Synopses-Driven Data integration & Federated Learning

Doctorate Project Plan 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

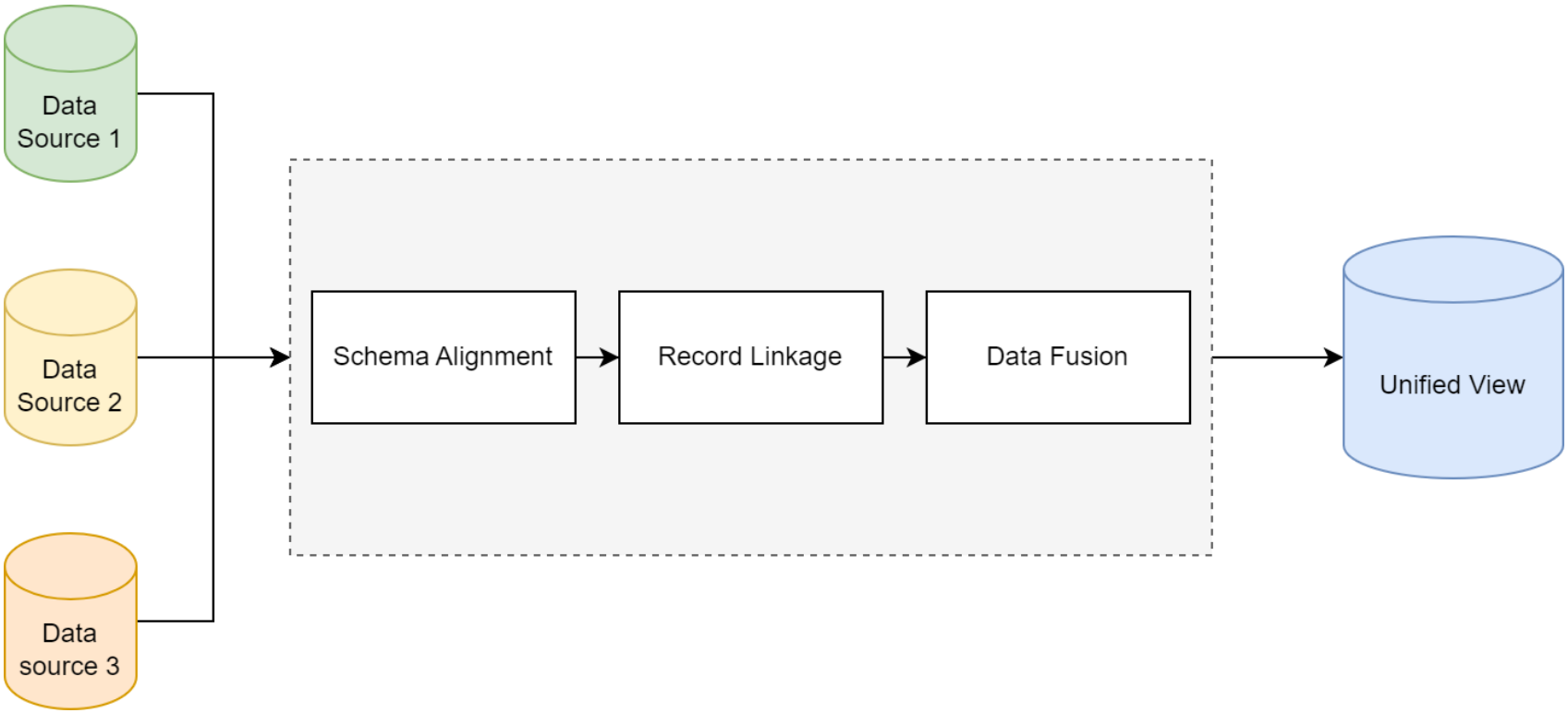
Privacy

- 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**

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

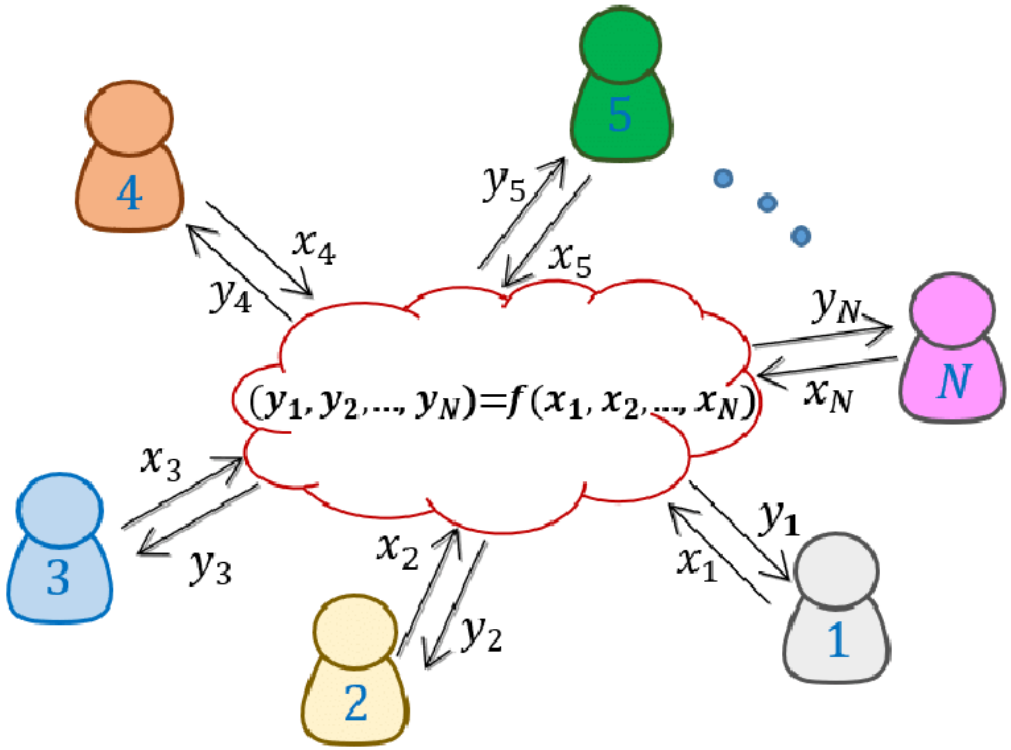
