Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Fellegi-Sunter Model: m and u Weights, a Worked Example

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · fellegi sunter model · fellegi-sunter · m and u probabilities · probabilistic record linkage · splink fellegi sunter

Original article: https://reliablecontext.com/articles/fellegi-sunter-model-match-weights-worked-example



In this article

1. Linkage as a classification problem
2. Agreement patterns and the two probabilities
3. The match weight is a log likelihood ratio
4. Worked example with four fields
5. Thresholds: match, possible match, non-match
6. Estimating m and u without labels: the EM algorithm
7. Term-frequency adjustments
8. Pitfalls
9. Sources

Linkage as a classification problem

Record linkage is the task of finding records, in one file or across several, that refer to the same entity when no shared unique identifier is available; it relies mainly on names, addresses and other fields that are not unique identifiers (Winkler 2006, §2). Probabilistic record linkage treats every candidate pair of records as an object to classify.

The standard model comes from Ivan Fellegi and Alan Sunter's "A Theory for Record Linkage", published in the *Journal of the American Statistical Association* in 1969 (DOI record). Their paper gave a formal mathematical footing to ideas Howard Newcombe had introduced in 1959 and 1962 (Winkler 2006, §3.1). Its abstract, as indexed by OpenAlex, states the problem as recognizing "those records in two files which represent identical persons, objects or events" and names three decisions, link (A1), possible link (A2) and non-link (A3), with two error types fixed "at stipulated levels of error". The full article is behind a paywall, so every formula below is taken from open sources that restate it: a US Census Bureau research report by William Winkler, the review by Binette and Steorts, and the documentation of Splink, an open-source Python package whose core linkage algorithm is based on the Fellegi-Sunter model. For a short definition, see the Fellegi-Sunter model glossary entry. This article works through the arithmetic by hand.

The model splits all record pairs into two sets: M, the true matches, and U, the true non-matches. Each pair is assigned to one of three actions: link, possible link (sent to clerical review), or do not link (Binette and Steorts, §4.2).

Agreement patterns and the two probabilities

Each pair is compared field by field. The outcome of those comparisons is an **agreement pattern** γ (gamma), also called a comparison vector. With three fields, each either agreeing or disagreeing, there are eight possible patterns (Winkler 2006, §3.1); with four fields there are sixteen.

The model has three kinds of parameter (Splink, The Fellegi-Sunter model):

Term	Definition	What it usually measures
m probability	Probability of an observation (for example, "surname agrees") given that the two records are a match	Data quality: typos, nicknames, people moving house
u probability	Probability of the same observation given that the records are not a match	Coincidence: how often two different entities share a value
λ (lambda), the prior	Probability that two records drawn at random match	Duplication and overlap in the input data
Bayes factor K	The ratio m/u for one observation	How much one observation shifts the odds
Match weight	log2 of a Bayes factor; a partial match weight is the weight of one field	Evidence on an additive scale

Splink describes m as largely a measure of data quality and u as a measure of coincidence or cardinality: date of birth has thousands of possible values and a low u, while a two-valued field such as sex agrees about half the time by chance (Splink, The Fellegi-Sunter model).

The match weight is a log likelihood ratio

Fellegi and Sunter considered the ratio R = P(γ | M) / P(γ | U). Winkler notes that R, or any monotonically increasing function of it such as its logarithm, is called a matching weight or score (Winkler 2006, eq. 1). Binette and Steorts write it as W(γ) = log m(γ) − log u(γ) (§4.2). Splink uses base-2 logarithms, so one unit of weight doubles the odds; this article does the same.

The pattern-level ratio becomes usable only through one assumption, **conditional independence**: given the true match status, agreement on one field tells you nothing about agreement on another. Under that assumption the log of the pattern ratio equals the sum of per-field log ratios (Winkler 2006, eq. 4). For binary comparisons, field i contributes:

- agreement weight: `log2(m_i / u_i)`, positive when m exceeds u
- disagreement weight: `log2((1 − m_i) / (1 − u_i))`, negative in the usual case

Winkler states that agreement weights are typically positive and disagreement weights negative for exactly this reason (§3.5). Winkler also notes the assumption is the naïve Bayes assumption of machine learning.

The prior enters as its own weight. Splink writes the total weight as `M_obs = log2(λ / (1 − λ)) + Σ log2(m_i / u_i)` and converts it to a probability with `P(match | observation) = 2^M_obs / (1 + 2^M_obs)` (Splink). The same quantity follows from Bayes' theorem as `λ·m_1…m_n / (λ·m_1…m_n + (1 − λ)·u_1…u_n)` (Linacre, The mathematics of the Fellegi Sunter model; Binette and Steorts, eq. 1). Useful anchors from Splink's chart: a total weight of 0 means probability 0.5, about 3 means 0.9, and about 7 means 0.99.

