Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

Record Linkage for Company Names: Fellegi-Sunter Explained

> Build a more reviewable acquisition pipeline with Docket. Our deal-origination software and managed research help private equity teams screen companies against a mandate and understand the evidence behind each finding.

Published by Docket · 2026-10-29 · entity resolution · entity resolution process · record linkage · company name matching

Original article: https://docket.capital/articles/fellegi-sunter-record-linkage-company-names



In this article

1. The problem: same company, many records
2. Deterministic vs probabilistic linkage
3. Blocking: deciding which pairs to compare
4. Comparison vectors for company records
5. m and u probabilities and match weights
6. Thresholds and clerical review
7. Measuring error rates
8. Practical limits for company data

The problem: same company, many records

A target list assembled from more than one source will contain the same company several times. A state registry lists "Harbor Valve & Fitting, Inc." A trade directory lists "Harbor Valve and Fitting." A commercial database lists "Harbour Valve & Fittings Co" at an address the company left two years ago. No identifier is shared, so a join on exact name returns nothing and a join on address returns the wrong neighbors.

This is the record linkage problem. Jared Murray describes probabilistic record linkage as "the task of merging two or more databases in the absence of a unique identifier" and notes that it is closely related to deduplicating a single database (Murray, arXiv:1603.07816). In deal research the same work is usually called entity resolution, one of several names for the field listed in a Census Bureau overview (Winkler, Census Bureau Research Report 2006-2).

Company records are a hard case. Winkler writes that business lists "are typically more difficult to match" than person or agriculture lists "because of the variants of names and the variants of addresses" (Winkler 2006). Exact comparison fails even on person data: in one major census application, Winkler reports that more than 25% of matches would not have been found by exact character-by-character comparison.

This article explains the Fellegi-Sunter model with company-name examples. For applied procedures, see the guides on identifying the same company across state registrations and finding the legal entity behind a brand.

Deterministic vs probabilistic linkage

Deterministic linkage applies rules: "same normalized name and same ZIP code," or "same state registration number." The documentation for Splink, which describes itself as "a Python package for probabilistic record linkage" (Splink), calls this "a rules-based approach for joining records together." It lists the strengths as computational cheapness and high precision, and the weakness as low recall, meaning missed true matches (Splink, Probabilistic vs Deterministic linkage). A Census Bureau working paper on linking employers adds that, because deterministic matching keeps only records with exactly one match, the matched file can end up as "a small and potentially non-random subset of the original file" (Abowd et al., CES-WP-19-08).

Probabilistic linkage weighs evidence from each field comparison against thresholds. Splink calls it "a evidence-based approach for joining records together" in which the user chooses the evidence threshold for a match (Splink).

The formal model is Ivan Fellegi and Alan Sunter's "A Theory for Record Linkage," published in the *Journal of the American Statistical Association* in December 1969 (Fellegi and Sunter 1969). Winkler describes it as a formal mathematical model for ideas introduced by Howard Newcombe in 1959 and 1962 (Winkler 2006).

Blocking: deciding which pairs to compare

Comparing every record with every other is usually infeasible: Splink notes that comparisons grow with approximately the square of the number of records, and that 1 million records "would generate around 500 billion pairwise record comparisons" (Splink, What are Blocking Rules?).

Blocking reduces that set. Winkler describes Newcombe's approach as "only considering pairs that agreed on a characteristic such as surname or date-of-birth," and adds: "Such reduction in the number of pairs is called blocking" (Winkler 2006). Pairs that share no blocking key are never scored; they are nonmatches by construction. The CES employer study states this for its own design, which blocked on 3-digit ZIP code, 10-digit phone number, telephone area code and city-state: a record sharing none of those values "is assumed to have 0 probability of being a true match" (CES-WP-19-08).

The trade-off: looser rules catch more true matches but generate more comparisons, and a name-based rule "would miss a true match where there was a typo in the name." Splink therefore accepts several rules and keeps every comparison that satisfies at least one (Splink, blocking rules). Winkler gives a similar multi-pass example (ZIP plus the start of last name; ZIP plus the start of street name; name plus street address), notes that blocking strategies have historically been "determined using trial and error," and warns that a set of passes generally will not bring together every pair that would score above the cutoff (Winkler 2006).