Data Integration

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



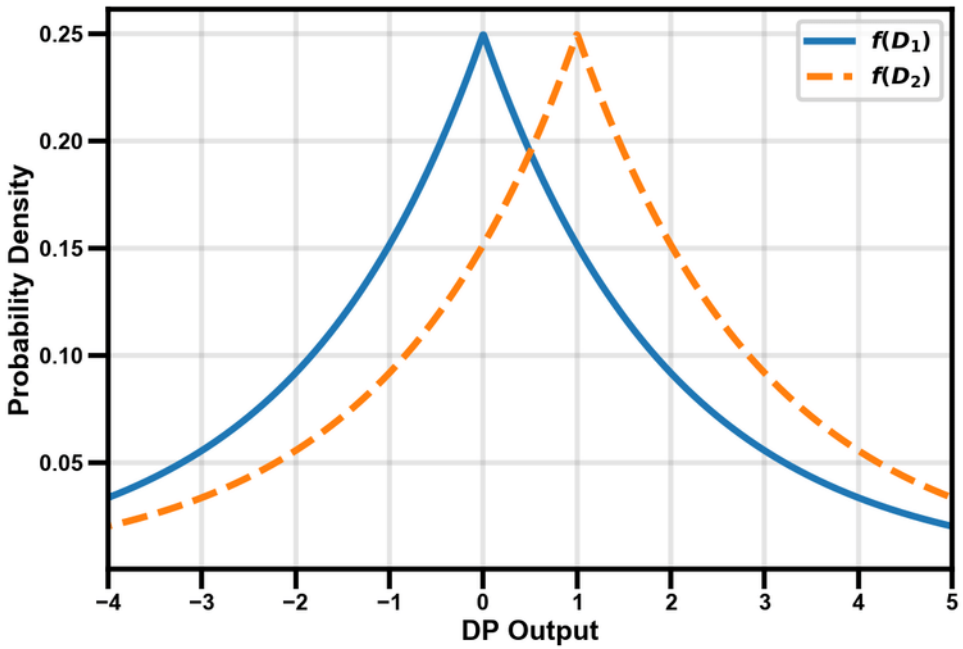
Secure Multi-Party Computation

- 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 Schema Alignment

Parties learn no information regarding other parties' schemas, except for the matching attributes

