

Joins as Selection Problems: Objective Misspecification in Record Linkage and Table Integration

April 3, 2026

Abstract

Standard record linkage scores candidate pairs and then solves an assignment problem. That formulation is exact only when the quality of the linked table is separable over its selected links. Many empirical joins are not evaluated that way. The analyst cares about a property of the linked table as a whole: hierarchical nesting, distributional calibration, referential integrity, temporal continuity, or consistency across multiple joins. We formalize linkage as a selection problem over feasible linked tables. Additive pairwise scoring yields a surrogate objective; the analyst’s actual problem typically adds a non-separable table-level criterion or imposes table-level constraints. This perspective connects naturally to Morris’s finite selection model (FSM): when the criterion is global, candidate links should be evaluated by their contribution to the evolving table, not in isolation. We show that every penalized optimum is optimal for the constrained problem at the coherence level it attains, that one-to-one disagreements with the truth are combinatorially coupled through permutation cycles, and that downstream bias is bounded by distortion in the analysis labels induced by the join. A household example illustrates the general framework. In a reproducible simulation, the structured optimizer raises person-level accuracy from 0.291 under Hungarian assignment to 0.542, raises exact household recovery from 0.110 to 0.576, and lowers absolute bias from 4.139 to 2.595. The example is specific; the lesson is general.

1 Introduction

Classical record linkage begins with pairwise evidence. Candidate record pairs are compared, similarity scores are assigned, and the final linkage is obtained by classification or by solving a one-to-one assignment problem (Fellegi and Sunter, 1969). When uniqueness constraints matter, the Hungarian method is a natural optimizer because the objective is additive over matched pairs (Kuhn, 1955). That workflow is often exactly right.

The difficulty is that many linked-data problems are not evaluated pair by pair. The object that matters is the linked table. A useful join may need to preserve household membership, panel continuity, foreign-key relationships, cluster composition, or known marginal distributions. These are properties of the output table as a whole, not sums over its rows. In such cases, solving the pairwise assignment problem exactly does not solve the analyst’s problem. It solves a surrogate.

The core abstraction is therefore broader than “household coherence matters.” It is a mismatch between the *unit of optimization* and the *unit of analysis*. Standard linkage optimizes over candidate links. The analyst’s loss function operates on the linked table. Whenever the table must satisfy a global property—hierarchical, distributional, temporal, relational, or compositional—pairwise optimization is misspecified.

This paper develops that claim in four steps. First, we formalize joins as selection problems over feasible linked tables. The additive assignment formulation is one special case, and it is exact only when the analyst’s criterion is representable as a linear functional of selected links. Otherwise the natural formulations are constrained or penalized set-level problems. Second, we connect this perspective to the finite selection model (FSM) introduced by Morris (1979) for the RAND Health Insurance Experiment and recently formalized and extended by Chattopadhyay et al. (2026). The original application was experimental design, but the principle is more general: when the criterion is non-separable, local optimization solves the wrong problem. Third, we state explicit conditions under which a structured or FSM-style criterion can reduce a bound on downstream bias: namely, when it reduces distortion in the analysis labels that enter the estimand. Fourth, we work out households as a minimal example and quantify the resulting gains in a fully reproducible simulation.

The household example is not the thesis of the paper. It is a clean instantiation of a more general idea. A person-level optimizer spends pairwise evidence record by record. A household-level analyst uses the joined table household by household. The same logic applies to entity resolution, calibrated linkage, and multi-table integration.

2 Joins as selection problems

2.1 General setup

Let A and B denote two source tables. Let $E \subseteq A \times B$ be the set of admissible candidate links after blocking or other preliminary restrictions. A *join* is a subset $J \subseteq E$ belonging to a feasible family \mathcal{J} that encodes application constraints such as one-to-one matching, cardinality limits, compatibility restrictions, or higher-order structure. The selected links induce a linked output table $T(J)$.

For each candidate link $e \in E$, let $s_e \in \mathbb{R}$ denote its pairwise score. The standard linkage objective is the additive score

$$S(J) = \sum_{e \in J} s_e. \quad (1)$$

The additive rule solves

$$J^{\text{add}} \in \arg \max_{J \in \mathcal{J}} S(J). \quad (2)$$

To describe the analyst’s actual problem, let $\Psi(T(J))$ be a table-level criterion. Depending on the application, Ψ may measure exact recovery of a hierarchy, agreement with known margins, referential integrity, temporal consistency, transitivity in entity resolution, or any other property of the linked table that is not simply the sum of rowwise scores.

