

Asjad Moiz Khan

Supervised by Dr Luisa Cutillo (University of Leeds); Andrew Swanton and Ewen Anderson (Experian)
MATH5872M: Dissertation in Data Science and Analytics, University of Leeds, School of Mathematics

Background and Motivation

Mortgage lenders verify applicant income by asking payroll providers directly, faster than manual payslip checks. But this requires the payroll provider to search their whole workforce database using the applicant's name, date of birth, and postcode, and some providers treat that search as needing explicit consent from every employee, not just the applicant, which they often don't have. Where consent is unavailable, the provider can't take part, and the applicant is left with the slower manual route. This project tests whether privacy-preserving matching can close that gap: a lender and a payroll provider confirming they mean the same person, without either side sharing that person's raw details.

Data and Methods

Two synthetic Experian datasets were used: a cleaner 10,000-record-pair set and a noisier 25,000-record-pair set. Name and date-of-birth noise were characterised empirically first, which shaped a combined bi-gram/token-set Dice similarity measure and a fuzzy date-of-birth rule tolerant of day/month swaps and single-digit typos. Both fields are compared using Cryptographic Longterm Key (CLK) Bloom filter encoding, so neither party ever sees the other's raw name or date of birth. The pipeline was validated with a stratified train/test split, then tested against three real attacks from the Vidanage et al. (2022) taxonomy, built and run directly against the encoded output, with a salting mitigation implemented and tested afterward.

Key Findings

- **Accuracy:** the unified pipeline reaches **99.7%** agreement with Experian's own matching system on the cleaner dataset and **81.3%** on the noisier one.
- **One deployable model:** a single fuzzy date-of-birth rule works across both datasets, independently matching the tolerance Experian's own production system already uses.
- **Security tested, not assumed:** three real attacks (dictionary, frequency, date-of-birth brute-force) all succeed near-100% against the unsalted encoding.
- **Salting closes two of three:** a low-cost fix collapses the dictionary and frequency attacks to near-zero success, at no accuracy cost.
- **One vulnerability remains open:** the date-of-birth brute-force attack isn't fixed by salting, and is reported as a limitation rather than resolved.
- **PSI comparison:** an exact-match alternative is roughly $1,000\times$ harder to brute-force but only reaches F1 0.318, too low to be usable on this data.

Value of the Research

This project shows privacy-preserving matching can reach accuracy close to an existing full-data system, using one practical method instead of a theoretical proposal. It also shows what such a system currently protects against and what it doesn't, measured directly instead of assumed: most of the security gaps found here close cheaply, one doesn't, and both are stated plainly. For payroll providers and lenders, that gives a concrete starting point for deciding whether privacy-preserving matching is worth taking further, and a clear list of what still needs fixing first.