Worked example with four fields

Scenario (fictional). Halden Therapeutics AG, a fictional company, is deduplicating a list of 10,000 investigator contact records. Assume it contains 1,000 duplicate pairs. The records, names and parameter values below are invented for illustration; they are not estimates from real data.

Step 1: m, u and the per-field weights

Field	m	u	m/u	Agreement weight	(1−m)/(1−u)	Disagreement weight
Surname	0.95	0.01	95.0	+6.57	0.0505	−4.31
First name	0.90	0.02	45.0	+5.49	0.1020	−3.29
Birth year	0.97	0.03	32.3	+5.01	0.0309	−5.01
Postcode	0.80	0.001	800.0	+9.64	0.2002	−2.32

Read the rows as stories. Postcode has the lowest m because people move, yet the largest agreement weight because two different people rarely share one (u = 0.001). Birth year has a high m but a u of 0.03, roughly what birth years spread evenly over about 33 years would produce. A postcode disagreement costs only 2.32 because it is common among true matches; a birth-year disagreement costs 5.01 because true matches almost never differ on it.

Step 2: the prior

10,000 records form 10,000 × 9,999 / 2 = 49,995,000 pairs. With 1,000 true matches, $\lambda = 1{,}000 / 49{,}995{,}000 \approx 2.0 \times 10^{-5}$, prior odds of about 1 in 49,994, and a prior weight of $\log_2(\lambda / (1 - \lambda))$ = **−15.61**. Before any field is compared, a random pair needs about 15.6 units of evidence just to reach even odds.

Step 3: five candidate pairs

Pair	Record 1	Record 2	Pattern (S, F, B, P)
A	Brunner, Lea, 1979, 4051	Brunner, Lea, 1979, 4051	agree, agree, agree, agree
B	Keller, Jonas, 1984, 8004	Keller, Jonas, 1984, 3011	agree, agree, agree, disagree
C	Baumann, Katharina, 1971, 6003	Baumann, Katja, 1971, 6003	agree, disagree, agree, agree
D	Huber, Martin, 1956, 9000	Huber, Simone, 1988, 9000	agree, disagree, disagree, agree
E	Steiner, Paul, 1990, 1201	Frei, Nadia, 1990, 4410	disagree, disagree, agree, disagree

Pair B, by hand: 6.57 + 5.49 + 5.01 − 2.32 = 14.76 from the fields. Add the prior: −15.61 + 14.76 = −0.85. Convert: 2^−0.85 ≈ 0.555, and 0.555 / 1.555 ≈ 0.356.

Pair	Sum of field weights	Total weight (with prior)	Match probability
A	+26.72	+11.11	0.99955

Pair	Sum of field weights	Total weight (with prior)	Match probability
B	+14.76	−0.85	0.356
C	+17.94	+2.33	0.834
D	+7.91	−7.70	0.0048
E	−4.91	−20.52	about 7 in 10 million

Weights are shown to two decimals; totals were computed from unrounded values, and each probability was checked against the direct Bayes formula.

Three points stand out. Pair B agrees on every personal field and is still less likely than not to be a match, because the prior is so small and the postcode evidence points the other way. Pair C survives a first-name disagreement because postcode agreement is worth 9.64. And "Katja" against "Katharina" counts as a full disagreement only because the comparison here is exact; string comparators that grade partial agreement exist precisely to handle such variants (Winkler 2006, §3.3), and in Splink each column can have several levels such as exact match, fuzzy match and all other (Linacre).

Thresholds: match, possible match, non-match

The Fellegi-Sunter decision rule uses two cut-offs. If $R > T_\mu$, designate the pair a match; if $T_\lambda \le R \le T_\mu$, a possible match held for clerical review; if $R < T_\lambda$, a non-match. The thresholds are set from a priori bounds on the false-match and false-non-match error rates, and the middle region is the clerical review region (Winkler 2006, eq. 2). Fellegi and Sunter showed that the optimal procedure, which reaches given error rates with the fewest possible links, is a likelihood ratio test (Binette and Steorts, §4.2). (The subscripts μ and λ here name the two error rates; they are not the prior λ.)

