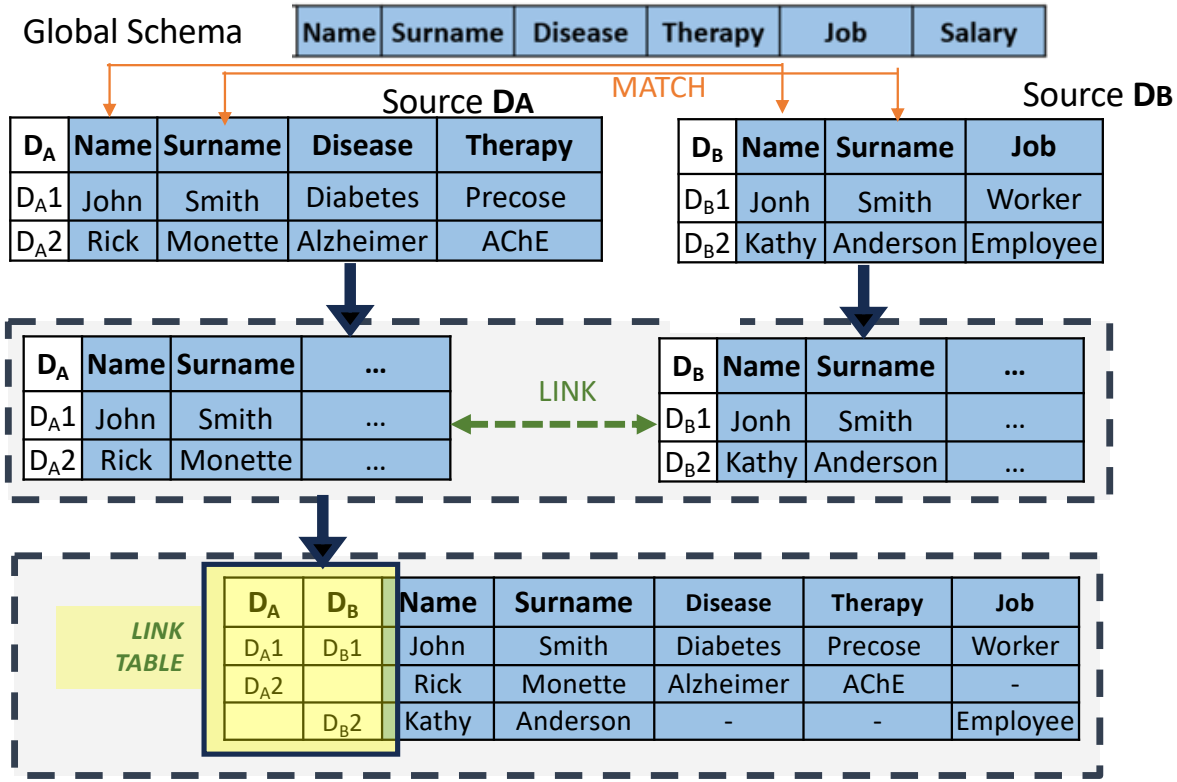
To **match attributes** of Local Sources and produce a **Global Schema**

Record Linkage

To **link records about the same real-world entity** from Local Sources.

Data Fusion

To **solve conflicts** and produce a **unique record for each real-world entity**



Privacy-Preserving Data Integration

Privacy-Preserving Data Integration (PPDI) is the task of **creating an accurate and unified representation of personal data** of multiple data sources while **preventing privacy disclosure of individuals** represented in the underlying data.

- **Personal data classification**: various types of personal data need to be managed differently to ensure the protection of the identity and sensitive data of individuals

- **Quasi-identifiers (QID)**: information that potentially identifies record owners when joined with other information; e.g., names, dates of birth, addresses.

- **Sensitive Personal Information (SPI)** contains confidential personal information that must be protected from privacy disclosure; e.g., disease or income, religion or political opinions.

	QID		SPI	
D_A	**Name**	**Surname**	**Disease**	**Therapy**
D_A1	John	Smith	Diabetes	Precose
D_A2	Rick	Monette	Alzheimer	AChE

Schema Matching

- **Schema Matching**: Automatically or semi-automatically discover correspondences between schemata.