Two standard formulations are then

$$\max_{J \in \mathcal{J}} S(J) \quad \text{subject to} \quad \Psi(T(J)) \geq \kappa, \quad (3)$$

and

$$\max_{J \in \mathcal{J}} U_\lambda(J) \equiv S(J) + \lambda \Psi(T(J)), \quad \lambda \geq 0. \quad (4)$$

The constrained form is natural when the analyst knows a minimum acceptable level of coherence or calibration. The penalized form is its Lagrangian relaxation.

The distinction between the additive problem (2) and the structured problems (3)–(4) is exact, not rhetorical. If the analyst’s criterion can be written as

$$S(J) + \lambda \Psi(T(J)) = c + \sum_{e \in E} w_e \mathbf{1}\{e \in J\}$$

on the feasible family \mathcal{J} for some weights w_e , then the problem is separable and can be absorbed into edge weights. The interesting case begins exactly when no such representation exists. Then the value of a selected link depends on what else is in the join.

2.2 Penalty and constraint

Proposition 2.1 (Penalized optimum implies constrained optimum at the attained level). *Fix $\lambda \geq 0$, and let $\hat{J}_\lambda \in \arg \max_{J \in \mathcal{J}} U_\lambda(J)$. Write $\kappa_\lambda = \Psi(T(\hat{J}_\lambda))$. Then \hat{J}_λ solves the constrained problem (3) with threshold $\kappa = \kappa_\lambda$.*

Proof. Take any $J \in \mathcal{J}$ such that $\Psi(T(J)) \geq \kappa_\lambda$. If $S(J) > S(\hat{J}_\lambda)$, then

$$U_\lambda(J) = S(J) + \lambda \Psi(T(J)) \geq S(J) + \lambda \kappa_\lambda > S(\hat{J}_\lambda) + \lambda \kappa_\lambda = U_\lambda(\hat{J}_\lambda),$$

contradicting optimality of \hat{J}_λ . □

Proposition 2.1 places the structured join problem beside balance-constrained matching and sampling formulations in which one optimizes an additive criterion subject to sample-level restrictions (Zubizarreta, 2012; Deville and Sørndal, 1992; Deville and Tillé, 2004). It also clarifies the exact scope of the penalty–constraint connection used in the paper: every penalized optimum is a constrained optimum at the coherence level it attains. The converse does not automatically hold without additional regularity.

A minimal misspecification inequality follows immediately.

Proposition 2.2 (When the additive optimum is not the analyst’s optimum). *Let $J_0, J_1 \in \mathcal{J}$ satisfy $S(J_0) > S(J_1)$ and $\Psi(T(J_1)) > \Psi(T(J_0))$. Then J_1 is preferred by the penalized objective (4) whenever*

$$\lambda > \frac{S(J_0) - S(J_1)}{\Psi(T(J_1)) - \Psi(T(J_0))}. \quad (5)$$

Proof. Subtract the two penalized objectives:

$$U_\lambda(J_1) - U_\lambda(J_0) = (S(J_1) - S(J_0)) + \lambda(\Psi(T(J_1)) - \Psi(T(J_0))).$$

The right-hand side is positive exactly when (5) holds. □

Proposition 2.2 is elementary, but it isolates the central point of the paper. Solving the wrong problem exactly does not recover the right answer. Exact optimization of the additive surrogate is compatible with strict suboptimality for the analyst’s actual objective.

2.3 Error cascades under one-to-one matching

The one-to-one case contains another structural fact that pairwise scoring obscures: disagreements with the truth are combinatorially coupled.

Specialize now to equal-size one-to-one linkage. Let Σ_n denote the set of permutations of $[n]$, and let $\sigma^* \in \Sigma_n$ denote the true bijection. For any candidate linkage $\sigma \in \Sigma_n$, define the disagreement set

$$D(\sigma, \sigma^*) = \{i \in [n] : \sigma(i) \neq \sigma^*(i)\}.$$

Proposition 2.3 (Cascade structure of one-to-one disagreements). *For any $\sigma, \sigma^* \in \Sigma_n$, the disagreement set $D(\sigma, \sigma^*)$ decomposes into disjoint cycles of length at least 2. In particular, if $\sigma \neq \sigma^*$, then $|D(\sigma, \sigma^*)| \geq 2$.*

Proof. Consider the permutation $\tau = (\sigma^*)^{-1} \circ \sigma$ on $[n]$. The indices fixed by τ are exactly those for which $\sigma(i) = \sigma^*(i)$. Every finite permutation decomposes into disjoint cycles, and every non-fixed point belongs to a cycle of length at least 2. The disagreement set is therefore the union of those nontrivial cycles. \square

Proposition 2.3 does not say that linkage errors are probabilistically independent or dependent. It says something more basic: under bijectivity, they cannot arise one at a time. A mistaken assignment necessarily displaces at least one more record. In structured linkage problems the cascade can be much larger: one bad household link, one broken foreign key, or one inconsistent temporal match can force a chain of compensating errors elsewhere in the table.

