Privacy-Preserving Record Linkage for Big Data

Kaoutar Chennaf, Gonçalo Moreira, Alberto Abello
Universitat Politècnica de Catalunya

Abstract

With Privacy Preserving Data Mining (PPDM) between organisations and governments being on the rise, exchange methods for privacy preserving of identifying data have seen the light. **Privacy Preserving Record Linkage (PPRL)** is one of them. It allows for secure data sharing among data owners while minimizing the risk of identifying individuals.

Introduction

Nowadays, a huge amount of personal data is stored and needs to be analysed. The analysis of private personal data becomes increasingly complicated and complex, since the data come from multiple sources. In order to solve this problem, the Privacy-preserving record linkage (PPRL) was developed.

Privacy-preserving record linkage (PPRL) aims at integrating person-related data without revealing sensitive information. PPRL is being required in many real-world areas such as public health surveillance to crime and fraud detection.

Challenges of PPRL

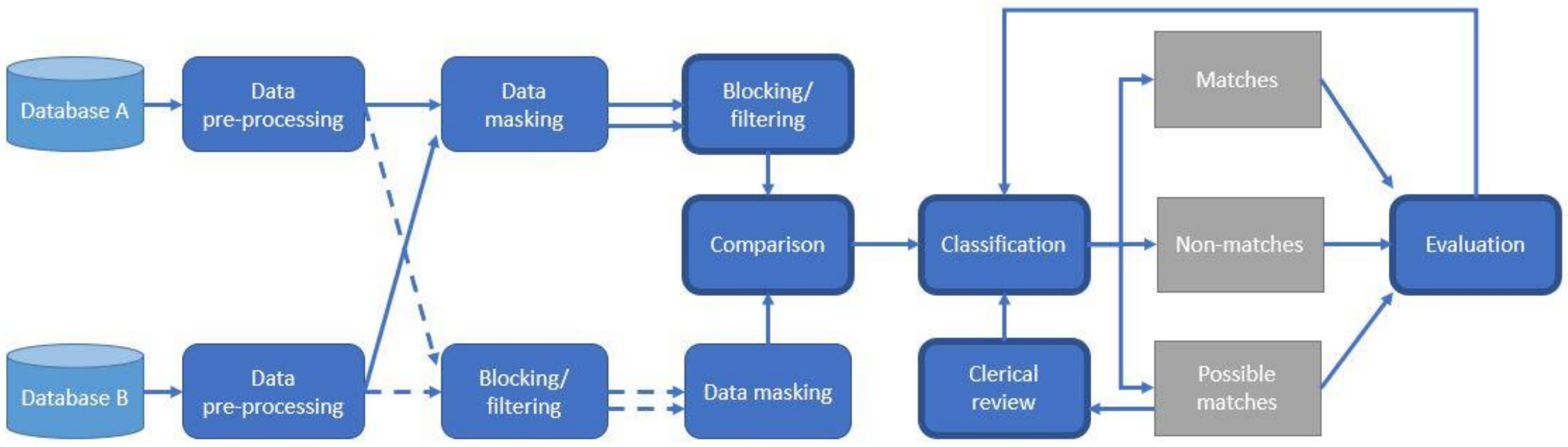
PPRL applied to Big Data poses several challenges:

- **Scalability:** the nº comparisons required for classifying record pairs or sets equals to the product of the size of the databases that are linked. This is a performance bottleneck since it potentially requires comparison of all record pairs/sets using expensive comparison functions. Due to the increasing size of Big Data (volume), comparing all records is not feasible in most real-world applications.
- **Linkage quality:** the challenge of dealing with typographical errors and other variations in data (variety and veracity).
- **Privacy:** needs to be considered in all steps as only the masked records can be used, making the task of linking databases across organizations more challenging.

Future Work

- Many of the discussed data masking techniques lack scalability which is a main challenge of PPRL and a characteristic of Big data.
- More research is required towards the development of techniques that allow for multiple large databases to be linked in privacy-preserving, effective, and efficient ways.
- Advanced classification and matching techniques for PPRL still need to be further developed to allow secure and efficient de-identification.
- There is also a requirement to develop a comprehensive architecture which combines data publisher and data recipient. In distributed environment, efficiency will pay an important role, so an efficient algorithm which tries to balance between sensitive disclosure, data utility and communication cost is required.