Because the model specifies m(γ) and u(γ) for every pattern, the error rates for a threshold pair can be computed directly under the model: the false-match rate is the total u probability of the patterns above the upper threshold, and the false-non-match rate is the total m probability of the patterns below the lower one. Take an upper threshold of +3 (probability ≈ 0.89) and a lower threshold of −3 (≈ 0.11):

Pattern (S, F, B, P)	Total weight	Match probability	m(γ)	u(γ)	Decision
A A A A	+11.11	0.9995	0.6635	6.0×10^{-9}	Match
A d A A	+2.33	0.834	0.0737	2.9×10^{-7}	Review
A A d A	+1.08	0.679	0.0205	1.9×10^{-7}	Review
d A A A	+0.23	0.540	0.0349	5.9×10^{-7}	Review
A A A d	−0.85	0.356	0.1659	6.0×10^{-6}	Review
11 other patterns	−7.70 or lower	≤ 0.0048	0.0415 in total	≈ 1.0 in total	Non-match

A = agree, d = disagree. Under the model, only 6×10^{-9} of non-matching pairs are linked automatically, about 0.3 of the 49,994,000 non-matches. The cost is elsewhere: 4.15% of true matches (about 41.5 of 1,000) fall below the lower threshold, and about 649 pairs go to review, 295 of them true matches and 354 not.

Moving the thresholds shows the trade-off:

Upper / lower threshold	Automatic links: true / false	Pairs to review	Missed matches
+3 / −3	663.5 / 0.3	648.8	41.5
0 / −3	792.6 / 54.4	465.5	41.5
+3 / −8	663.5 / 0.3	1,126.3	39.2

Lowering the upper threshold to 0 saves about 183 reviews and admits about 54 false links. Lowering the lower threshold to −8 adds about 477 reviews to recover about 2 matches. These are expected counts under a model whose assumptions hold by construction. Real error rates depend on whether conditional independence and the parameter estimates hold; Winkler reports that automatic error-rate estimation methods worked well only in a narrow range of situations (§3.5), so a clerically reviewed sample remains the usual check.

Estimating m and u without labels: the EM algorithm

In practice nobody hands you m and u. Fellegi and Sunter observed that the distribution of observed patterns is a mixture, $P(\gamma) = P(\gamma \mid M)P(M) + P(\gamma \mid U)P(U)$. With three fields under conditional independence, this gives seven equations in seven unknowns. Winkler (1988) showed how to estimate the parameters with the EM (expectation-maximisation) algorithm, an unsupervised method; he advises that the share of matches in the pairs it is applied to should be above 5% (Winkler 2006, §3.5). Jaro (1989) independently proposed EM for binary comparisons (Binette and Steorts).

The intuition: matches and non-matches form two fairly distinct clusters, so a crude rule can label most pairs. The expectation step predicts which pairs match from the current parameters; the maximisation step re-estimates the parameters from those predictions, weighting each pair by its match probability. The procedure maximises the model's likelihood but is guaranteed only to reach a local maximum (Linacre, EM intuition).

Splink recommends a hybrid approach (Splink, Training rationale):

1. Estimate λ directly from strict deterministic rules plus a user-supplied guess of their recall.
2. Estimate u by comparing random pairs, which are almost all non-matches.
3. Estimate m with EM on blocked pairs, which contain far more matches than random pairs. Because a blocking column is forced to agree, each pass cannot estimate that column's m, so Splink runs several passes on different blocking rules and averages the estimates.

As a check, we simulated 200,000 synthetic pairs from the example's parameters with a 10% match share and ran plain EM from deliberately poor starting values (every m = 0.9, every u = 0.1, λ = 0.5). EM returned λ = 0.099 and m estimates of 0.950, 0.894, 0.970 and 0.801 against true values of 0.95, 0.90, 0.97 and 0.80, with u estimates within 0.0005 of the truth. This is a best case: the simulated fields are conditionally independent by construction.

Term-frequency adjustments

A single u for surname treats agreement on a common name and on a rare one as equally informative. Winkler gives the example of "Martinez", which is relatively rare in Minneapolis and relatively frequent in Los Angeles, and of "Zabrinsky" against "Smith" (Winkler 2006). Binette and Steorts note that Fellegi and Sunter themselves considered value-specific comparisons (§4.2).

Splink's term-frequency adjustment works as follows. For a value X with term frequency T_X (its share of records), the value-specific m is m·T_X and the value-specific u is approximately T_X^2. The Bayes factor becomes K·(u / T_X), so the adjustment adds `log2(u / T_X)` to the match weight (Splink, Term-frequency adjustments, formula shown as a figure). In Splink's own toy example, with u = 77/190 and a first name of frequency 0.1, the adjustment is +2.02.

In the example, the surname u is 0.01:

Surname frequency T_X	Adjustment log2(0.01 / T_X)	Adjusted agreement weight
0.04 (common surname)	−2.00	+4.57
0.01 (average)	0.00	+6.57
0.0005 (rare surname)	+4.32	+10.89

Splink notes that the adjustment does not depend on m, so EM does not have to estimate it, and it can be switched off (Splink).