Illustrative example (fictional). A team deduplicates 40,000 company records. All pairs: 40,000 × 39,999 / 2 = 799,980,000. The team blocks on the first three letters of the normalized name plus state, and suppose the result is 2,000 blocks of 20 records. Each block has 20 × 19 / 2 = 190 pairs, so the total is 2,000 × 190 = 380,000. A second pass on phone number would recover pairs whose names differ in the first three letters.

Comparison vectors for company records

Inside a block, each field comparison is recorded as a level; the levels form the pair's comparison vector, which Winkler calls an agreement pattern. His simple example is a space of eight patterns built from agreement or disagreement on the largest name component, street name and street number (Winkler 2006).

A workable starting set for company records:

Field	Preparation	Levels
Company name	Standardize suffixes and common words, lowercase, strip punctuation	exact / close / different
Street address	Standardize street types and abbreviations	agree / disagree
Phone	Digits only	agree / disagree
Web domain	Lowercase, drop "www."	agree / disagree
Registry or tax ID	As issued	agree / disagree (where present)

Standardization. Winkler describes replacing variant spellings with one, for example variants of "Incorporated" with "Inc." and "Street" with "St," then parsing names and addresses into comparable components; his name-parsing examples include business-name components (Winkler 2006).

String comparators. The "close" level needs a similarity score. Winkler explains that linkage needs "a function that represents approximate agreement," scored 1 for exact agreement and between 0 and 1 for partial agreement. He describes the Jaro comparator, which accounts for insertions, deletions and

transpositions, and a variant he reports dramatically improves matching compared with not using string comparators (Winkler 2006). The CES study used Jaro-Winkler scores for employer name and establishment address as its first two predictors (CES-WP-19-08).

m and u probabilities and match weights

The model attaches two probabilities to each comparison level. Splink defines them (Splink, The Fellegi-Sunter Model):

- **m probability:** "the probability of a given observation given the records are a match."
- **u probability:** "the probability of a given observation given the records are not a match."
- **lambda (λ):** "the prior probability that any two records match."

The ratio m/u is the Bayes factor; its base-2 logarithm is the partial match weight. For a two-level field, disagreement has probabilities (1 - m) and (1 - u), so:

- agreement weight = log2(m / u)
- disagreement weight = log2((1 - m) / (1 - u))

Splink's key assumption is that comparisons on different columns are independent, which makes weights additive. The total match weight is the prior weight, log2(λ / (1 - λ)), plus the partial weights, and converts to a probability as 2^M / (1 + 2^M) (Splink). Winkler writes the core quantity as the ratio R = P(γ | M) / P(γ | U) and calls R, or any monotonically increasing function of it such as a log, the matching weight (Winkler 2006).

Worked example (fictional numbers)

Illustrative example (fictional). Every m and u value below is invented. Weights are rounded to two decimals; totals and probabilities were computed before rounding.

Field	Level	m	u	m / u	Weight (log2)
Name	exact	0.70	0.002	350	8.45
Name	close	0.20	0.010	20	4.32
Name	different	0.10	0.988	0.101	-3.30
Address	agree	0.80	0.010	80	6.32
Address	disagree	0.20	0.990	0.202	-2.31
Phone	agree	0.60	0.001	600	9.23
Phone	disagree	0.40	0.999	0.400	-1.32
Domain	agree	0.50	0.0005	1,000	9.97
Domain	disagree	0.50	0.9995	0.500	-1.00

The three name levels sum to 1.00 across m and across u. For binary fields, disagreement is the complement: for address, (1 - 0.80) / (1 - 0.01) = 0.20 / 0.99 = 0.202, and log2(0.202) = -2.31.

Assume a fictional prior of 1 true match per 1,000 pairs within blocks: λ = 0.001, prior weight log2(0.001 / 0.999) = -9.96. A missing field contributes 0.

Four fictional candidate pairs for "Harbor Valve & Fitting, Inc.":

Pair	Name	Address	Phone	Domain	Field weights	Prior	Total	Match probability
A: "Harbor Valve and Fitting Co"	exact (8.45)	agree (6.32)	agree (9.23)	missing (0)	24.00	-9.96	14.04	0.9999
B: "Harbour Valve & Fittings"	close (4.32)	disagree (-2.31)	disagree (-1.32)	agree (9.97)	10.66	-9.96	0.70	0.62
C: "Harbor Valves Supply"	close (4.32)	agree (6.32)	disagree (-1.32)	disagree (-1.00)	8.32	-9.96	-1.64	0.24
D: "Bayline Marine Electric"	different (-3.30)	agree (6.32)	disagree (-1.32)	disagree (-1.00)	0.70	-9.96	-9.27	0.0016

Pair B as arithmetic: 4.32 - 2.31 - 1.32 + 9.97 = 10.66; 10.66 - 9.96 = 0.70; 2^0.70 / (1 + 2^0.70) = 0.62. Pair D's displayed values sum to -9.26; the unrounded total is -9.27.