- **Standard schema matching** methods are:
  - **Label-based** Methods: Rely on the names of schema elements
  - **Instance-based** Methods: Compare the actual data values

- We developed and implemented both such methods

❖ **MOMIS** is a Data Integration system which implements mainly label-based **schema matching** based on the **annotation** of attributes w.r.t. a thesaurus (such as Wordnet or domain specific)

❖ **SparkER** is an Entity Resolution framework developed for Apache Spark which also implements efficient instance-based **schema matching** technique for heterogeneous big data sources

Privacy-Preserving Schema Matching

- Within the **privacy context**, accessing the plain schema poses minimal risks and then **traditional schema matching methods** are generally used (very few techniques in literature) … ***nevertheless*** …

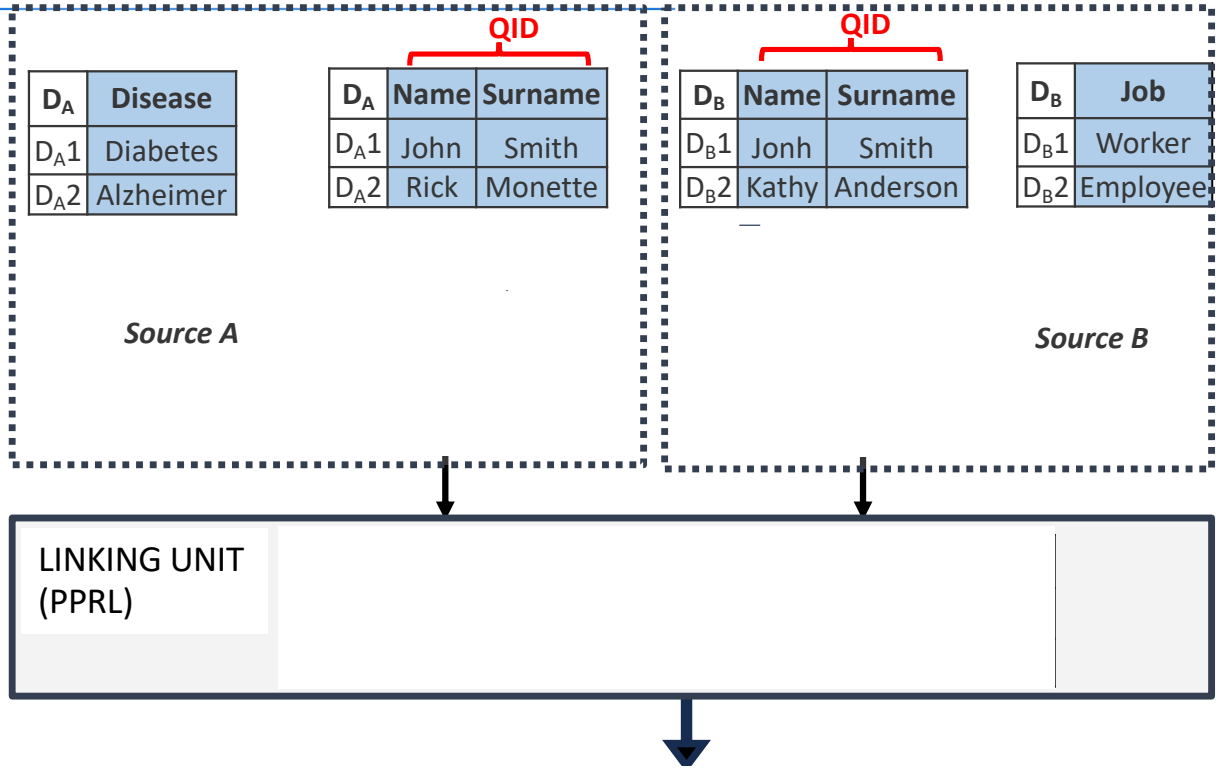
- Traditional methods analyze the entire local schema, while in PPDI, different types of personal data need to be managed differently.
  - **Our vision**: To enhance the overall PPDI process, it is convenient to explore schema alignment techniques based on the different classifications of schema elements (QID vs, SPI)

- It is a common practice to manually pre-determine a subset of QID coE first point Is that Fest point Is Right provammon to all sources. However, this is only feasible for low-dimensionality schemata and not in the context of Big Data.
  - **Our vision**: Using additional schema information, obtained, for example, through annotation, it is possible to automatically classify attributes in QID and SPI.

- Some PPSM methods proposed in the literature aim to prevent privacy disclosure of data by applying only schema-level matching, i.e. consider only the attribute names and their associated descriptions.
  - **Our vision:** Using pseudonyms instead of clear values, it is possible to apply instance-based matching without revealing any private information about the individual.

Privacy-Preserving Record Linkage (PPRL)

- To link records about the same individual without revealing any private, sensitive information
- PPRL is mainly based on the **Pseudonymization** of QIDs (*encoded values*)
  ➔ **GDPR requirement**
- **QIDs are sent to the linking unit as pseudonyms**, no clear plain text personal information is transmitted.
- **Pseudonym-based matching** ensures privacy preservation: the output is a link table with no private, sensitive information



General Data Protection Regulation (GDPR)

- Whenever sensitive personal data about individuals are to be integrated, privacy and confidentiality have to be considered.

- Data protection in Europe is set off by the European General Data Protection Regulation (GDPR) which became active in May 2018 and is a comprehensive legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU).

- An Appropriate technique to implement data-protection principles in an effective manner is Pseudonymization.