Pitfalls

The independence assumption

Conditional independence rarely holds exactly. Winkler gives two failures in opposite directions (§3.5). If one field is a near copy of another, such as surname and its Soundex code, summing both agreement weights inflates the total. If agreement on last name, date of birth and ZIP+4 already guarantees agreement on first name among matches, the independence calculation makes the total weight too low. He also notes that some non-matches, such as a husband and wife, agree on many fields. Pair D in the example is that case: surname and postcode agree for two relatives at one address, and among non-matches that combination is far more common than the product 0.01 × 0.001 assumes.

Binette and Steorts record that the problem was first noted by Smith and Newcombe (1975), that log-linear models with interaction terms can model the dependence, and that Xu et al. (2019) found the independence model performs comparably when the linking fields discriminate well (§4.2). Splink advises input columns that are not highly correlated and gives city and postcode as an example of a nearly correlated pair (Splink).

Interaction with blocking

Blocking compares only pairs that agree on some characteristic, such as surname or date of birth, so that the full cross product never has to be scored (Winkler 2006, §2; Papadakis et al.). A true match that no blocking rule brings together receives no weight and never reaches either threshold, so the false-non-match rates computed above exclude it; auditing for that loss is covered in how to audit blocking for missed matches. Blocking also affects training: Splink warns that estimating m on blocked pairs can give a biased selection of matches when conditional independence fails (Training rationale).

Pairs are not independent of each other

The model scores each pair in isolation. One-to-one linkage, where each record in one file can link to at most one record in the other, and transitivity (A–B and B–C imply A–C) both create dependencies between pairs that the basic likelihood ratio test does not see (Binette and Steorts, §4.2).

Parameters do not travel

Winkler reports that optimal m and u probabilities can vary substantially between regions, for example between an urban and an adjacent suburban area, largely because of different typographical error rates (§3.5). Re-estimate parameters for each source and period rather than reusing them. For when a weighted model is the right choice at all, see deterministic versus probabilistic entity resolution.

Whatever records the outcome, a reader should be able to recover the m and u values, prior and thresholds behind each link; that is the property a system of record such as Reliable Context should let them check.

Sources

1. Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage", JASA 64(328):1183-1210, DOI record: bibliographic origin of the model (paywalled; not used for formulas). Abstract read via OpenAlex, https://openalex.org/W2073471108.
2. Winkler, W. E. (2006), "Overview of Record Linkage and Current Research Directions", US Census Bureau Research Report Statistics #2006-2: likelihood ratio, decision rule and thresholds, additive weights under conditional independence, EM estimation, frequency-based matching, blocking, dependence examples.
3. Binette, O. and Steorts, R. C., "(Almost) All of Entity Resolution", arXiv:2008.04443: three actions, W(γ), optimality, posterior probability, Jaro's EM, assumptions and extensions.
4. Splink, "The Fellegi-Sunter model": m, u and λ definitions, log2 match weights, conversion to probability, weight-probability anchors.
5. Splink, "Training rationale": direct λ and u estimation, EM for m on blocked pairs, bias warning.
6. Splink, "Term-Frequency Adjustments": value-specific m and u, the log2(u/T_X) adjustment, toy example.
7. Splink documentation home: Splink based on Fellegi-Sunter; advice on correlated columns.
8. Linacre, R., "The mathematics of the Fellegi Sunter model": Bayes-theorem form of the match probability, comparison levels.

9. Linacre, R., "The Intuition Behind the Use of Expectation Maximisation to Train Record Linkage Models": EM steps, local maxima.
10. Papadakis, G. et al., "A Survey of Blocking and Filtering Techniques for Entity Resolution", arXiv:1905.06167: blocking as an efficiency technique.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.census.gov/content/dam/Census/library/working-papers/2006/adrm/rrs2006-02.pdf
2. https://doi.org/10.1080/01621459.1969.10501049
3. https://openalex.org/W2073471108
4. https://arxiv.org/abs/2008.04443
5. https://moj-analytical-services.github.io/splink
6. https://moj-analytical-services.github.io/splink/topic_guides/theory/fellegi_sunter.html
7. https://www.robinlinacre.com/maths_of_fellegi_sunter
8. https://www.robinlinacre.com/em_intuition
9. https://moj-analytical-services.github.io/splink/topic_guides/training/training_rationale.html
10. https://moj-analytical-services.github.io/splink/topic_guides/comparisons/term-frequency.html
11. https://arxiv.org/abs/1905.06167

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.