3 The finite selection model and when it helps

3.1 FSM as a general principle

Morris’s finite selection model was introduced for experimental design in the RAND Health Insurance Experiment (Morris, 1979; Newhouse and Insurance Experiment Group, 1993). There is a finite ground set of units, a feasible family of allocations, and a global design criterion. Treatment groups take turns selecting units in a fair random order. On each turn, the active group chooses the available unit that most improves the common criterion. The recent reformulation by Chattopadhyay et al. (2026) makes explicit that the important object is the incremental gain in a *global* criterion, not an isolated unit score.

Write \mathcal{U} for a finite ground set and $\mathcal{F} \subseteq 2^{\mathcal{U}}$ for the feasible family. Let $Q : \mathcal{F} \rightarrow \mathbb{R}$ be the objective. The FSM constructs a selection $X \in \mathcal{F}$ sequentially. At step t , with partial selection X_t , the next admissible element is evaluated by its incremental contribution

$$\Delta_Q(u \mid X_t) = Q(X_t \cup \{u\}) - Q(X_t).$$

If Q is additive, then $\Delta_Q(u \mid X_t)$ does not depend on X_t and the problem collapses to ordinary scoring. The only reason to invoke FSM thinking is that the value of u depends on the current partial selection.

That principle carries over directly to linkage. The ground set can be candidate record links, candidate block matches, or tuples in a multiway join. The feasible family encodes one-to-one,

hierarchical, calibration, or consistency restrictions. The criterion can be the penalized or constrained objective from Section 2. Randomized turn-taking is essential in experiments because design validity itself depends on randomization. In linkage, randomness is optional. What matters is the conceptual move from independent pairwise scoring to incremental evaluation against a table-level criterion.

The present paper does not implement a literal randomized FSM. We use an exact structured optimizer in the household example so that the comparison isolates *objective choice* rather than approximation error. An FSM-style sequential algorithm would be the natural next step when exact optimization over the structured feasible set is computationally infeasible.

3.2 When a structured criterion can reduce bias

To state when a structured criterion can help downstream inference, let U_i^\star denote the analysis label for record i under the true join. Depending on the application, U_i^\star may be a household identifier, a treatment label, a classroom identifier, a time index, a foreign-key parent, or some other attribute imported through linkage. Under a candidate join J , let $\tilde{U}_i(J)$ denote the label induced by the linked table $T(J)$.

Suppose the analyst's target is

$$\theta^\star = \frac{1}{n} \sum_{i=1}^n q(Y_i, U_i^\star), \quad (6)$$

and the linked-table analogue is

$$\theta(J) = \frac{1}{n} \sum_{i=1}^n q(Y_i, \tilde{U}_i(J)). \quad (7)$$

Assume there is a metric ρ on the label space and a constant $L < \infty$ such that

$$|q(y, u) - q(y, u')| \leq L|y| \rho(u, u') \quad \text{for all } y, u, u'. \quad (8)$$

Proposition 3.1 (Bias is bounded by distortion in analysis labels). *Under (6)–(8),*

$$|\theta(J) - \theta^\star| \leq \frac{L}{n} \sum_{i=1}^n |Y_i| \rho(\tilde{U}_i(J), U_i^\star). \quad (9)$$

Consequently, for any random linkage rule R producing J_R ,

$$|\mathbb{E}[\theta(J_R)] - \theta^\star| \leq \frac{L}{n} \sum_{i=1}^n \mathbb{E}[|Y_i| \rho(\tilde{U}_i(J_R), U_i^\star)]. \quad (10)$$

Proof. Using the triangle inequality and (8),

$$|\theta(J) - \theta^\star| \leq \frac{1}{n} \sum_{i=1}^n |q(Y_i, \tilde{U}_i(J)) - q(Y_i, U_i^\star)| \leq \frac{L}{n} \sum_{i=1}^n |Y_i| \rho(\tilde{U}_i(J), U_i^\star),$$

which gives (9). Taking expectations yields (10). \square

Proposition 3.1 states the condition transparently. A structured or FSM-style rule can lower an upper bound on downstream bias when it reduces distortion in the analysis labels that enter the estimand. It need not help when the added structure is irrelevant to the estimand, unavailable, or misspecified.

