MORE IS BETTER? EXPLORING LATENT CORRESPONDENCE IN FEDERATED HYPERGRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated hypergraph learning with latent cross-client hyperedge correspondence poses fundamental identifiability and computational challenges, as the same local release can correspond to multiple global hypergraph structures. We characterize the tractability of recovering these latent correspondences, showing tractable pairwise assignment but generally intractable multi-client hyperedge matching. Building on this characterization, we derive a cross-entropy decomposition that reveals when local prediction can outperform predictors fitted with additional cross-client information. Motivated by this principle, we introduce **Fed**erated **A**daptive **S**pectral **T**raining with **R**epresentation **A**lignment (**FedASTRA**), which learns correspondence-invariant diffusion representations through adaptive heat diffusion and masked latent prediction while communicating model parameters only. We establish its asymptotic optimality with respect to the permitted local information and prove global population-risk convergence under explicit conditions on recovery, representation, and update alignment. Experiments on eight real-world datasets demonstrate the effectiveness of FedASTRA against identifier-free federated graph and hypergraph methods. On ogbn-papers100M, FedASTRA reaches the performance-communication Pareto frontier, achieving 50.07 Accuracy and 29.73 Macro-F1 while communicating only 59.60 MiB per round.

1 INTRODUCTION

Federated learning enables collaborative model training without sharing data records (McMahan et al., 2017). For applications involving group-wise interactions, hypergraphs provide a natural representation of memberships in care teams, coauthor sets, and biochemical interactions (Feng et al., 2019; Chien et al., 2022; Wang et al., 2023). However, when hypergraph data are anonymized across clients, local memberships may remain observable while their cross-client correspondence is lost, making the underlying global hypergraph potentially non-identifiable from local observations alone. Consider two hospitals that share care episodes, as illustrated in Figure 1. After episode identifiers are removed, Hospital A retains anonymous records of backache and BP/ECG, while Hospital B retains records of surgery and headache/drug. Same-order matching yields compatible episodes of sizes two and four, whereas cross-order matching yields two episodes of size three. Thus, even this minimal two-client example admits distinct global compatible structures consistent with the same local observations, raising a natural question: *which global hypergraph properties remain identifiable under latent hyperedge correspondence?*

We answer this question by establishing sharp identification bounds and computational boundaries under latent hyperedge correspondence. For two clients, the compatible global structures reduce to bipartite assignments, yielding sharp identification bounds and a double-centering criterion for point identification. Beyond two clients, correspondence becomes a multiway hypergraph-matching problem. It is NP-complete for every fixed client count above two, yet fixed-parameter tractable jointly in client count and matching size. When the client count is unbounded, it becomes W[1]-complete when parameterized only by the matching size.

This characterization raises another question: *when does additional cross-client information improve prediction, and when can exploiting it be detrimental?* Additional information may provide useful label signal, but uncertain correspondence can also increase prediction variability and fitting error. We decompose cross-entropy risk into information gain and fitting error, yielding a sufficient condi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

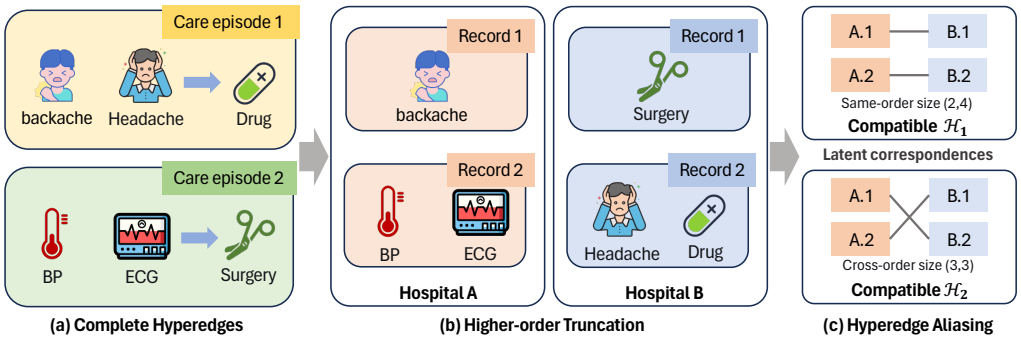


Figure 1: **From complete hyperedges to latent hyperedge correspondences.** (a) A complete care episode is represented as a hyperedge (blood pressure (BP); electrocardiogram (ECG)). (b) Complete episode may span multiple hospitals, for example, BP measurement and ECG examination may be performed at hospital A, while the subsequent surgery is carried out at hosptial B. Each hospital observes only a local fragment of the original hyperedge, leading to **higher-order truncation**. (c) Once cross-hospital record correspondences are latent, the same local releases can admit multiple compatible global matchings, resulting in **hyperedge aliasing**.

tion under which a local predictor outperforms a fitted predictor with additional information under the same evaluation law.

