Private Synopses-Driven Data Integration

Thesis Progress Report Presentation

PhD Candidate: Eros Fabrici
Supervisor: Prof. Minos Garofalakis
Co-Supervisors: Prof. Josep Lluís Berral-García, PhD Besim Bilalli

Athena Research Center & Universitat Politècnica de Catalunya

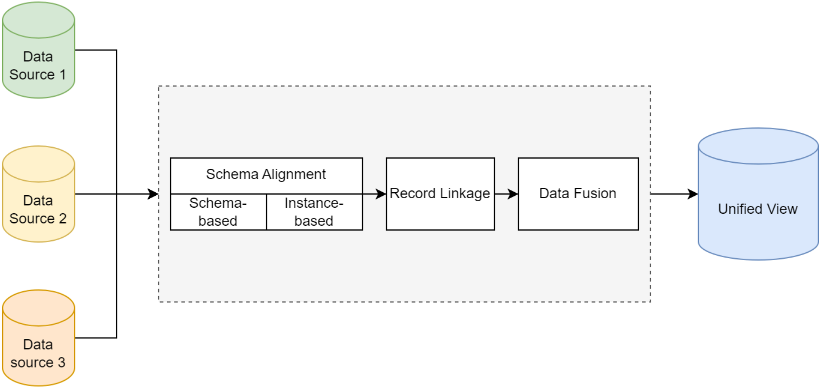
Contents

Motivation

- Big data $\rightarrow$ various challenges in data management
- Privacy breaches over the last decade
  - Need of new regulation $\rightarrow$ GDPR
  - Increase of work in the research fields
    - **Privacy Preserving Big Data Mining/Analytics**
    - **Privacy Preserving Data Synthesis/Release**
    - **Privacy Preserving Machine Learning**
    - **Federated Learning**
- A common pre-processing step is *Data Integration*

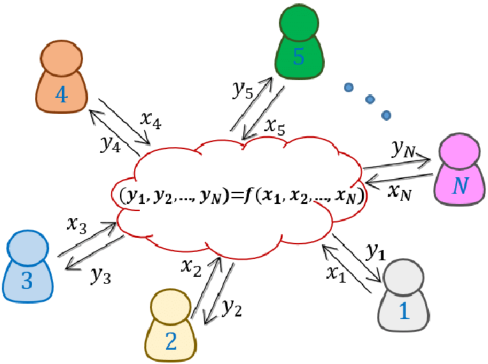
Data Integration

Data Integration is the process of bringing different disparate sources into a unified view



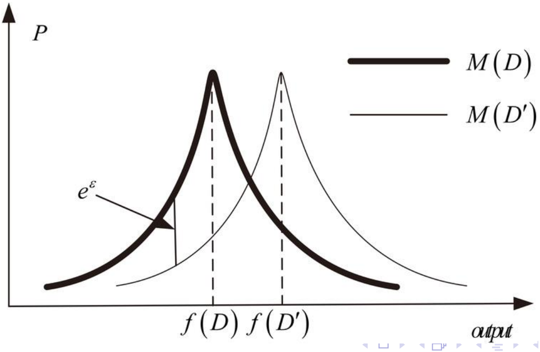
Secure Multi-Party Computation

- Models for parties to jointly compute a function over their input while keeping those inputs private
- Sub-field of cryptography



Differential Privacy

- State-of-the-art standard for Big Data Analytics/Machine Learning
- Learn nothing about an individual while learning useful information about the population



Privacy-Preserving Data Integration

- Tutorial on the state of the art
  1. Categorization on techniques used in PPRL

Table 1: Comparison of Private Record Linkage algorithms

	Paper	Privacy Guarantee: Blocking/Filtering	Privacy Guarantee: Matching	3^{rd} P.[1]	SA[2]	Matching function
Formal Privacy	[5, 32]	n.a.	DP[3] (RR[4] and Embedding)	✓	✗	Dice, E.[5]
	[2, 3, 12, 21]	n.a.	SMC[6] and Cryptography	✓	✗	TFIDF , Any, EM[7]
	[14, 17, 18, 27]	DP, k-anonymity	SMC	✓	✓	Dist. Based
Ad-Hoc Privacy	[10, 20, 31]	n.a.	Hashing (Phonetic, BF[9])	✓	✗	EM, Dice
	[30, 34]	n.a.	Embedding (Complex P.[8], SparseMap [16])	✓	✓	Dist. Based

[1] 3^{rd} Party; [2] Schema-Aware; [3] Differential Privacy; [4] Randomized Response; [5] Euclidean Distance; [6] Secure Multiparty Computation; [7] Exact Matching; [8] Complex Plane; [9] Bloom Filter

- *Limited work on PP-Schema alignment*
  - Can sketch algorithms and DP be applied?