- A basic approach is to use encrypted hashes [16]
- Other approaches use *probabilistic* [12] or *data summaries* [6]

Takeaways

- Very limited research for privacy-preserving schema alignment
- exploring DP and the usage of data summaries for matching schema is an open research question

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

Privacy-Preserving Record Linkage

During the RL phase, parties only learn the set of matching records

- Works can be grouped into 4 categories:
  - Masking/Embedding
  - SMC/Cryptography
  - Differential Privacy
  - Hybrid

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

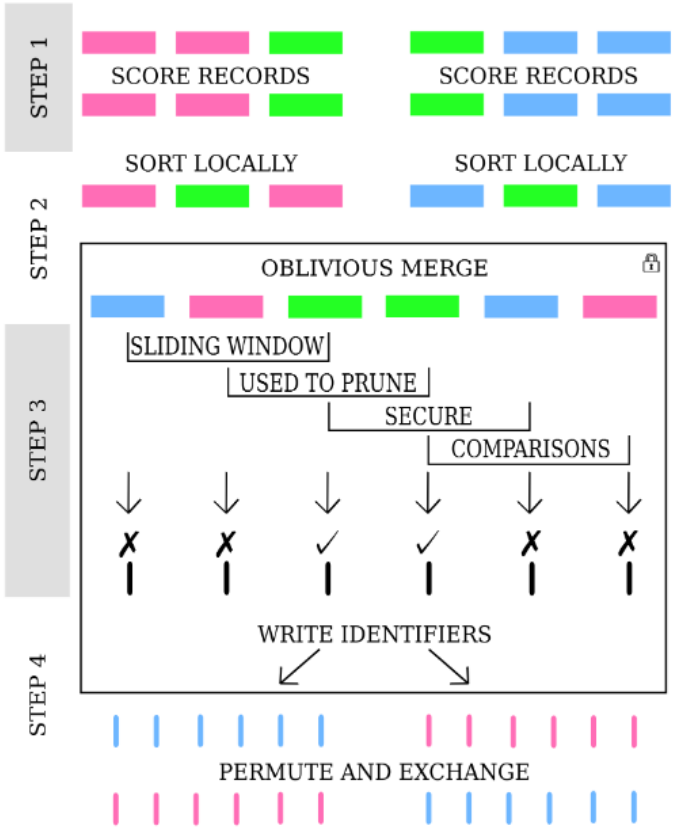
Masking and Embedding

- Masking: usage of probabilistic data structures → **Bloom Filters** [7, 17]
  - Frequency attacks [3, 5, 4, 14]
- Embedding: map records to a space
  - Euclidean/Complex Spaces [19, 16] → limited accuracy
  - Distributed Representation of Tuples [8]: promising approach for linkage quality

No formal privacy guarantee

SMC and Cryptography

- Two main relevant works
- [1] uses commutative encryption
- [13] represent one the state-of-the-art works



→ Still computationally expensive

Differential Privacy

- [2] introduce the use of DP in RL
  - The idea is to generate a base for a euclidean space that is DP
  - Prefix mining algorithm: extract the top-k frequent n-grams, then add Laplace noise to the counts
- [18] use Randomized Response to perturb the BFs

Research Questions

As embedding and DP were explored, a research question would be to use modern embedding technologies (deep learning) and DP

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

Hybrid

- Combining k-anonymity and SMC [10]
- Combining DP and SMC [11, 9, 15]

→ k-anonymity doesn't guarantee a good level of privacy
→ DP+SMC provides a good level of privacy, still improvements to reduce the number of comparisons are needed

Research Questions

1. How DP affects synopses-driven algorithms for mining the structure of federated datasets?
2. With regards to Record Linkage, privacy-preserving technologies have been studied extensively, but
   - Hybrid approaches are still limited due to the expensive SMC computations $\rightarrow$ improvements in the private blocking/indexing algorithms are possible
   - deep learning embeddings brought the democratization of RL to a new level $\rightarrow$ can DP be applied to learning word embeddings in a federated scenario?

