Private Synopses-Driven Data Integration

Eros Fabrici @ Athena Research Center & UPC,
eros.fabrici@athenarc.gr

Overview

- Data Integration (DI) is a long-standing research topic
- Privacy in Data Integration is a more recent topic, where the focus is to make the process private across the different parties involved.
- The literature focused in Record Linkage.
- The objective of this PhD is to explore the usage of *sketches* and *differential privacy* for inferring the relationships between the data in a data-driven fashion to be applied on DI process.

Privacy-Preserving Data Integration

The privacy threats in the DI lie in the first two phases. In particular, there is very limited research in privacy-preserving instance-based schema alignment. Privacy-Preserving Record Linkage (PPRL) is the most studied problem. Table 1 shows a categorization of the approaches propesed in the literature for PPRL.

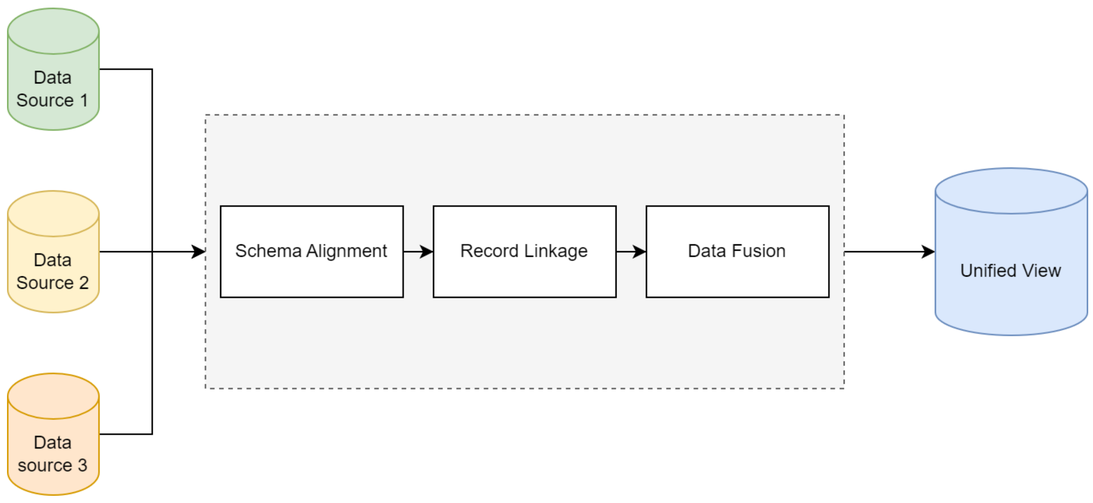
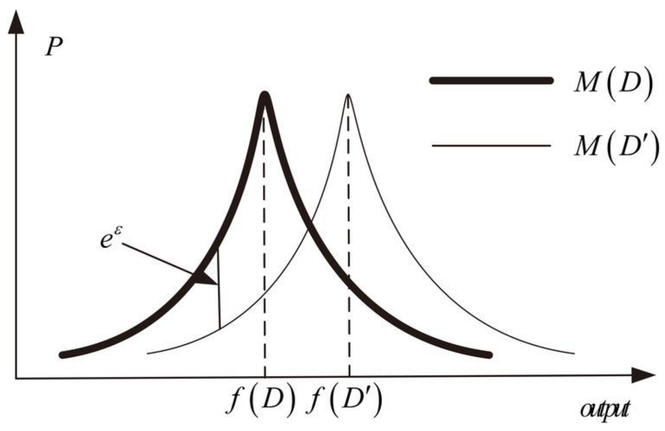


Figure 1: Data Integration Phases

Differential Privacy

Differential Privacy is a technique for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset. It is *de facto* standard in the field of privacy for ML, synthethic data and querying.



Sketches for estimating joint quantities

Sketches are probabilistic data structures to summarize the data and to obtain estimations on pre-defined queries (e.g. cardinality estimation) with theoretical bounds. The two main families of interests are *order-invariant hash based* for cardinality and joint quantities estimation and *linear sketches* for estimating distributional frequency.