Privacy-Preserving Schema Alignment

In an instance-based SA scenario, parties only learn the set of matching attributes

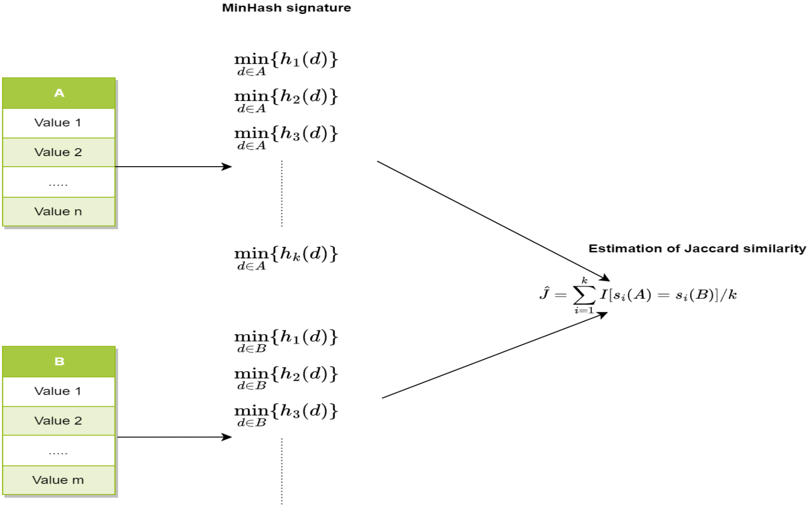
- A basic approach is to use encrypted hashes [7]
- Other approaches use *information theoretic measures* [5] or *sketches* [1, 6, 4]

Contact: Eros Fabrici (eros.fabrici@athenarc.gr)

Sketches for estimating joint quantities

- Sketches are probabilistic data structures that compress the data and permit to estimate statistics over the data with guaranteed error bounds
- Sketches that are related to DI are:
  1. MinHash (or Bottom-k)
  2. HyperLogLog
  3. Linear Sketches (e.g. random projection on qgrams)

MinHash


MinHash signature
A
Value 1
Value 2
.....
Value n
$\min_{d \in A}\{h_1(d)\}$
$\min_{d \in A}\{h_2(d)\}$
$\min_{d \in A}\{h_3(d)\}$
$\min_{d \in A}\{h_k(d)\}$
Estimation of Jaccard similarity
$\hat{J} = \sum_{i=1}^{k} I[s_i(A) = s_i(B)]/k$
B
Value 1
Value 2
.....
Value m
$\min_{d \in B}\{h_1(d)\}$
$\min_{d \in B}\{h_2(d)\}$
$\min_{d \in B}\{h_3(d)\}$

HyperLogLog

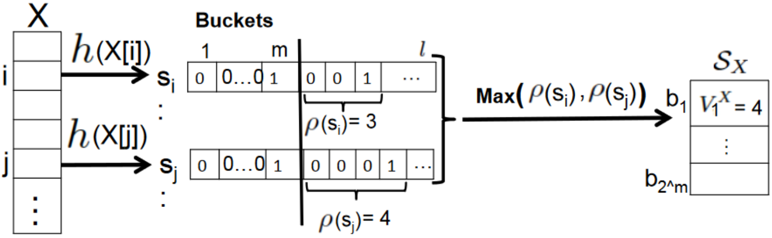
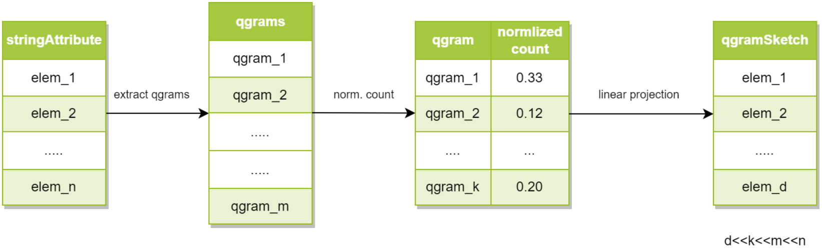


Figure: Contructing the HLL sketch of the column X as S_X [6]

Qgram sketches

- [1] use also Qgram sketches for finding similarities between columns.