Idea - Schema Alignemnt

- Use DP mechanisms on sketching algorithms like
  - MinHash
  - Qgram sketches
- experiment utility/privacy tradeoff

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

Idea II - Schema Alignment

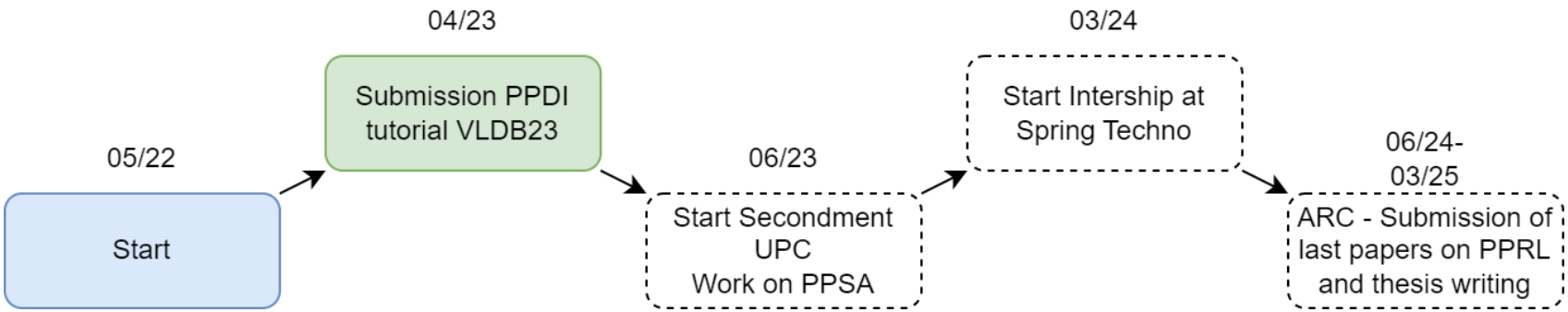
- [12] use undirected graphs representation of tables (node=entropy of column, edge=MI between two columns), then run a graph matching algorithm
- Private Synthetic Data generation makes use of Bayesian Networks (BN) [20]
- BN estimates the full distribution of a dataset, graphically as a DAG
- **Experiment with the graph matching approach with the DP BN graph**

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

Idea III - Record Linkage

- [15] use a private indexing algorithm that distributed the DP noise to the leaves
  - It is worth investigating and experimenting with more advanced indexing algorithms
- Outstanding results of DL embeddings for RL [8]
  - Study the applicability of DP on Deep Learning word embeddings to support RL and Schema Alignment

Timeline


05/22
Start
04/23
Submission PPDI tutorial VLDB23
06/23
Start Secondment UPC
Work on PPSA
03/24
Start Intership at Spring Techno
06/24-03/25
ARC - Submission of last papers on PPRL and thesis writing

References I

[1] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant. "Information Sharing Across Private Databases". en. In: ().

[2] Luca Bonomi et al. "Frequent grams based embedding for privacy preserving record linkage". en. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. Maui, Hawaii, USA: ACM Press, 2012, p. 1597. ISBN: 978-1-4503-1156-4. DOI: 10.1145/2396761.2398480. URL: http://dl.acm.org/citation.cfm?doid=2396761.2398480 (visited on 11/14/2022).

[3] Peter Christen et al. "Efficient cryptanalysis of bloom filters for privacy-preserving record linkage". In: *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I 21*. Springer. 2017, pp. 628–640.

[4] Peter Christen et al. "Pattern-mining based cryptanalysis of Bloom filters for privacy-preserving record linkage". In: *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III*. Springer. 2018, pp. 530–542.