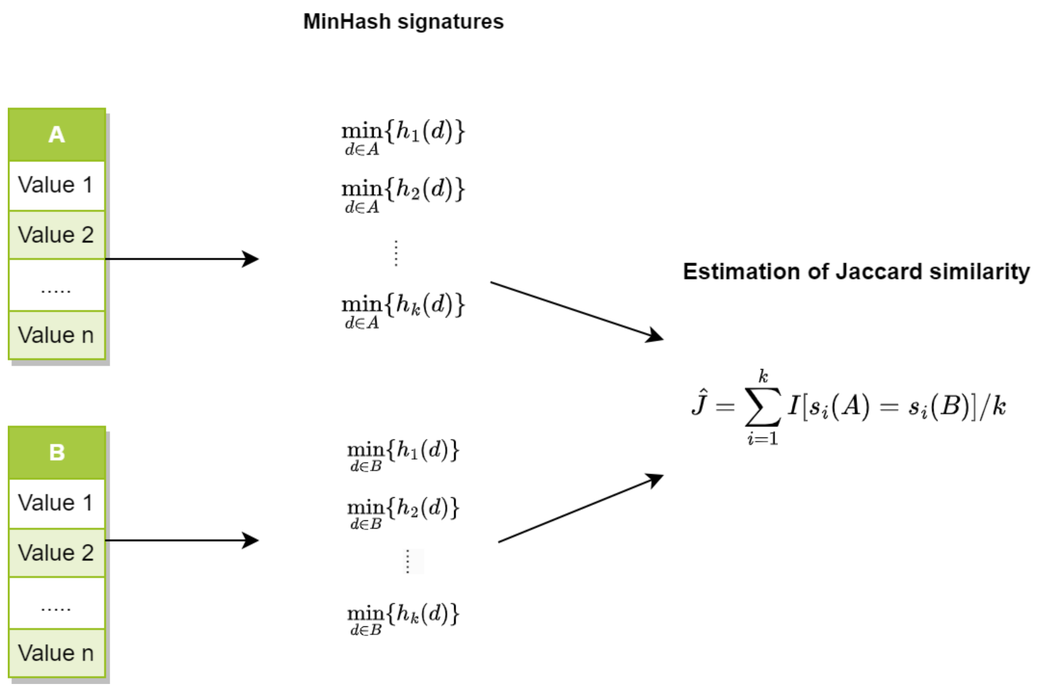


Figure 2: Estimating Jaccard similarity with MinHash

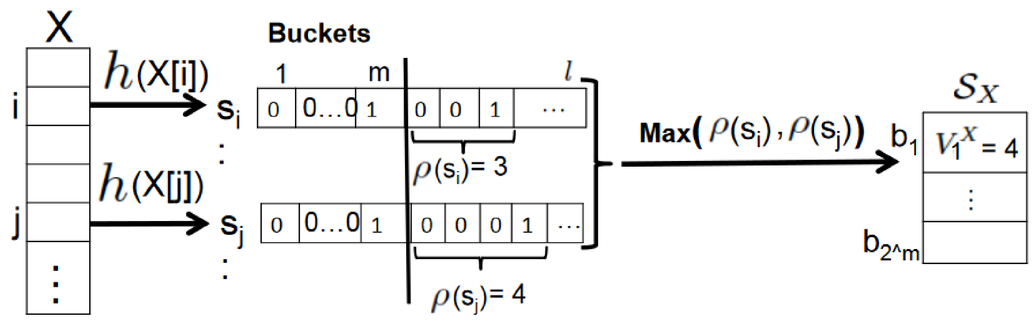


Figure 3: HyperLogLog, for cardinality and joint quantities estimations

Current Directions

The current direction is to implement a framework for a fully data-driven, private DI, where state-of-the-art sketches will be tested. An empirical study will be carried out to find the best sketches for the purpose. Efficient privacy budget allocation tailored to the DI problem will be studied as well. The aim of the tool is to infer mappings across the datasets' attributes, while guaranteeing privacy.

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