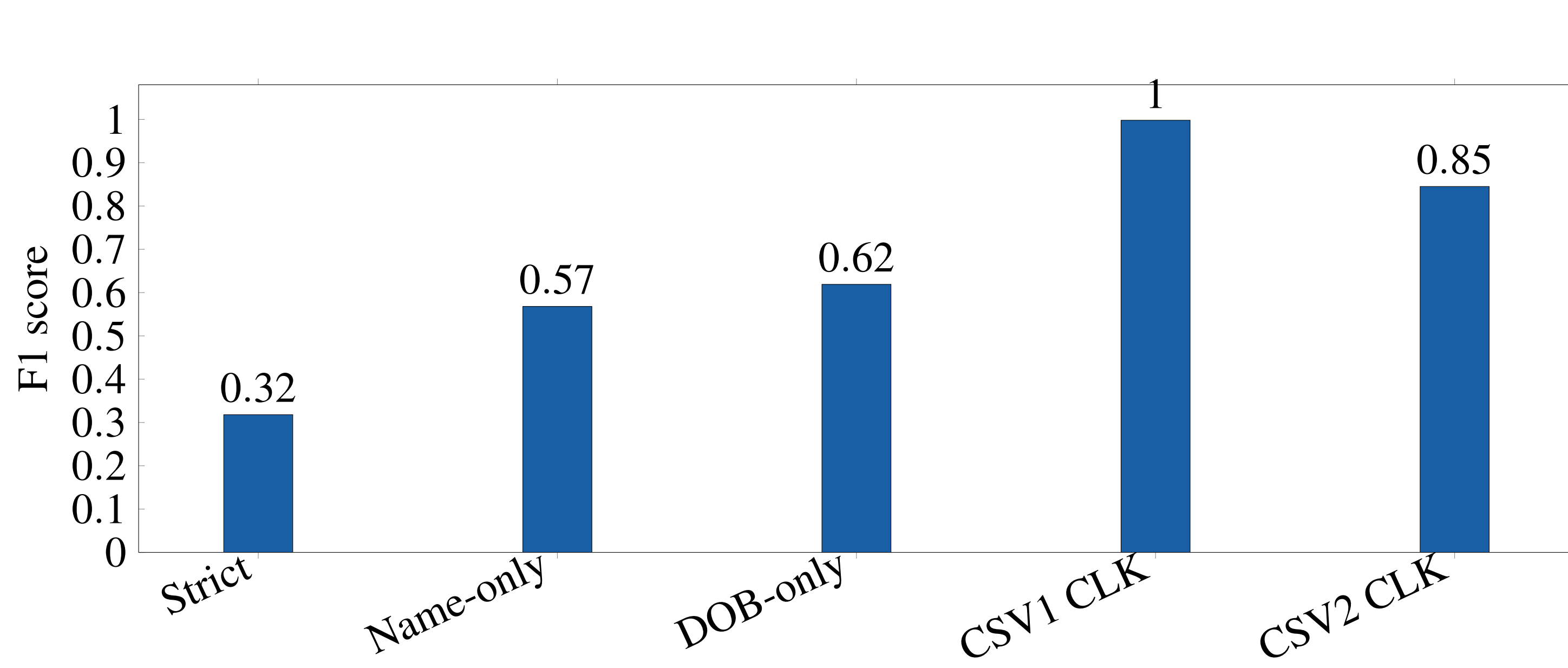


Figure 1: Matching accuracy (F1) across baselines and the unified CLK pipeline, both datasets.