This applies to the use of tolerant **privacy-preserving techniques** to create **pseudonyms** of the data to be integrated.

Privacy-Preserving Record Linkage (PPRL)

The "Standard" PPRL Approach

1. **Pseudonymization**: crucial for privacy; transforms QID into pseudonyms (**encoded values**) by using one or more encoding function, for example Bloom Filters.
2. **Blocking**: crucial for scalability; reduces the number of comparisons needed between encoded values and generates candidate encoded record pairs
3. **Comparison and Classification**: Crucial for linkage quality; involves comparing candidate encoded record pairs and classifying them into match or not match.

- In the context of PPRL, blocking can either be conducted **locally** (at each data sources), on clear plain text or **globally** (at the linking unit) - on encoded values.
- In the context of **Big Data**, it is crucial to perform **blocking locally** to reduce the amount of data that needs to be transferred from each source to the linking unit.
  - ❖ **SparkER** is an Entity Resolution framework developed for Apache Spark which implements advanced blocking techniques (token blocking, meta blocking, …) for big data sources

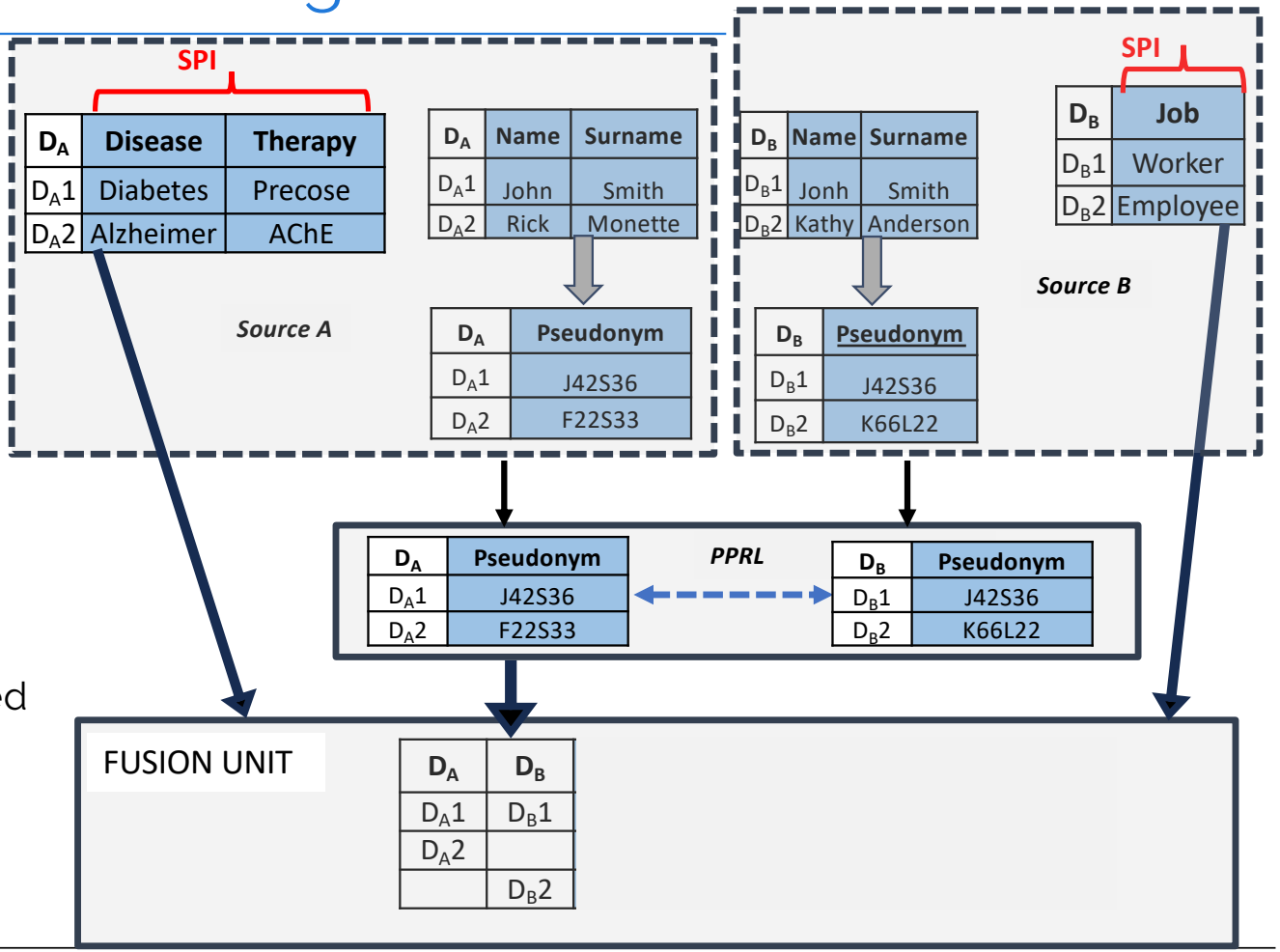
➢ **Our vision :** advanced local blocking techniques can be adapted to the Big Data privacy context to optimize the PPRL process.