In the simplest case, if a wrong join only changes a binary analysis label and outcomes are linear in that label, bias is controlled directly by how often the join changes the label.

Corollary 3.2 (Binary label misclassification and attenuation). *Suppose $U_i^* = Z_i \in \{0, 1\}$ is a balanced binary label, $\tilde{U}_i(J) = W_i(J)$, and outcomes satisfy*

$$Y_i = \alpha + \tau Z_i + u_i, \quad \mathbb{E}[u_i \mid Z_i] = 0. \quad (11)$$

Assume $W_i(J_R) = Z_i \oplus E_i$ under linkage rule R , where $E_i \sim \text{Bernoulli}(\varepsilon_R)$ is independent of (Z_i, u_i) and \oplus denotes exclusive OR. Then the linked difference-in-means estimator based on $W_i(J_R)$ satisfies

$$\mathbb{E}[\hat{\tau}_R] = (1 - 2\varepsilon_R)\tau. \quad (12)$$

Hence any linkage rule with a smaller induced misclassification rate ε_R is less biased in absolute value.

Proof. Under balance and symmetric misclassification,

$$\Pr(Z_i = 1 \mid W_i = 1) = 1 - \varepsilon_R, \quad \Pr(Z_i = 1 \mid W_i = 0) = \varepsilon_R.$$

Therefore

$$\mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0] = \tau((1 - \varepsilon_R) - \varepsilon_R) = (1 - 2\varepsilon_R)\tau.$$

□

Remark. Corollary 3.2 is deliberately narrow. It relies on a balanced binary label, a linear outcome model, and symmetric misclassification. Outside that setting, linkage error need not attenuate. With continuous labels, asymmetric mistakes, or more complex estimands, bias can move in either direction.

4 Three sources of non-separability in table linkage

The household example below is just one instance of a broader phenomenon. At least three recurring sources of non-separability appear in record linkage and data integration.

Hierarchical or relational structure. Households, classrooms, firms with subsidiaries, orders with line items, parent–child tables, and entity-resolution clusters all impose structure above the individual record. In entity resolution this is the canonical issue: pairwise match decisions are followed by transitive closure or clustering, but transitivity is a property of the cluster, not of any one pair. Collective entity-resolution methods and correlation clustering make that non-separability explicit by optimizing over partitions or graphs rather than isolated pairs (Christen, 2012; Bansal et al., 2004). Household linkage is a matched-block version of the same problem.

Distributional calibration. Sometimes the linked file should reproduce known margins or moments: age distributions, school sizes, treatment shares, or other calibration totals. Those constraints are free information about the target table. Ignoring them and maximizing an unconstrained additive

score wastes information in exactly the way that ignoring auxiliary covariates wastes information in sampling or matching. Calibration estimators, entropy balancing, and the cube method all solve this kind of problem by enforcing sample-level balance, not unit-level fit (Deville and Sørndal, 1992; Hainmueller, 2012; Deville and Tillé, 2004). A *calibrated join* is the analogous object for linkage.

Cross-table or temporal consistency. Real integrations rarely stop at one two-table join. If A is linked to B and B to C , then the implied A – C relation should make sense. Likewise, person-period joins in panels should preserve feasible temporal sequences. These are consistency conditions across joins or across time. They are non-separable even when each component join is one-to-one.

Across all three cases, the common failure mode is the same. A flat pairwise score is used to optimize an objective that actually lives on the output table. The more informative the higher-order structure, the more costly that misspecification can be.

5 A worked household example

Households are a useful worked example because they make the mismatch between optimization and analysis especially transparent. The join is selected person by person, but the downstream analysis imports a household-level label.

5.1 Exact household-preserving linkage

Suppose the records in each file are partitioned into households,

$$\mathcal{G}_A = \{g_1, \dots, g_K\}, \quad \mathcal{G}_B = \{h_1, \dots, h_K\},$$

with compatible household sizes. The exact household-preserving feasible family consists of those one-to-one linkages that map each g_k entirely into one h_ℓ .

For any admissible household pair (g_k, h_ℓ) , define the compatibility score

$$C_{k\ell} = \max_{\varphi \in \Pi(g_k, h_\ell)} \sum_{i \in g_k} s_{i, \varphi(i)}, \quad (13)$$

where $\Pi(g_k, h_\ell)$ is the set of bijections from members of g_k to members of h_ℓ .

