UDC 004.056:336.71(076)

ENHANCING PERSONAL DATA DE-IDENTIFICATION USING DES-BASED TRANSFORMATION TECHNIQUES

C. Li
Educational Institution "Belarusian State University of Informatics and Radioelectronics", Minsk, Republic of Belarus

Abstract. Operational data sharing often requires stable record linkage while reducing identity disclosure risk. This paper outlines a compact two-layer de-identification design: DES-family deterministic pseudonymization (preferably Triple-DES) to transform direct identifiers into reversible tokens, and k-anonymity-guided generalization/suppression to limit linkage via quasi-identifiers. We show how DES-style permutations can be adapted to format-constrained fields through rank-then-encipher methods, and we summarize key security and utility trade-offs.

Keywords: de-identification; pseudonymization; tokenization; DES; Triple-DES; format-preserving encryption; rank-then-encipher; k-anonymity; quasi-identifiers; re-identification risk.

Introduction

De-identification aims to reduce the chance that released data can be linked back to individuals, while keeping data useful for analysis and operations. In practice, removing names is insufficient because combinations of quasi-identifiers (e.g., date of birth, region, sex) can enable linkage to external datasets, motivating formal models such as k-anonymity [4,5]. At the same time, many workflows require consistent linkage across tables and time; therefore, deterministic, keyed transformations (pseudonymization) are commonly used to replace direct identifiers with stable tokens [5,6].

This paper focuses on DES-based transformation because DES is a well-studied block cipher whose security debate highlights the importance of key length and brute-force feasibility [1,2]. Given modern threats, single-DES (56-bit) should not be relied on; where legacy constraints exist, Triple-DES is the preferred DES-family option for token generation [1,2,3].

Transformation of Personal Data for Anonymization

We propose a two-layer pipeline that separates (A) protection of direct identifiers from (B) control of quasi-identifier linkage risk. Layer A creates reversible pseudonyms (tokens) for identifiers such as account numbers or patient IDs; Layer B generalizes or suppresses quasi-identifiers until the released dataset satisfies a k-anonymity threshold [4,5].

A block cipher is a keyed permutation over fixed-size blocks; under a secret key it maps each input to a unique output and supports inversion with

the same key. DES was standardized for interoperability and scrutinized widely; its 56-bit key length makes exhaustive search increasingly feasible, as early analyses argued [1,2].

For de-identification, we use the permutation property to create deterministic tokens: pack an identifier into one or more 64-bit blocks, apply Triple-DES as the core permutation, and re-encode the result into the field's allowed character set. Determinism preserves referential integrity for joins, while access to keys can be restricted for controlled re-identification [1,3,6].

Many real identifiers have constrained formats (e.g., fixed-length digits). Format-preserving techniques address this by using Rank-then-Encipher (RtE): ranking maps a formatted value to an integer, a block-cipher-based permutation encrypts the integer, and unranking maps it back to the original format [7,8]. RtE aims to preserve only format properties; preserving extra message-specific characteristics can leak information and weaken protection [7].

In a DES-family setting, Triple-DES can serve as the underlying permutation inside RtE when system constraints require it, but designers should apply domain separation (per-field keys or context tweaks) to prevent cross-dataset linkability, especially for small domains where enumeration is practical [8].

k-Anonymity requires that each quasi-identifier pattern in the released dataset correspond to at least k records, reducing the chance that a record uniquely matches an external reference [4]. A practical enforcement method is to iteratively generalize (e.g., date→year, ZIP→region) and suppress outliers until the minimum equivalence-class size reaches k. Empirical work on health datasets shows that algorithm choice affects information loss and runtime, and that optimized methods can improve utility compared to simpler heuristics [6].

Security: deterministic tokens enable linkage, but they remain vulnerable if the token domain is small or if the underlying cipher key is weak; therefore single-DES is unsuitable and Triple-DES should be used when DES is unavoidable [1,2]. Utility: stronger generalization for k-anonymity reduces re-identification risk but degrades analytic precision; selecting k and hierarchies should be guided by intended use and acceptable risk thresholds [5, 6].

Conclusion

A practical way to enhance de-identification is to combine cryptographic pseudonymization of direct identifiers with formal control of quasi-identifier linkage. DES-family permutations (preferably Triple-DES) can generate stable, reversible tokens and can be adapted to format-constrained fields via RtE methods. k-Anonymity then reduces linkage risk by ensuring non-unique quasi-identifier patterns, with an explicit privacy–utility trade-off

References

1. Smid, M. E., & Branstad, D. K. (2002). Data encryption standard: past and future. *Proceedings of the IEEE*, *76*(5), 550-559.

2. Diffie, W., & Hellman, M. E. (2006). Special feature exhaustive cryptanalysis of the NBS data encryption standard. *Computer*, *10*(6), 74-84.

3. Schneier, B. (2007). *Applied cryptography: protocols, algorithms, and source code in C*. john wiley & sons.

4. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, *10*(05), 557-570.

5. El Emam, K. (2011). Methods for the de-identification of electronic health records for genomic research. *Genome medicine*, *3*(4), 25.

6. El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., ... & Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, *16*(5), 670-682.

7. Weiss, M., Rozenberg, B., & Barham, M. (2015). Practical solutions for format-preserving encryption. *arXiv preprint arXiv:1506.04113*.

8. Durak F. B., Horst H., Horst M., Vaudenay S. (2021) FAST: Secure and High Performance Format-Preserving Encryption and Tokenization. In: ASIACRYPT 2021, LNCS 13092, 465–489. Springer.