Motivated by this tradeoff, together with the communication and privacy costs of richer cross-client exchanges (Shi et al., 2025), we introduce FedASTRA, which learns correspondence-invariant local diffusion representations through adaptive heat filtering and masked latent prediction (LeCun, 2022; Assran et al., 2023), with a client-local exponential moving-average (EMA) target and parameter-only FedAvg (McMahan et al., 2017). We show that effectively exploiting the task-relevant local information can be sufficient without resolving latent cross-client correspondences or completing the global hypergraph, yielding asymptotic optimality relative to the permitted local information. Under explicit representation and update-alignment conditions, we further establish global population-risk convergence along the implemented federated updates.

Our contributions are fourfold. **First,** we establish sharp identification bounds and a computational hierarchy for latent hyperedge correspondence: polynomial-time two-client assignment, NP-complete yet fixed-parameter tractable matching for any fixed client count above two, and W[1]-complete matching when the client count is unbounded. **Second,** we characterize when additional cross-client information helps or hurts prediction through a cross-entropy decomposition that connects information gain, fitting error, prediction variation, and assignment geometry. **Third,** we develop FedASTRA and establish asymptotic optimality relative to the permitted local information, together with global population-risk convergence under explicit representation and update-alignment conditions. **Fourth,** experiments on eight real-world datasets demonstrate consistent improvements over identifier-free federated graph and hypergraph methods; on ogbn-papers100M, FedASTRA reaches the performance-communication Pareto frontier.

2 Related Work

Federated graph learning. Existing federated graph learning methods differ in whether cross-client graph correspondence is available. Identifier-free methods leave cross-client node identities latent (Zhang et al., 2021; Baek et al., 2023; Li et al., 2023; Fu et al., 2025), as in FedSage (Zhang et al., 2021) operating on local subgraphs, while FedSage+ (Zhang et al., 2021) generating cross-client neighbor features. Methods with explicit correspondence, such as FedGCN (Yao et al., 2023) and FedCog (Lei et al., 2023), exploit shared cross-client node or edge identities to incorporate additional graph information. For hypergraph, FedHGL (Luo et al., 2025) instead assumes shared hyperedge identifiers and utilze local differential privacy (LDP) (Kasiviswanathan et al., 2011) to support structure completion. Despite progress, these methods trade cross-client communication and privacy exposure for predictive performance. In contrast, we study an identifier-free setting in which cross-client hyperedge correspondence is latent and only local incidence is available.

Hypergraph representations learning. Most hypergraph neural networks assume observed node-hyperedge memberships and focus on how to encode higher-order structure, including spectral propagation on hypergraph operators (Feng et al., 2019; Wang et al., 2023), graph-reduction approaches that approximate hyperedges with pairwise structures (Yadati et al., 2019), and direct higher-order aggregation through message passing or permutation-invariant set functions (Dong et al., 2020; Chien et al., 2022; Huang & Yang, 2021). Beyond supervised learning, self-supervised methods learn transferable hypergraph representations through multi-level contrastive learning (Lee & Shin, 2023) and generative pretext tasks (Kim et al., 2024). More recently, large-scale pretraining and foundation models Gao et al. (2026) have emerged for cross-task and cross-domain transfer. FedASTRA targets a more realistic federated scenario by learning effective representations from anonymous client-local incidence when cross-client hyperedge correspondence remains latent.