Proposition 5.1 (Decomposition of the exact household-preserving optimum). *The exact household-preserving additive problem is equivalent to the household-level assignment*

$$\max_{\pi \in \Sigma_K} \sum_{k=1}^K C_{k, \pi(k)}. \quad (14)$$

An optimizer of the person-level problem is obtained by solving (14) and then using, within each matched household pair $(g_k, h_{\pi(k)})$, the bijection that attains $C_{k, \pi(k)}$.

Proof. Any household-preserving person-level bijection can be written as a household permutation $\pi \in \Sigma_K$ together with within-household bijections $\varphi_k \in \Pi(g_k, h_{\pi(k)})$. The additive score therefore decomposes as

$$\sum_{k=1}^K \sum_{i \in g_k} s_{i, \varphi_k(i)}.$$

For a fixed π , the terms indexed by k are independent, so each is maximized by $C_{k,\pi(k)}$. Maximizing over π yields (14). \square

Proposition 5.1 is why the benchmark in the simulation is useful. It is not a heuristic household-first rule. It is the exact optimizer of the additive score over the structurally correct feasible set.

5.2 A motivated two-household example

The mechanism appears in the smallest nontrivial case. Let file A contain households $g_1 = \{a_1, a_2\}$ and $g_2 = \{a_3, a_4\}$, and let file B contain $h_1 = \{b_1, b_2\}$ and $h_2 = \{b_3, b_4\}$. Suppose the true correspondence is $g_1 \leftrightarrow h_1$ and $g_2 \leftrightarrow h_2$, but one member in each household has a particularly plausible decoy in the neighboring household. A person-level score matrix of that form is

$$\begin{pmatrix} 0.95 & 0.04 & 0.07 & 0.10 \\ 0.05 & 0.92 & 0.06 & 1.01 \\ 0.08 & 0.07 & 0.94 & 0.06 \\ 0.09 & 1.00 & 0.05 & 0.93 \end{pmatrix}. \quad (15)$$

The additive optimizer prefers the crossover assignment

$$a_1 \mapsto b_1, \quad a_2 \mapsto b_4, \quad a_3 \mapsto b_3, \quad a_4 \mapsto b_2,$$

with score 3.90, over the household-preserving identity assignment with score 3.74. So this is a genuine case in which the additive optimizer sacrifices the analyst's objective in order to gain pairwise score.

Now suppose the downstream analysis imports a household treatment label from file B . If household g_1 is treated and g_2 is control, the crossover join swaps one treated and one control label. In the balanced symmetric setting of Corollary 3.2, that induced misclassification attenuates the linked treatment effect. The point of the example is not that all linkage error attenuates. It is that small pairwise gains can be inferentially costly once the analysis label lives at household level.

5.3 Local ambiguity and household breakage

To connect the example to ambiguity sweeps, consider one file- A household $g = \{a_1, \dots, a_m\}$ with true counterpart $h = \{b_1, \dots, b_m\}$ and a nearby decoy household $h' = \{b'_1, \dots, b'_m\}$. For role r , define the margin

$$D_r(\alpha) = s(a_r, b_r) - s(a_r, b'_r), \quad (16)$$

where α indexes ambiguity. Positive $D_r(\alpha)$ means the true within-household candidate looks better than the decoy for role r .

Assume the local block is role-aligned, so exact household recovery occurs if and only if every role prefers the true candidate. This gives a direct link between person-level ambiguity and household breakage.

Proposition 5.2 (Local household crossover probability). *Suppose $D_1(\alpha), \dots, D_m(\alpha)$ are independent with common distribution function F_α . Then*

$$\Pr_\alpha(\text{exact recovery of } g) = [1 - F_\alpha(0)]^m, \quad (17)$$

so

$$\Pr_{\alpha}(\text{household } g \text{ is broken}) = 1 - [1 - F_{\alpha}(0)]^m. \quad (18)$$

If additionally $D_r(\alpha) \sim N(\mu(\alpha), \sigma^2(\alpha))$, then

$$\Pr_{\alpha}(\text{household } g \text{ is broken}) = 1 - \Phi\left(\frac{\mu(\alpha)}{\sigma(\alpha)}\right)^m. \quad (19)$$

Hence household breakage rises with ambiguity whenever $\mu(\alpha)/\sigma(\alpha)$ falls with α , and it rises with household size m for any fixed nondegenerate ambiguity level.