Two points carry over to real data. The shared domain in pair B outweighs a moved address and a changed phone, because in this parameter set non-matching pairs almost never share a domain. And pair D shares an address yet scores low: companies share buildings and registered-agent addresses, so a real model's u for address agreement should reflect that.

Independence is the weak assumption for company data: address and phone often change together when a business moves, so their disagreements may be double-counted. The CES authors, discussing their employer setting, call the independence assumption "untenable" (CES-WP-19-08).

Real m and u values come from labeled pairs or from estimation without labels. Winkler cites methods for optimal linkage both with and without training data (Winkler 2006); the CES paper notes approaches that use the expectation-maximization (EM) algorithm "without a training dataset," while its own model was trained on about 1,000 hand-reviewed pairs (CES-WP-19-08). Winkler also describes frequency-based weighting, in which agreement on a rare value counts for more than agreement on a common one. Agreement on a distinctive company name is stronger evidence than agreement on "ABC Plumbing."

Thresholds and clerical review

The decision rule has two cutoffs. As Winkler states it (Winkler 2006):

- if R is above the upper cutoff, designate the pair a match;
- if R is between the cutoffs, designate it a possible match "and hold for clerical review";
- if R is below the lower cutoff, designate it a nonmatch.

The cutoffs "are determined by a priori error bounds on false matches and false nonmatches." Raising the upper cutoff lowers false matches; lowering the lower cutoff lowers false nonmatches. Both send more pairs to review.

Illustrative example (fictional), continued. Set the upper cutoff at a match probability of 0.95 (total weight 4.25) and the lower at 0.05 (weight -4.25). Pair A is a match, pair D a nonmatch. Pairs B and C go to a reviewer, who checks evidence outside the four fields, such as a registry filing or the company's website, and records the decision with that evidence.

At scale the savings can be large: Winkler reports that in the 1992 U.S. Census of Agriculture, "the computerized procedures were able to reduce the clerical review from 75 clerks over 3 months to 6500 hours of review." The listed entities were farms corresponding to individuals, partnerships and corporations. Based on a large validation follow-up, the computerized procedures identified more duplicates automatically than the clerks had in the previous census, and duplication in the final file was 2%, against 10% five years earlier (Winkler 2006).

Clerical review has weaknesses of its own. The CES paper summarizes Belin and Rubin's (1995) criticism: the error properties of manual review are unknown, standards may vary across reviewers, and review may not yield many unequivocal matches. The CES team had each training pair reviewed by two different human experts (CES-WP-19-08). For a deal team: a written rubric, a second reviewer on a sample and a disagreement log.

Pairwise decisions also need clustering: if A links to B and B links to C, the three become one entity even if A and C scored low against each other. Winkler cites graph-based methods that partition linkage relationships "into clusters representing individual entities" (Winkler 2006). Write the clustering rule down; one false link can merge two companies.

Measuring error rates

A linkage run makes two kinds of error:

- **False match:** two companies linked as one, merging their ownership, revenue or contact facts.
- **False nonmatch:** one company left as two records, counted and contacted twice.

Precision is the share of linked pairs that are true matches; recall is the share of true matches that were linked. Splink frames the deterministic trade-off in these terms (Splink).

Measuring them takes labeled pairs:

1. Sample randomly from each band: auto-match, clerical review, and auto-nonmatch above a floor score.
2. Label each sampled pair using evidence beyond the compared fields; have a second reviewer label a subset.
3. Estimate the false match rate from the auto-match sample and the false nonmatch rate from the nonmatch sample.

4. Treat pairs that blocking never compared as a separate risk. Winkler describes a capture-recapture method for estimating matches missed by blocking and cautions that it can be inaccurate (Winkler 2006).
5. Adjust cutoffs and re-run.

The audit sampling worksheet covers sample sizes. Rates measured on a labeled sample are verified for that sample; misses outside the blocks remain open.

