ANONYMIZATION: DATA PRIVACY VS DATA UTILITY

Jezuela Gega, Mariana Mayorga Llano
eBISS Barcelona, 2023

Introduction

The General Data Protection Regulation (GDPR) defines anonymous data as information that does not identify individuals.
Data anonymization techniques aim to protect privacy by removing direct identifiers and minimizing the risk of re-identification.
Balancing data utility and privacy is a critical challenge when anonymizing extensive datasets collected from individuals.
Various techniques, such as adding noise, swapping values, and releasing aggregated data, are used for statistical disclosure limitation (SDL) and protecting confidentiality.
Algorithms like K-Anonymity, α-deassociation, L-Diversity, and T-closeness, as well as graph and stream techniques, offer approaches to address data utility and privacy concerns.

Methods

Scientific goal

Compare techniques to strike a balance between ensuring privacy and maintaining data utility in the context of anonymization.

Methodology

Review and analyse different anonymization techniques: Conduct an extensive literature review to identify and evaluate various anonymization techniques employed in different domains. Explore techniques such as generalization, suppression, noise addition, and k-anonymity, among others.

Comparative analysis: Compare the performance of different techniques based on their ability to strike a balance between privacy and data utility. Assess the trade-offs between privacy protection and the usability of the anonymized data, considering factors such as the structure veracity and instance veracity.

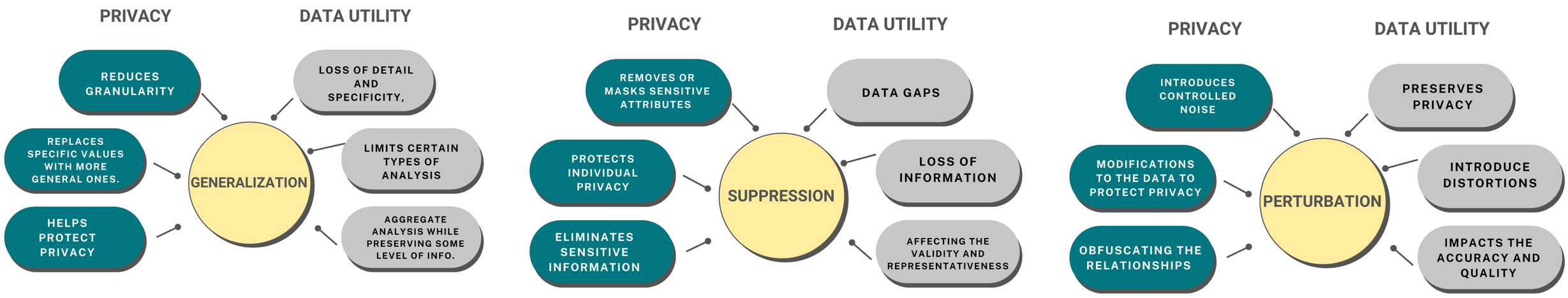

Figure A: Types of techniques and their implications in privacy and data utility.

Comparative

	Technique	Privacy Risks Mitigation: Singling Out Records	Privacy Risks Mitigation: Linkability Detection	Privacy Risks Mitigation: Inference Detection	Utility Implications: Structure Veracity	Utility Implications: Instance Veracity
Generalization	*Bucketization*	Not possible within a group.	Limited within groups of k users.	No prevention of inference attacks.	Loss of granularity and precision, potentially limiting analysis accuracy.	Integrity of data quantity and distribution. No false nor altered results.
Generalization	*Hierarchy Based Generalization*					
Generalization	*Aggregation*					
Generalization	*Clustering Based Generalization*					
Suppression	*Attribute Removal*	Still possible, depending on the attribute			Alterations of data quantity and distribution.	No false nor altered results.
Suppression	*Data Masking*	Possible with high risk of occurrence.	Possible with high risk of occurrence.	Possible with high risk of occurrence.	Integrity of data quantity and distribution.	No false nor altered results.
Perturbation	*Noise Addition*	Still possible, but less reliable, possible false assertions.	Still possible with reduced reliability, possible false assertions.	Lower success rate for inference attacks and possible false results.	Alterations of data quantity and distribution.	May introduce inaccuracies and affect the reliability of data analysis.
Perturbation	*Data Swapping*	Still possible, but less reliable, possible false assertions.	Still possible with reduced reliability, possible false assertions.	Lower success rate for inference attacks and possible false results.	Integrity of data quantity and distribution.	May introduce inaccuracies and affect the reliability of data analysis or subsequent decision-making processes.
Perturbation	*Data Transformation*	Still possible, but less reliable.	Still possible with reduced reliability.	Inferences may still be drawn with probabilistic outcomes.	Integrity of data quantity and distribution.	Can alter attribute order, potentially leading to incorrect conclusions or unreliable results.

Figure B: Privacy risks mitigation and utility implications.

Measurements

Data Privacy

- Risk of disclosure
- Probability of reidentification
- Linkability risk
- ...

Data Utility

- Information Loss
- Distortion
- Accuracy
- ...

Figure C: Measurements for Privacy and Utility.

Alternative Approaches

STREAMING

- **K.Means** and **BIRCH** [O'Callaghan], clustering-based for streaming data.
- **SABREw** [Cao], uses t-closeness, encryption and statistical methods.
- **CASTLE** [Cao], a cluster-based scheme for categorical data
- **FAANST** [Zakerzadeh] to work with numerical data based on cluster-based k-anonymity

GRAPH

- **Two-level vertex anonymisation** [Zhou], where the first level is applied for the labels and the second in the edges.
- **Social network greedy anonymisation** [Campan] based on clustering approach.
- Creating a new degree sequence which satisfies k-anonymity in a supergraph [Clarkson]

OTHER

- **Airavat** [Roy] is a **MapReduce** with a mapper, that can be untrusted or trusted, and a trusted reducer.
- **UPGMA** cluster-combination method to reduce communication costs and achieve a better level of privacy.

Figure D: Alternative anonymization approaches

Conclusion

1.Finding a single anonymization technique that suits all use cases and data analysis purposes is challenging, as different techniques serve different needs.
2.Care must be taken when implement different anonymization techniques independently over a same dataset to avoid the risk of identification if all versions are brought together.
3.While various anonymization techniques exist, some have limitations and are still in early stages of development, particularly for non-relational, streaming, and dynamic data.
4.Continued research and development are necessary to adapt anonymization techniques to effectively handle the evolving nature of data, ensuring privacy protection and data utility in different processing scenarios.

References

Advisors, S. C. (2020, July 21). Anonymization and GDPR compliance; an Overview. GDPR Summary. https://www.gdprsummary.com/anonymization-and-gdpr/

Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey", IEEE Access, vol. 9, pp. 8512-8545, 2021. https://ieeexplore.ieee.org/abstract/document/9298747

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). Fast Data Anonymization with Low Information Loss. Retrieved from https://www.vldb.org/conf/2007/papers/research/p758-ghinita.pdf

Li, N., Li, T. and Venkatasubramanian, S. (2007) 't-closeness: privacy beyond k-anonymity and l-diversity', in ICDE 2007. IEEE 23rd International Conference on Data Engineering, 2007, IEEE, pp.106–115. Retrived from https://www.researchgate.net/publication/4251020_t-Closeness_Privacy_Beyond_k-Anonymity_and_l-Diversity

Contact authors
Jezuela.gega@estudiantat.upc.edu
Mariana.mayorga@estudiantat.upc.edu