PPRL Process and Tools



Database A
Data pre-processing
Data masking
Blocking/ filtering
Matches
Comparison
Classification
Non-matches
Evaluation
Database B
Data pre-processing
Blocking/ filtering
Data masking
Clerical review
Possible matches

Fig 1. PPRL process

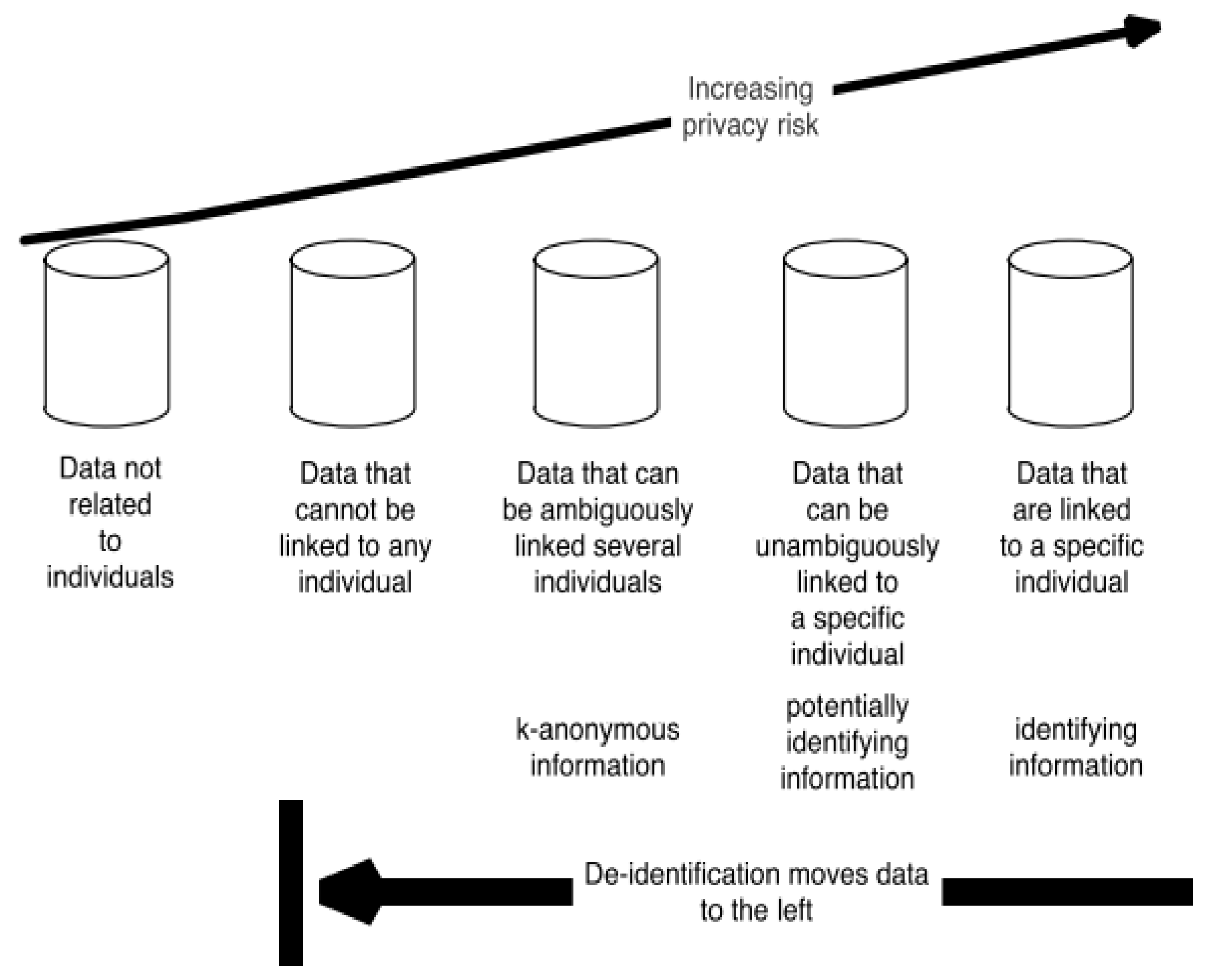
Data Masking Techniques

Terminologies:
Privacy classification of data:

- Identity Attributes
- Sensitive Attributes
- Non-sensitivie Attributes

De-identification techniques:

- Anonymization
- Pseudonymization (Data masking)



Increasing privacy risk
Data not related to individuals
Data that cannot be linked to any individual
Data that can be ambiguously linked several individuals
Data that can be unambiguously linked to a specific individual
Data that are linked to a specific individual
k-anonymous information
potentially identifying information
identifying information
De-identification moves data to the left

Fig 2. Terminologies

Data Masking techniques:

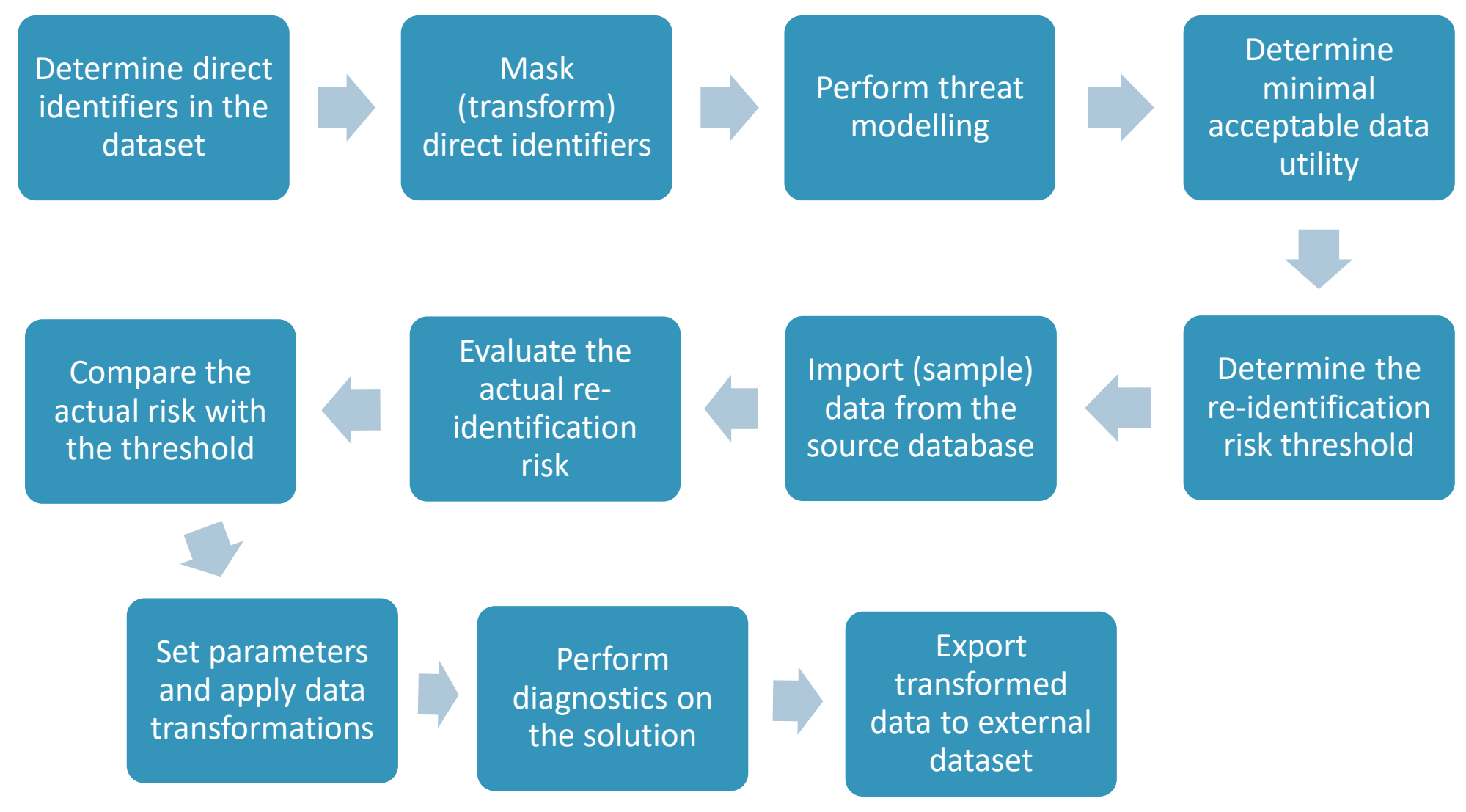
- Auxiliary:
  - Pseudo random function (PRF)
  - Reference values
  - Noise addition
  - Differential privacy
- Blocking:
  - Phonetic encoding,
  - Generalization techniques
- Matching:
  - Secure hash-encoding
  - Statistical linkage key (SLK)
  - Embedding space
  - Encryption schemes,
  - Bloom filter
  - Count-min sketches