Sketches and DP

- Estimating Jaccard and other joint quantities with HLL sketches make use of MLE, which is possible to make DP by using the *subsample and aggregate* technique [3]
  - [2] showed that hash-based order invariant cardinality estimators are DP
  - subsample with probability $1 - e^{-\epsilon}$ and real cardinality big enough
- Linear sketches are DP by initializing them with Gaussian noise [8]

Research Objectives

- Evaluation on how DP affects instance-based schema alignment
  - Implementation of a framework for running experiments, starting from Private Instance-Based Schema Matching
- Propose a set of algorithms for private instance-based schema alignment
  - Use of sketches for tackling computational performance limitations
  - Use of DP for privacy guarantees
- Implement a proof of concept
  - Potential use case: Data Integration for Federated Learning

Contact: Eros Fabrici (eros.fabrici@athenarc.gr)

Framework for PPDI

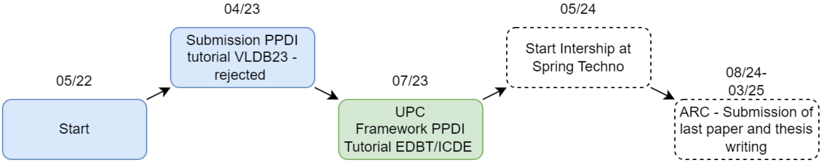
Implement a **framework** for testing sketches for data-driven PPDI

- **Schema Alignment**
  - Implement state-of-the-art sketches and run empirical experiments on mining schemas
  - Distribute a global privacy budget efficiently
    - formal framework for distributing the privacy budget across different sketches in an efficient way

Publication Plan

1. *Privacy-Preserving Data Integration: a Tutorial*
    - Tutorial Paper
    - Venue: ICDE 2024
2. *Private Sketch-based, Instance-based Schema Alignment: an Empirical Evaluation*
    - Conference Paper
    - Venue: EDBT 2024
3. *Privacy-Preserving Alignment of Schemas: an End-to-End Protocol*
    - Conference Paper
    - Venue: VLDB/SIGMOD 2024
4. *Private Data Integration for Federated Learning*
    - Demo Paper
    - Venue: VLDB/SIGMOD 2025

Timeline


05/22
Start
04/23
Submission PPDI tutorial VLDB23 - rejected
07/23
UPC Framework PPDI Tutorial EDBT/ICDE
05/24
Start Intership at Spring Techno
08/24-03/25
ARC - Submission of last paper and thesis writing

References I

[1] Tamraparni Dasu et al. "Mining Database Structure; or, How to Build a Data Quality Browser". In: *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*. SIGMOD '02. Madison, Wisconsin: Association for Computing Machinery, 2002, pp. 240–251. ISBN: 1581134975. DOI: 10.1145/564691.564719. URL: https://doi.org/10.1145/564691.564719.

[2] Charlie Dickens, Justin Thaler, and Daniel Ting. "Order-Invariant Cardinality Estimators Are Differentially Private". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 15204–15216. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/623307df18da128262aaf394cdcfb235-Paper-Conference.pdf.

[3] Cynthia Dwork and Jing Lei. "Differential Privacy and Robust Statistics". In: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. STOC '09. Bethesda, MD, USA: Association for Computing Machinery, 2009, pp. 371–380. ISBN: 9781605585062. DOI: 10.1145/1536414.1536466. URL: https://doi.org/10.1145/1536414.1536466.

[4] Otmar Ertl. "SetSketch: filling the gap between MinHash and HyperLogLog". en. In: *Proc. VLDB Endow.* 14.11 (July 2021), pp. 2244–2257. ISSN: 2150-8097. DOI: 10.14778/3476249.3476276. URL: https://dl.acm.org/doi/10.14778/3476249.3476276 (visited on 10/05/2022).

[5] Jaewoo Kang and Jeffrey F Naughton. "On Schema Matching with Opaque Column Names and Data Values". en. In: (), p. 12.

[6] Azade Nazi et al. "Efficient estimation of inclusion coefficient using hyperloglog sketches". en. In: *Proc. VLDB Endow.* 11.10 (June 2018), pp. 1097–1109. ISSN: 2150-8097. DOI: 10.14778/3231751.3231759. URL: https://dl.acm.org/doi/10.14778/3231751.3231759 (visited on 12/15/2022).

References II

[7] Monica Scannapieco et al. "Privacy preserving schema and data matching". en. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*. Beijing, China: ACM Press, 2007, p. 653. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247553. URL: http://portal.acm.org/citation.cfm?doid=1247480.1247553 (visited on 11/14/2022).

[8] Fuheng Zhao et al. "Differentially Private Linear Sketches: Efficient Implementations and Applications". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 12691–12704. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/525338e0d98401a62950bc7c454eb83d-Paper-Conference.pdf.