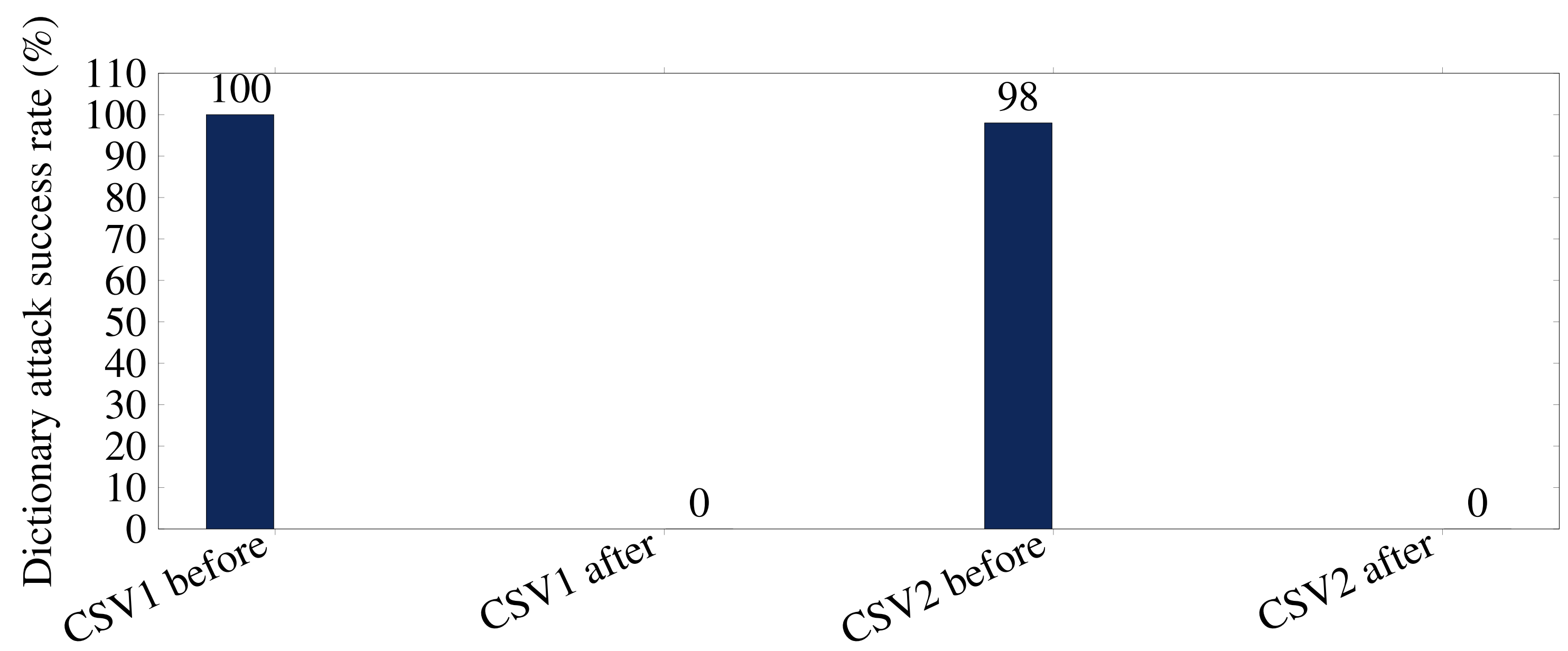


Figure 2: Dictionary attack success rate before and after salting, both datasets.

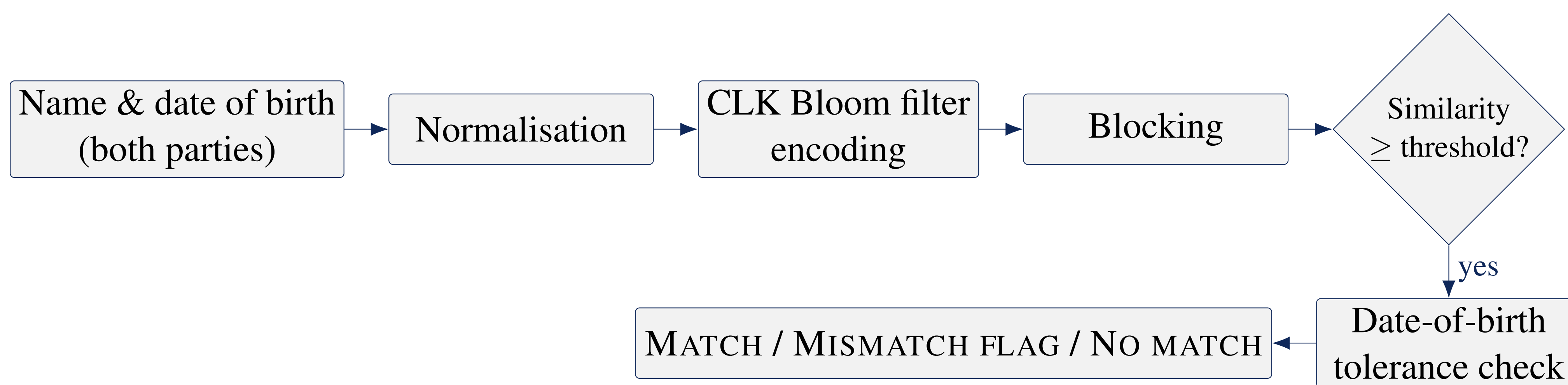


Figure 3: The unified matching pipeline. Only the final decision is ever shared between parties; the underlying name and date of birth are never transmitted in raw form.

Selected References

- Schnell, R., Bachteler, T., & Reiher, J. (2011). *A Novel Error-Tolerant Anonymous Linking Code*. German Record Linkage Center.
- Vidanage, A., Ranbaduge, T., Christen, P., & Schnell, R. (2022). A taxonomy of attacks on privacy-preserving record linkage. *Journal of Privacy and Confidentiality*, 12(1).
- Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data. In *Handbook of Big Data Technologies*. Springer.
- Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques. *IEEE TIFS*, 16.

Acknowledgements

With thanks to Dr Luisa Cutillo (University of Leeds) for academic supervision, and Andrew Swanton and Ewen Anderson (Experian) for industrial supervision, the datasets used in this work, and feedback on the matching methodology throughout the project.