Privacy Preserving Data Fusion

- Data Fusion is the process of merging linked (duplicated) record from sources into a single unified record.
  - Standard Approach: **Conflict Resolution Functions** to choose most promising values

- In the **privacy** context, Data Fusion is generally performed only on **Sensitive Personal Information (SPI)**,

- Privacy-preserving data publishing techniques, such as **k-anonymity**, need to be applied to the fused dataset to prevent any reidentification.



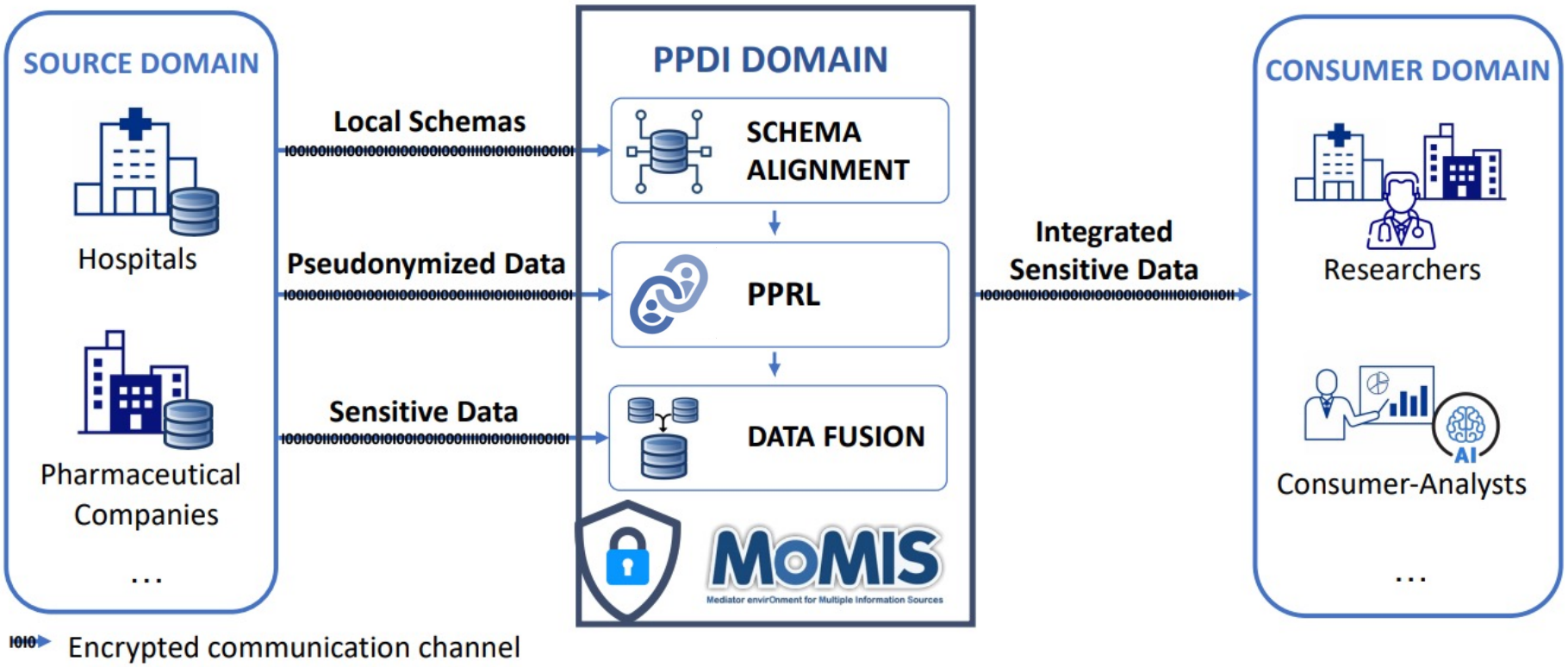
Privacy-Preserving Virtual Data Fusion

- **Virtual Integration**: leave the data at the sources and access it at query time by supporting query over the integrated schema and by applying online **Query Reformulation**
  - A query is transformed in a set of sub-queries, one for each data source
  - The results are collected by a mediator, merged and shown to the user.

- MOMIS is a Virtual (Big) Data Integration system which implements Query Reformulation techniques to perform Data Fusion based on Conflict Resolution Functions
  - Developed at the DBGroup, made available as open source by DataRiver

➢ **Our Vision**: within the **privacy** context *Conflict Resolution Functions* and *Query Reformulation* techniques can be extended for ensuring K-Anonymity

PPDI framework for Health



THANK YOU FOR YOUR ATTENTION