Open implementations exist. Enamorado, Fifield and Imai (*American Political Science Review*, 2019) describe "a fast and scalable algorithm to implement a canonical model of probabilistic record linkage," and their abstract states that an open-source software package is available (Enamorado, Fifield and Imai 2019). Splink is another implementation of the Fellegi-Sunter model.

Practical limits for company data

The model answers one question: do two records describe the same entity? Company data raises others.

- **Name changes and successors.** Whether a renamed or merged company is "the same" is a definitional choice first.
- **Brands vs legal entities.** A trade name, its operating LLC and a holding company can share phone, domain and address while being distinct entities. The model will link them unless the design separates them.
- **Shared addresses.** Office parks, registered agents and virtual offices raise u for address agreement.
- **Many small candidates.** The CES study found matching blocks populated with many small employers, each receiving a trivial match probability, a problem made worse when EINs could not be used for blocking (CES-WP-19-08).

Identifiers beat probabilities when they exist. The CES paper describes Census Bureau longitudinal business data as starting with exact identifier linkage (Census File Number to EIN) and using Fellegi-Sunter methods "when EINs produced many-to-many linkages," or none, as with births, deaths and changes in legal identifiers (CES-WP-19-08).

Decision rule:

- [ ] Link on a shared, reliable identifier first (registry number, tax ID, license number).
- [ ] Use probabilistic linkage where no identifier is shared, or where identifiers conflict.
- [ ] Send the middle band to clerical review with a written rubric.
- [ ] Measure false match and false nonmatch rates on a labeled sample; record blocking misses as open.
- [ ] Keep the evidence for every reviewed decision with the record.

Further reading: Fellegi and Sunter's 1969 paper is the original statement of the model; Winkler's 2006 overview is an open survey from the Census Bureau; Murray's 2016 paper reviews how indexing, blocking and filtering interact with the Fellegi-Sunter framework; and Peter Christen's *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection* (Springer, 2012) is a book-length treatment (Christen 2012).

In Docket, the Triage agent handles this step for a client's acquisition targets: it resolves identity and removes duplicates.

Source links

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

1. https://arxiv.org/abs/1603.07816
2. https://www.census.gov/content/dam/Census/library/working-papers/2006/adrm/rrs2006-02.pdf
3. https://moj-analytical-services.github.io/splink
4. https://moj-analytical-services.github.io/splink/topic_guides/theory/probabilistic_vs_deterministic.html
5. https://www2.census.gov/ces/wp/2019/CES-WP-19-08.pdf
6. https://doi.org/10.1080/01621459.1969.10501049
7. https://moj-analytical-services.github.io/splink/topic_guides/blocking/blocking_rules.html
8. https://moj-analytical-services.github.io/splink/topic_guides/theory/fellegi_sunter.html
9. https://doi.org/10.1017/s0003055418000783
10. https://doi.org/10.1007/978-3-642-31164-2

THE PUBLISHER

About Docket

Docket provides deal-origination research software and managed research for private equity firms. We help investment teams investigate acquisition targets using structured screening criteria, retained sources and reviewable company evidence. Teams can work through a self-serve platform or use managed research, depending on how they want research delivered.

Research against a defined mandate

A useful target list needs more than company names. Docket focuses on the questions that determine whether a company fits an investment mandate, including the evidence needed to support or qualify each answer. Our research approach makes the connection between screening criteria, source material and conclusions visible to the team reviewing the work.

Triage, Scout and Audit

Docket's three named research agents perform complementary tasks. Triage resolves company identity and screens fit. Scout collects sourced answers against the mandate. Audit checks retained evidence, addresses contradictions and leaves unsupported answers visibly unresolved. This structure helps reviewers distinguish established findings from missing information and questions requiring further investigation.

Evidence that supports investment-team judgment

Our research library covers market mapping, screening criteria, private-company data, succession and ownership, source evaluation and evidence standards. These resources explain the methods and limitations behind origination research. Findings support a team's judgment; they do not establish that a company is for sale or guarantee a transaction or investment outcome.

Work with Docket

Visit Docket to explore the platform, managed research and the current contact path. Read about screening criteria, evidence standards, AI research agents and market mapping.

Public examples are illustrative unless explicitly identified otherwise. Research preparation and authorized outreach are separate activities; confidential target lists and customer outcomes should never be inferred from an educational example.

Website: https://docket.capital

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.