Auditing Linkage Flows: Design-Based Inference for Additive Many-to-Many Record Linkage

Extended Abstract

Motivation

Computational social science increasingly uses retrieval-and-linkage pipelines to produce additive measurements rather than binary matches. These pipelines, which range from string matching and embedding nearest-neighbor search to LLM-assisted entity resolution, quantify how much content flows between units: legislative text into enacted law, judicial language borrowed from earlier opinions, donor activity attached to an organization, or content a social-media actor reuses from upstream sources (Eatough and Preece, 2025; Ornstein, 2026). The substantive object is a source-level flow, a row sum over latent pair contributions, not a pair label. This shift has an uncomfortable consequence for validation: a pipeline can score every retained pair accurately and still understate the total flow for actors whose influence is diffuse, because pair-level precision and recall are silent on contributions the retriever never returned. The bias is not statistically neutral. If retrieval errors concentrate on short or institutionally indirect contributions, linkage systems reshape who appears productive, influential, or represented, reproducing the visibility bias the measurement project was meant to correct, and echoing parallel concerns in LLM-assisted social measurement, where unaudited predictions yield biased downstream estimates even at high accuracy (Egami et al., 2024; Gligorić et al., 2025). We develop a design-based audit framework for additive many-to-many linkage: a weighted linkage-flow estimand, an error decomposition that separates retained-pair scoring from two flavors of omission, and a model-assisted residual-correction estimator whose validity rests on the audit design rather than the linkage model.

Estimand and error decomposition

Let $i = 1, \ldots, N$ index source units, $C^*_{ij} \in \mathbb{R}$ the true contribution of target j to source i, and $\mathcal{U}_i$ the substantive universe of admissible targets. The weighted linkage-flow estimand is

$$\theta(w) = \tfrac{1}{N} \sum_{i=1}^{N} \sum_{j \in \mathcal{U}_i} w_{ij} C^*_{ij}, \tag{1}$$

with w_{ij} a known analysis weight. Indicator weights yield group-specific flows; estimating numerator and denominator separately yields ratios such as the share of enacted text attributed to a group. Most applied pipelines contain three nested sets: $\mathcal{U}_i$, an auditable retrieval frame $\mathcal{F}_i \subseteq \mathcal{U}_i$ (a broad block reviewable with known probabilities), and the scored set $\mathcal{S}_i \subseteq \mathcal{F}_i$ the production pipeline actually returns. For a machine row total $Q_i(w) = \sum_{j \in \mathcal{S}_i} w_{ij} q_{ij}$ built from predicted contributions q_{ij}, error against the universe target decomposes additively:

$$Q_i(w) - Y_i^U(w) = \underbrace{\sum_{j \in \mathcal{S}_i} w_{ij}(q_{ij} - C^*_{ij})}_{\text{retained-pair scoring}} - \underbrace{\sum_{j \in \mathcal{F}_i \setminus \mathcal{S}_i} w_{ij} C^*_{ij}}_{\text{within-frame omission}} - \underbrace{\sum_{j \in \mathcal{U}_i \setminus \mathcal{F}_i} w_{ij} C^*_{ij}}_{\text{out-of-frame omission}} . \tag{2}$$

A retained-pair audit addresses only the first term. It is silent on mass sitting in the auditable block but never scored (the failure mode of top-k truncation and approximate-nearest-neighbor retrieval), and silent on mass outside the frame. Retained-pair validation therefore identifies $Y_i^S(w) = \sum_{j \in \mathcal{S}_i} w_{ij} C^*_{ij}$, not $Y_i^F(w)$ or $Y_i^U(w)$, unless the omitted terms vanish.

