

Probabilistic record linkage across federated benefit rolls

2026-10-11

Abstract. A working note on detecting duplicate enrollment when every jurisdiction keeps its own records: the Fellegi–Sunter model, unsupervised weight estimation with EM, and the blocking and threshold choices that decide whether the result is usable.

1 The problem

A national program administered by many jurisdictions rarely has a shared, reliable identifier. Each jurisdiction maintains its own enrollment file, formats names and addresses its own way, and corrects errors on its own schedule. Detecting a person enrolled in more than one place is therefore a *record linkage* problem: deciding, for pairs of records drawn from different files, whether they refer to the same individual.

Exact matching on a few fields fails in both directions. It misses true duplicates whenever a single character differs, and it over-matches common names that share a date of birth. Probabilistic linkage replaces the yes-or-no rule with a calibrated score and makes the error trade-off explicit.

2 The Fellegi–Sunter model

Let A and B be two files. Every candidate pair $(a, b) \in A \times B$ belongs to one of two unknown classes: matches M or non-matches U . For each pair we compute a comparison vector $\gamma = (\gamma_1, \dots, \gamma_K)$, where $\gamma_k = 1$ if field k agrees and 0 otherwise.¹

The model is parameterized by two probabilities per field:

$$m_k = \Pr(\gamma_k = 1 \mid M), \quad u_k = \Pr(\gamma_k = 1 \mid U). \quad (1)$$

Here m_k captures data quality (how often a true match agrees on field k), and u_k captures coincidence (how often two different people agree by chance). Assuming conditional independence of fields given the class, the log-likelihood ratio of a pair is a sum of per-field weights:

$$w(\gamma) = \sum_{k=1}^K \left[\gamma_k \log_2 \frac{m_k}{u_k} + (1 - \gamma_k) \log_2 \frac{1 - m_k}{1 - u_k} \right]. \quad (2)$$

I. P. Fellegi and A. B. Sunter [1] showed that the decision rule with two thresholds $T_\mu > T_\lambda$,

$$\text{decide} \begin{cases} M & \text{if } w(\gamma) \geq T_\mu \\ \text{review} & \text{if } T_\lambda < w(\gamma) < T_\mu \\ U & \text{if } w(\gamma) \leq T_\lambda \end{cases} \quad (3)$$

¹Real deployments use more than two levels per field, for example exact, approximate, and disagree. The algebra is unchanged; each level gets its own probability.

minimizes the size of the clerical review region for fixed error rates on the two automatic decisions. That property is what makes the method operational: it tells you exactly how much human review a given tolerance for error will cost.

Field	m_k	u_k	Agree	Disagree
Last name	0.95	0.010	+6.57	-4.31
First name	0.90	0.020	+5.49	-3.29
Date of birth	0.97	0.003	+8.34	-5.05
Sex	0.98	0.500	+0.97	-4.64
ZIP code	0.80	0.020	+5.32	-2.29

Table 1: Illustrative parameters and the resulting weights from Equation 2, in bits.

With the values in Table 1, a pair that agrees on every field scores +26.69. A pair that agrees on everything except ZIP code, the common case of a household that has moved, still scores $26.69 - 5.32 - 2.29 = 19.08$. A pair that agrees only on sex scores well below zero. The weights encode intuition that is otherwise hard to apply consistently: agreement on a rare value is strong evidence, while agreement on a common one is almost none.



Figure 1: The decision rule in Equation 3 partitions the weight axis into three regions.

3 Estimating weights without labels

Labeled pairs are expensive and, across jurisdictions, usually unavailable. The parameters can instead be estimated directly from the unlabeled comparison vectors by treating class membership as missing data and applying the EM algorithm [2], [3]. Let $p = \Pr(M)$ be the match prevalence among candidate pairs. In the E-step, each pair j receives a posterior match probability

$$g_j = \frac{p \prod_k m_k^{\gamma_{jk}} (1 - m_k)^{1 - \gamma_{jk}}}{p \prod_k m_k^{\gamma_{jk}} (1 - m_k)^{1 - \gamma_{jk}} + (1 - p) \prod_k u_k^{\gamma_{jk}} (1 - u_k)^{1 - \gamma_{jk}}}. \quad (4)$$

In the M-step, the parameters are re-estimated as weighted proportions:

$$\hat{m}_k = \frac{\sum_j g_j \gamma_{jk}}{\sum_j g_j}, \quad \hat{u}_k = \frac{\sum_j (1 - g_j) \gamma_{jk}}{\sum_j (1 - g_j)}, \quad \hat{p} = \frac{1}{N} \sum_j g_j. \quad (5)$$

Two practical cautions apply. First, EM converges to a local optimum, so initialize with m_k high and u_k low; otherwise the classes can swap. Second, the u_k estimates are dominated by the overwhelming number of non-matching pairs and are usually reliable, while the m_k estimates depend on there being enough true matches among the candidates. That makes blocking a statistical decision, not only a performance one.

4 Blocking

Comparing every pair is quadratic in file size and infeasible at national scale. Blocking restricts comparison to pairs that share a key. Every key trades pair volume against recall: a true match that disagrees on the blocking key is never compared and can never be found.

Blocking key	Recall risk	Notes
Date of birth	Low	Very selective; misses transposed digits.
Phonetic last name + birth year	Low–moderate	Tolerant of spelling; misses name changes.
ZIP code	High	Misses movers, the population of most interest.

Table 2: Qualitative trade-offs of common blocking keys.

The standard remedy is to run several independent blocking passes and take the union of candidate pairs. As Table 2 suggests, keys should be chosen so that their failure modes do not overlap.

5 Choosing thresholds

Under the model, the expected number of false matches accepted above T_μ is

$$\mathbb{E}[\text{false matches}] = \sum_{j:w_j \geq T_\mu} (1 - g_j), \quad (6)$$

which can be read directly from the EM output. In practice we set T_μ so that this quantity is a small fraction of accepted pairs, set T_λ from the review capacity actually available, and audit a random sample from each region. Because conditional independence rarely holds exactly (first name and sex are correlated, for example), posterior probabilities tend to be overconfident. Audits are what keep the thresholds honest.

6 Practical notes

- **Normalize before comparing.** Case, punctuation, suffixes, nicknames, and address standardization remove more spurious disagreement than any model refinement.
- **Use graded string comparison.** Approximate agreement levels based on string comparators such as Jaro–Winkler substantially improve linkage of names [4].
- **Resolve transitively.** Pairwise decisions must be reconciled into clusters; a single wrong link can merge two large groups.
- **Treat outputs as leads.** A high score indicates a likely duplicate record, not a determination about a person. Adverse action should always follow verification.

Bibliography

- [1] I. P. Fellegi and A. B. Sunter, “A Theory for Record Linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [3] W. E. Winkler, “Using the EM Algorithm for Weight Computation in the Fellegi–Sunter Model of Record Linkage,” in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1988, pp. 667–671.
- [4] W. E. Winkler, “String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage,” in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1990, pp. 354–359.