[5] Peter Christen et al. "Precise and fast cryptanalysis for Bloom filter based privacy-preserving record linkage". In: *IEEE Transactions on Knowledge and Data Engineering* 31.11 (2018), pp. 2164–2177.

[6] Tamraparni Dasu et al. "Mining Database Structure; Or, How to Build a Data Quality Browser". en. In: (), p. 12.

[7] Elizabeth A Durham et al. "Composite bloom filters for secure record linkage". In: *IEEE transactions on knowledge and data engineering* 26.12 (2013), pp. 2956–2968.

[8] Muhammad Ebraheem et al. "Distributed Representations of Tuples for Entity Resolution". In: *Proc. VLDB Endow.* 11.11 (July 2018), pp. 1454–1467. ISSN: 2150-8097. DOI: 10.14778/3236187.3236198. URL: https://doi.org/10.14778/3236187.3236198.

References II

[9] Xi He et al. *Composing Differential Privacy and Secure Computation: A case study on scaling private record linkage*. en. arXiv:1702.00535 [cs]. Sept. 2017. URL: http://arxiv.org/abs/1702.00535 (visited on 10/10/2022).

[10] Ali Inan et al. "A Hybrid Approach to Private Record Linkage". en. In: *2008 IEEE 24th International Conference on Data Engineering*. Cancun, Mexico: IEEE, Apr. 2008, pp. 496–505. ISBN: 978-1-4244-1836-7 978-1-4244-1837-4. DOI: 10.1109/ICDE.2008.4497458. URL: http://ieeexplore.ieee.org/document/4497458/ (visited on 01/18/2023).

[11] Ali Inan et al. "Private record matching using differential privacy". en. In: *Proceedings of the 13th International Conference on Extending Database Technology*. Lausanne Switzerland: ACM, Mar. 2010, pp. 123–134. ISBN: 978-1-60558-945-9. DOI: 10.1145/1739041.1739059. URL: https://dl.acm.org/doi/10.1145/1739041.1739059 (visited on 01/18/2023).

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

[13] Basit Khurram and Florian Kerschbaum. "SFour: A Protocol for Cryptographically Secure Record Linkage at Scale". en. In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. Dallas, TX, USA: IEEE, Apr. 2020, pp. 277–288. ISBN: 978-1-72812-903-7. DOI: 10.1109/ICDE48307.2020.00031. URL: https://ieeexplore.ieee.org/document/9101375/ (visited on 11/02/2022).

[14] Martin Kroll and Simone Steinmetzer. "Automated cryptanalysis of bloom filter encryptions of health records". In: *arXiv preprint arXiv:1410.6739* (2014).

References III

[15] Fang-Yu Rao et al. "Hybrid Private Record Linkage: Separating Differentially Private Synopses from Matching Records". en. In: *ACM Trans. Priv. Secur.* 22.3 (July 2019), pp. 1–36. ISSN: 2471-2566, 2471-2574. DOI: 10.1145/3318462. URL: https://dl.acm.org/doi/10.1145/3318462 (visited on 10/05/2022).

[16] 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).

[17] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters". In: *BMC medical informatics and decision making* 9.1 (2009), pp. 1–11.

[18] Rainer Schnell and Christian Borgs. "Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage". en. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. Barcelona, Spain: IEEE, Dec. 2016, pp. 218–224. ISBN: 978-1-5090-5910-2. DOI: 10.1109/ICDMW.2016.0038. URL: http://ieeexplore.ieee.org/document/7836669/ (visited on 01/18/2023).

[19] Mohamed Yakout, Mikhail J. Atallah, and Ahmed Elmagarmid. "Efficient Private Record Linkage". en. In: *2009 IEEE 25th International Conference on Data Engineering*. ISSN: 1084-4627. Shanghai, China: IEEE, Mar. 2009, pp. 1283–1286. ISBN: 978-1-4244-3422-0. DOI: 10.1109/ICDE.2009.221. URL: http://ieeexplore.ieee.org/document/4812521/ (visited on 11/14/2022).

[20] Jun Zhang et al. "PrivBayes: Private Data Release via Bayesian Networks". en. In: ().