Audit design and estimator

The design has two layers. **Outer:** sample source rows with known inclusion probabilities $\pi_i = \Pr_d(R_i = 1) > 0$; the outer unit must be the row because the residual being corrected is a row residual. **Inner:** within each audited row, run a probability sample over the three pair strata: retained pairs in $\mathcal{S}_i$ (often audited exhaustively), within-frame omissions in $\mathcal{F}_i \setminus \mathcal{S}_i$, and, when the universe target

matters, discovery pairs in $\mathcal{U}_i \setminus \mathcal{F}_i$. Inner pair inclusion probabilities ρ_{ij} must be known but need not be equal; active allocation can oversample near-threshold or high-uncertainty pairs. Combining audited strata gives an unbiased row substitute $\widetilde{Y}_i^T(w)$ for the chosen target $T \in \{S, F, U\}$. The model-assisted estimator is

$$\widehat{\theta}_T(w) = \frac{1}{N}\sum_{i=1}^{N} Q_i(w) + \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i}\{\widetilde{Y}_i^T(w) - Q_i(w)\}. \tag{3}$$

Whenever the outer probabilities are known and the audit procedure is credible, $\mathbb{E}_d\{\widehat{\theta}_T(w)\} = N^{-1}\sum_i Y_i^T(w)$. Correctness of the linkage model, calibration of q_{ij}, and full recall of the retriever are not required: validity comes from the audit, efficiency from the surrogate. The construction inherits the logic of design-based supervised learning and prediction-powered inference, applied to a row total from a retrieval pipeline rather than a single predicted label (Angelopoulos et al., 2023). Standard errors are computed by a plug-in design variance estimator that combines the outer row-sampling variance with the inner stratified pair-sampling variance. This ensures that confidence intervals reflect uncertainty from omitted strata.

Simulation

Monte Carlo experiments cross retrieval quality (strong/weak omission) with calibration quality (mild/severe retained-pair bias) at audit budgets $B \in \{1.2, 2.4, 4.8, 9.6\}\times 10^3$ and $N = 1{,}200$ rows per replication. Four estimators are compared against θ_U: machine-only, retained-pair audit, frame audit, and discovery audit. The pattern is sharp. Machine-only and retained-pair audits remain biased at every budget; retained-pair bias at the largest budget ranges from -0.44 (strong retrieval) to -1.17 (weak retrieval), with empirical coverage of nominal 95% intervals at zero. Frame and discovery audits are essentially unbiased for their targets, with RMSE between 0.024 and 0.069 and coverage between 0.94 and 0.96.

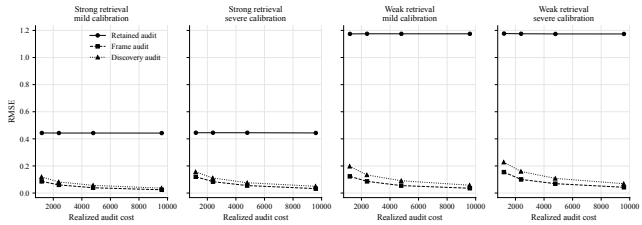

Figure 1: Cost frontier.

Empirical applications

The primary application is LawProM-style legislative text reuse (Eatough and Preece, 2025): rows are member-section units, C_{ij}^* is copied text mass into target section j, w_{ij} selects party, chamber, or demographic subgroup, $\mathcal{F}_i$ is a broad temporal-and-topic block, and $\mathcal{S}_i$ is whatever the production system retains. A secondary application is one-to-many organization-to-donor linkage (Ornstein, 2026), where ratio targets such as ideology-weighted donation share stress the same audit-identification logic; the same logic governs LLM-assisted entity resolution over online actors. Limitations are intrinsic: validity rests on the audit rather than the model, discovery sampling is costly when true links are rare, and the additive structure does not address non-additive causal questions.

Angelopoulos, Anastasios N. et al. (2023). "Prediction-Powered Inference". In: *Science* 382.6671, pp. 669–674.

Eatough, Mandi and Jessica R. Preece (2025). "Crediting Invisible Work: Congress and the Lawmaking Productivity Metric (LawProM)". In: *American Political Science Review* 119.2, pp. 566–584.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei (2024). *Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses*. Working paper, version dated November 17, 2024.

Gligorić, Kristina et al. (2025). "Can Unconfident LLM Annotations Be Used for Confident Conclusions?" In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 3514–3533.

Ornstein, Joseph T. (2026). "Probabilistic Record Linkage Using Pretrained Text Embeddings". In: *Political Analysis* 34.2, pp. 205–216.