Proof. Exact recovery requires $D_r(\alpha) > 0$ for all $r = 1, \dots, m$. Independence yields

$$\Pr_{\alpha}(D_1(\alpha) > 0, \dots, D_m(\alpha) > 0) = \prod_{r=1}^m \Pr_{\alpha}(D_r(\alpha) > 0) = [1 - F_{\alpha}(0)]^m.$$

The second claim follows by complementing, and the Gaussian formula follows by evaluating the one-dimensional probability with Φ . \square

This result is deliberately local. It is not a theorem about arbitrary global Hungarian assignments. Its role is narrower: it explains why exact household recovery can deteriorate much faster than person-level accuracy once nearby decoys become plausible.

6 Simulation design

The simulation is intentionally specific because the goal is to illustrate the general argument in one clean setting. Each replication generates 70 latent two-person households. Households are paired into ambiguity clusters so that nearby households share similar surname-like and age-like latent characteristics. Within each household a binary treatment Z_h is drawn with probability 1/2 and inherited by both members. Outcomes follow

$$Y_{hr} = 20 + 6Z_h + 0.5 \text{Age}_{hr} + \varepsilon_{hr},$$

with Gaussian noise.

Two observed files are generated from the same latent population. File A is relatively clean and carries the outcome. File B is noisier and carries the treatment label used in downstream estimation. Both files contain noisy surname, first-name, and age variables. An ambiguity parameter controls how similar neighboring households look in the observed linkage space.

All methods use the same person-level score matrix,

$$s_{ij} = -1.5 |\text{surname}_i - \text{surname}_j| - 1.2 |\text{fname}_i - \text{fname}_j| - 0.35 |\text{age}_i - \text{age}_j|.$$

We compare three procedures:

Greedy person-level. Select the highest-scoring remaining person pair at each step.

Hungarian person-level. Solve the additive assignment problem exactly.

Structured exact optimizer. Build the household compatibility matrix from (13), solve the household-level assignment (14), and recover the implied within-household matches.

The evaluation metrics follow the theory:

1. **Person accuracy:** the share of persons linked to their true counterpart.
2. **Exact household recovery:** the share of file-*A* households whose members are both linked to the correct file-*B* household.
3. **Absolute bias:** the absolute error of the linked estimator

$$\hat{\tau}(J) = \bar{Y}_{W(J)=1} - \bar{Y}_{W(J)=0},$$

where $W(J)$ is the treatment label imported from file *B* under join *J*.

Baseline results use ambiguity level 1.5 with 150 Monte Carlo replications. The ambiguity sweep uses levels 0.5, 1.0, 1.5, 2.0, and 2.5 with 60 replications each. The script `scripts/generate_results.py` reproduces every table and figure.

7 Results

Table 1 reports the baseline comparison. Standard errors across replications appear in parentheses.

Table 1: Baseline simulation results at ambiguity level 1.5.

Method	Person accuracy	Household exact match rate	Absolute bias
Greedy person-level	0.250 (0.003)	0.083 (0.003)	4.412 (0.058)
Hungarian person-level	0.291 (0.004)	0.110 (0.004)	4.139 (0.054)
Set-aware household-first	0.542 (0.005)	0.576 (0.005)	2.595 (0.067)

The first comparison is algorithmic: Hungarian improves on greedy linkage because it solves the additive person-level objective exactly. The second comparison is substantive: relative to Hungarian, the structured optimizer improves person-level accuracy by 0.251, improves exact household recovery by 0.466, and reduces absolute bias by 1.544 points.

Those gains line up with the theory. Proposition 5.1 says that once households define the relevant feasible set, the right exact benchmark is a household-level assignment, not a person-level one. Proposition 5.2 says that household recovery compounds person-level ambiguity multiplicatively. Corollary 3.2 says that, in the binary symmetric setup used here, reducing induced treatment-label misclassification lowers downstream bias. The simulation turns those statements into magnitudes.

Figure 1 plots exact household recovery across the ambiguity sweep. Figure 2 plots absolute downstream bias.

At low ambiguity, the pairwise surrogate is often close to the structured problem, so the gap is moderate. As ambiguity rises, cross-household temptations become more common and exact household recovery deteriorates much faster for the person-level procedures. The structured optimizer also maintains a clear advantage in downstream bias throughout the sweep.