Privacy concerns:
Adversary Models:

- Honest-but-curious (HBC) or semi-honest
- Malicious
- Covert and accountable computing

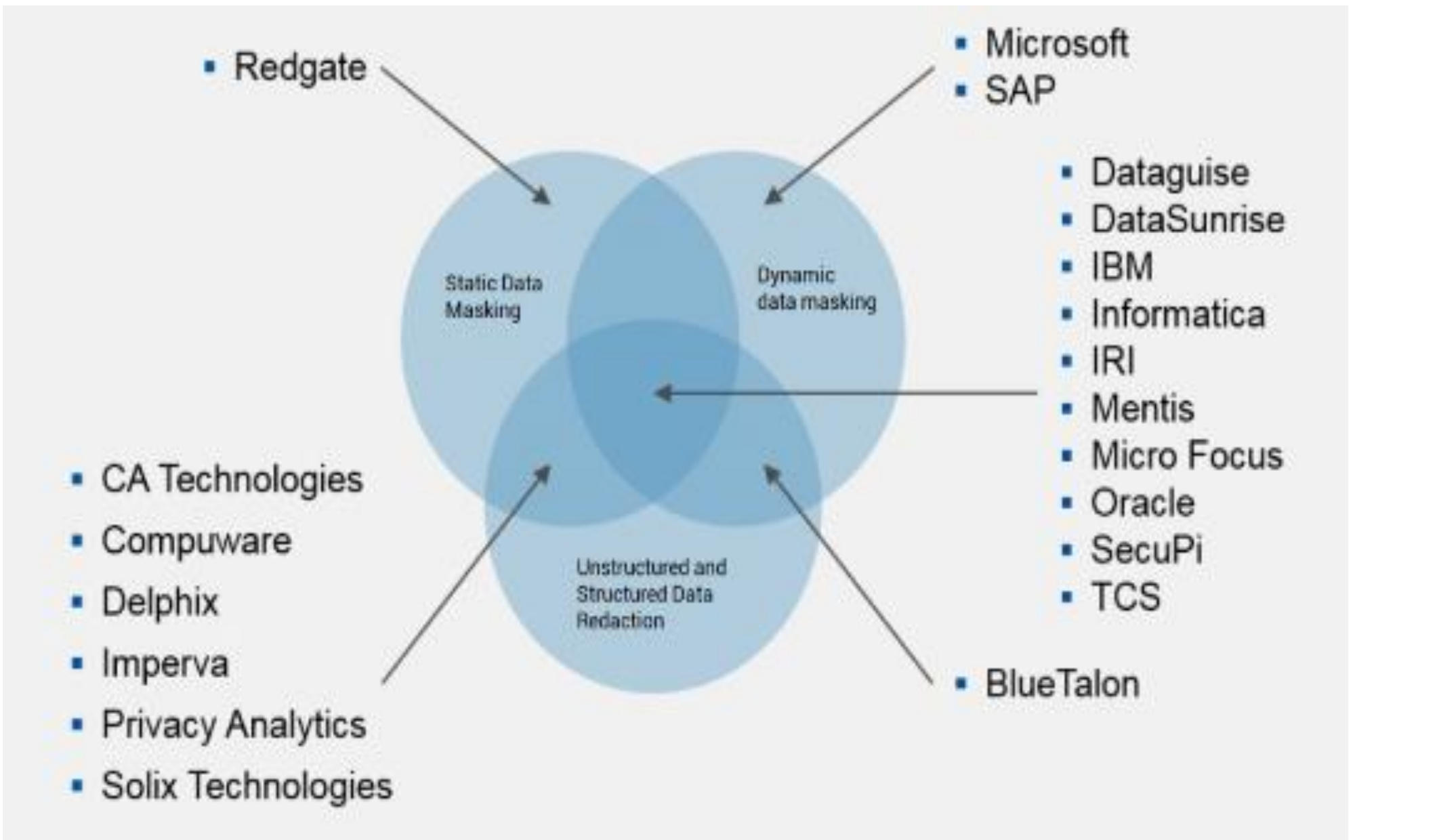
Attacks:

- Dictionary attack
- Frequency attack
- Cryptanalysis attack
- Composition attack
- Collusion



Determine direct identifiers in the dataset
Mask (transform) direct identifiers
Perform threat modelling
Determine minimal acceptable data utility
Determine the re-identification risk threshold
Import (sample) data from the source database
Evaluate the actual re-identification risk
Compare the actual risk with the threshold
Set parameters and apply data transformations
Perform diagnostics on the solution
Export transformed data to external dataset

Fig 3. Data Masking process



Redgate
Microsoft
SAP
Static Data Masking
Dynamic data masking
Unstructured and Structured Data Redaction
Dataguise
DataSunrise
IBM
Informatica
IRI
Mentis
Micro Focus
Oracle
SecuPi
TCS
CA Technologies
Compuware
Delphix
Imperva
Privacy Analytics
Solix Technologies
BlueTalon

Fig 4. Data Masking Tools

Conclusions

- There is a considerably increasing need of novel PPRL .
- PPRL open challenges: Improving Scalability, Improving Linkage Quality, Dynamic Data and Real-Time Matching, Improving Security and Privacy, Evaluation, Frameworks, and Benchmarks
- Many of the PPRL current approaches focus on two parties and are not oriented for multi-parity
- Benchmarks and evaluation models of current data masking techniques are rather immature
- Each PPRL technique focus on one at a time challenge only : scalability, linkage quality, or privacy

References

1. Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In Handbook of Big Data Technologies (pp. 851-895). Springer, Cham.
2. Franke, M., Sehili, Z., & Rahm, E. (2018). Parallel Privacy-preserving Record Linkage using LSH-based Blocking. In IoTBDS (pp. 195-203).
3. Franke, M., Sehili, Z., Gladbach, M., & Rahm, E. (2018). Post-processing methods for high quality privacy-preserving record linkage. In Data Privacy Management, Cryptocurrencies and Blockchain Technology (pp. 263-278). Springer, Cham.
4. Kuang, L., Wang, Y., Ma, P., Yu, L., Li, C., Huang, L., & Zhu, M. (2017). An improved privacy-preserving framework for location-based services based on double cloaking regions with supplementary information constraints. Security and Communication Networks, 2017.
5. El-Ghafar, R. M. A., Gheith, M. H., El-Bastawissy, A. H., & Nasr, E. S. (2017, December). Record linkage approaches in big data: A state of art study. In 2017 13th International Computer Engineering Conference (ICENCO) (pp. 224-230). IEEE.
6. Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. Information Systems, 38(6), 946-969.
7. Rani, V. U., Rao, M. S., & Srujana, K. S. Detection and Privacy Preservation of Sensitive Attributes Using Hybrid Approach for Privacy Preserving Record Linkage. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 54-58.
8. Motiwalla, L., & Li, X. B. (2013). Developing privacy solutions for sharing and analyzing healthcare data. International journal of business information systems, 13(2).
9. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
10. Adam, N. R., & Worthmann, J. C. (1989). Security-control methods for statistical databases: a comparative study. ACM Computing Surveys (CSUR), 21(4), 515-556.