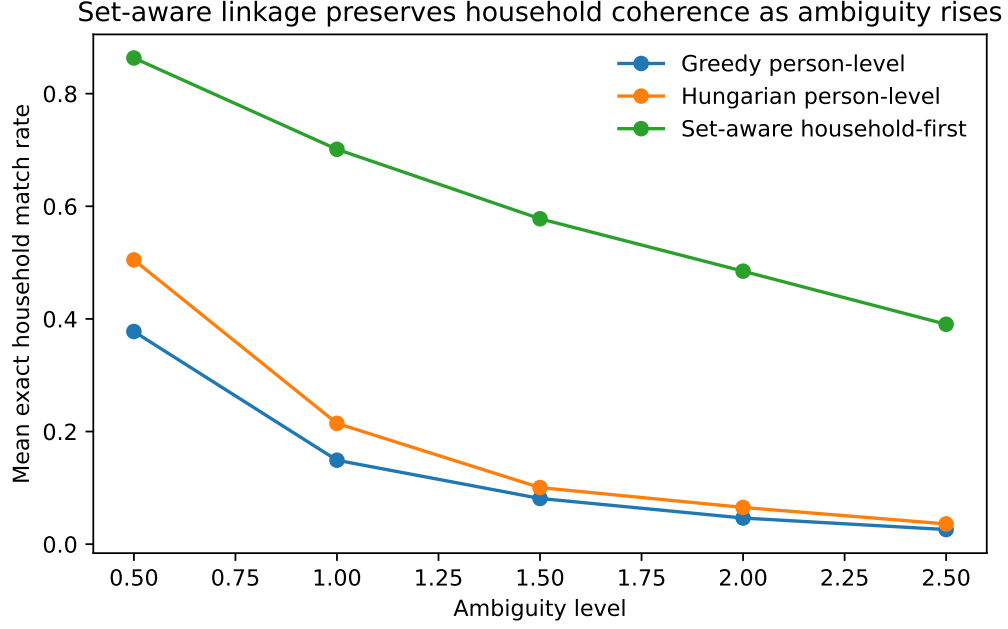


Figure 1: Exact household recovery as a function of ambiguity. All methods degrade as ambiguity rises, but the structured optimizer retains a large advantage throughout.

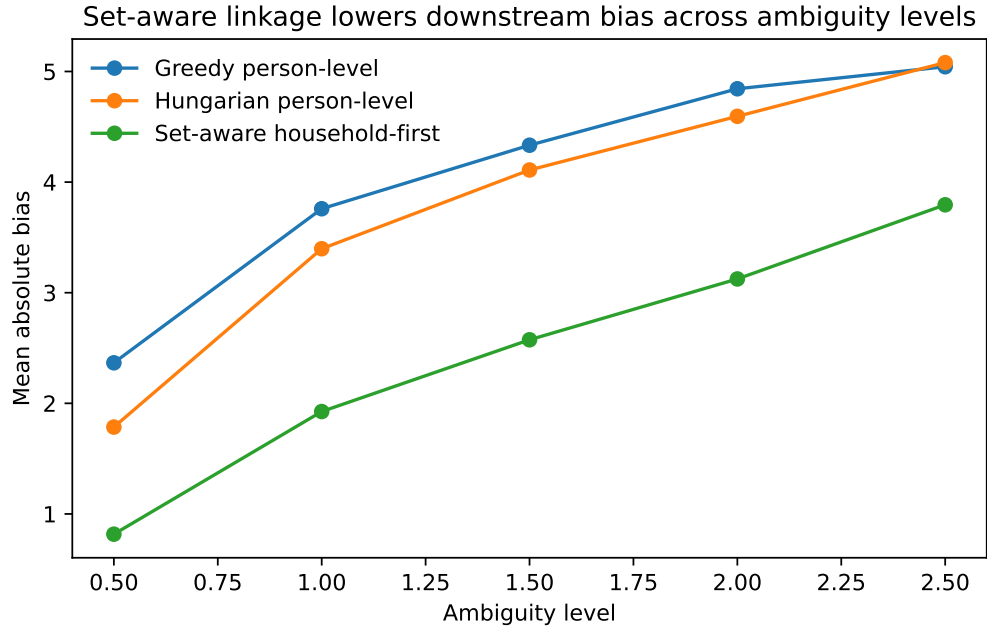


Figure 2: Absolute downstream bias across the ambiguity sweep. The structured optimizer dominates throughout, with the advantage widening as the join becomes harder.

These differences are not generated by richer pairwise features or by tuning. All three procedures use the same underlying record-level score matrix. What changes is whether the optimizer is permitted to spend pairwise score by damaging a property of the linked table that the downstream analysis actually uses.

8 Discussion

The paper’s main claim is general: many joins are better understood as selection problems over feasible linked tables than as additive optimization over candidate pairs. Households are one clean example because the gap between person-level optimization and household-level analysis is easy to see, but the same logic applies to table integration problems throughout statistics and computer science.

Three practical implications follow.

First, linkage should be designed against the downstream estimand, not against pairwise accuracy alone. Pairwise precision and recall can be high while the analysis labels in the linked table are badly distorted. Proposition 3.1 and Corollary 3.2 make that distinction explicit.

Second, higher-order constraints are information, not inconvenience. If the output table should preserve a hierarchy, satisfy calibration totals, or remain consistent across composed joins, that information can and should be used by the optimizer. Ignoring it is analogous to ignoring covariates in experimental design or auxiliary variables in survey sampling.

Third, the computer-science lesson is broader than linkage alone. The paper is really about decomposable versus non-decomposable objectives in combinatorial optimization for data integration. When the true objective has block structure, graph structure, calibration constraints, or compositional consistency, optimizing a flat additive relaxation and hoping the structure emerges is a gamble. Hierarchical assignment, constrained optimization, and FSM-style sequential selection are three ways to respect the object that actually matters.

That said, structure is not automatically beneficial. It helps when the added information is both *known* and *aligned* with the analyst’s loss function. If the structure is irrelevant to the estimand, weakly informative, or misspecified, enforcing it can add variance or even increase bias. The lesson is not that all linkage should be household-aware, calibration-aware, or temporally constrained. The lesson is that the relevant table-level criterion should be stated explicitly and then optimized on purpose.

9 Reproducibility

The project folder contains a single script, `scripts/generate_results.py`, that reproduces every table and figure included in this paper. Running the script writes CSV summaries to `results/`, L^AT_EX table fragments to `tables/`, and PDF figures to `figures/`. The `Makefile` compiles the paper. The bibliography is stored in `refs.bib`, with citation checks documented in `refs_verified.md`.

References

- Bansal, Nikhil, Avrim Blum, and Shuchi Chawla (2004). “Correlation Clustering”. In: *Machine Learning* 56.1–3, pp. 89–113. DOI: [10.1023/B:MACH.0000033116.57574.95](https://doi.org/10.1023/B:MACH.0000033116.57574.95).
- Chattopadhyay, Ambarish, Carl N. Morris, and Jose R. Zubizarreta (2026). “Balanced and Robust Randomized Treatment Assignments: The Finite Selection Model for the Health Insurance Experiment and Beyond”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society*. Advance article. DOI: [10.1093/jrsssa/qnag040](https://doi.org/10.1093/jrsssa/qnag040).
- Christen, Peter (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer. DOI: [10.1007/978-3-642-31164-2](https://doi.org/10.1007/978-3-642-31164-2).
- Deville, Jean-Claude and Carl-Erik S”arndal (1992). “Calibration Estimators in Survey Sampling”. In: *Journal of the American Statistical Association* 87.418, pp. 376–382. DOI: [10.1080/01621459.1992.10475217](https://doi.org/10.1080/01621459.1992.10475217).
- Deville, Jean-Claude and Yves Tillé (2004). “Efficient balanced sampling: The cube method”. In: *Biometrika* 91.4, pp. 893–912. DOI: [10.1093/biomet/91.4.893](https://doi.org/10.1093/biomet/91.4.893).
- Fellegi, Ivan P. and Alan B. Sunter (1969). “A Theory for Record Linkage”. In: *Journal of the American Statistical Association* 64.328, pp. 1183–1210. DOI: [10.1080/01621459.1969.10501049](https://doi.org/10.1080/01621459.1969.10501049).
- Hainmueller, Jens (2012). “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies”. In: *Political Analysis* 20.1, pp. 25–46. DOI: [10.1093/pan/mpr025](https://doi.org/10.1093/pan/mpr025).
- Kuhn, Harold W. (1955). “The Hungarian method for the assignment problem”. In: *Naval Research Logistics Quarterly* 2.1–2, pp. 83–97. DOI: [10.1002/nav.3800020109](https://doi.org/10.1002/nav.3800020109).
- Morris, Carl (1979). “A finite selection model for experimental design of the health insurance study”. In: *Journal of Econometrics* 11.1, pp. 43–61. DOI: [10.1016/0304-4076\(79\)90053-8](https://doi.org/10.1016/0304-4076(79)90053-8).
- Newhouse, Joseph P. and Insurance Experiment Group (1993). *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.
- Zubizarreta, Jose R. (2012). “Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery”. In: *Journal of the American Statistical Association* 107.500, pp. 1360–1371. DOI: [10.1080/01621459.2012.703874](https://doi.org/10.1080/01621459.2012.703874).