3 Learning Difficulty under Latent Hyperedge Correspondence

3.1 Observation model

Let $\mathcal{H} = (V, \mathcal{E}, X, r, w)$ be a finite node-attributed, typed, weighted multihypergraph. Here V is the node set, $\mathcal{E}$ is the multiset of hyperedge instances, X is the node-feature matrix, and r and w collect hyperedge types and weights. hyperedge instance $e \in \mathcal{E}$ has a support node set $A_e \subseteq V$, a type r_e in the finite type set $\mathcal{R}$, and a positive weight w_e. Multiple instances may share the same support and type, and a type label carries no instance identity. For an integer $1 \le q \le |V|$, let $\binom{V}{q}$ be the collection of q-node subsets. Aggregating all hyperedge instances with the same support A and type r, we define

$$w_r(A) = \sum_{e: A_e = A,\, r_e = r} w_e, \tag{1}$$

$$M^{\mathcal{H}}_{q,r}(S) = \sum_{A \supseteq S} w_r(A), \qquad r \in \mathcal{R}, \quad S \in \binom{V}{q}. \tag{2}$$

Thus $M^{\mathcal{H}}_{q,r}(S)$ is the weighted inclusion count of the q-node subset S among hyperedges of type r. The full collection $\{M^{\mathcal{H}}_{q,r}(S)\}$ is the invertible Boolean-zeta transform of the aggregate weights $\{w_r(A)\}$. In particular, the $q = 2$ and $q = 3$ coordinates capture observable pairwise and triple co-membership patterns before client partitioning. Find the notation collection in Appendix B.1.

A fixed node-to-client assignment $z : V \to [K]$, with $k \in [K] := \{1, \dots, K\}$, partitions V into local node set $V_k = \{v \in V : z(v) = k\}$ for each client k and induces local hyperedge fragments $F_{e,k} = A_e \cap V_k$. At a visibility threshold $c \ge 1$, client k retains each fragment $F_{e,k}$ with $|F_{e,k}| \ge c$, together with its type r_e, weight w_e, and a status indicator $\upsilon_{e,k} \in \{0, 1\}$ specifying whether hyperedge e is fully contained in client k. Crucially, cross-client hyperedge identifiers are withheld. Let Γ collect the public assumptions governing the observation process, including the visibility rule and any exact constraints on the number of retained fragments. Write $[T_k]_{\text{row}}$ for the equivalence class of client k's retained fragment table under arbitrary row permutations. The analyst release is then defined by the deterministic observation map:

$$Y_{\text{A}} = O_{z,\Gamma}(\mathcal{H}) = \left(\left((V_k, X_{V_k}), \left[(F_{e,k}, r_e, w_e, \upsilon_{e,k}) : |F_{e,k}| \ge c \right]_{\text{row}} \right)_{k=1}^{K}, \Gamma \right). \tag{3}$$

Uppercase Y_{A} denotes the analyst release as a random element and lowercase y_{A} a realized release. Within the latent hypergraph class $\mathcal{M}_\Gamma$ permitted by the public assumptions Γ, define the analyst compatibility class

$$\mathcal{F}_{\text{A}}(y_{\text{A}}, \Gamma) = \{\mathcal{H}' \in \mathcal{M}_\Gamma : O_{z,\Gamma}(\mathcal{H}') = y_{\text{A}}\}. \tag{4}$$

In contrast to the richer analyst release, the learner, representing the actual training algorithm, receives a coarsened release. Accordingly, let B_k be the binary node-fragment incidence matrix induced by T_k, with one column for each row of T_k. Define $[B_k]_{\text{col}} = \{B_k \Pi : \Pi \text{ is a column permutation}\}$ and thus arbitrary row permutations of T_k correspond to column permutations $B_k \mapsto B_k \Pi$. The learner coarsened release is therefore

$$Y_{\text{L}} = \rho(Y_{\text{A}}) = \left((X_k, [B_k]_{\text{col}}) \right)_{k=1}^{K}, \qquad X_k := X_{V_k}. \tag{5}$$

It retains only local node features and anonymous incidence structure, omitting hyperedge types, weights, status indicators, and row identities. For a fixed training protocol $\mathcal{P}$ containing the released labels, task split, participation schedule, and hyperparameters, and algorithmic randomness ξ, any learner operating under this observation model is a measurable function of $(Y_{\mathrm{L}}, \mathcal{P}, \xi)$. In native comparisons, the same realized release is provided to all learners.

For each q-node subset S, the client partition determines whether its co-membership information is retained. For $q \geq c$, we define

$$\widetilde{M}_{q,r}(S, z) = M^{\mathcal{H}}_{q,r}(S)\mathbf{1}\{z \text{ is constant on } S\}, \tag{6}$$

so the contribution of S is retained exactly when all nodes in S are assigned to the same client. We quantify the erased fraction by

$$\tau_q(\mathcal{H}, z) = 1 - \frac{\sum_{r\in\mathcal{R}} \|\widetilde{M}_{q,r}(\cdot, z)\|_1}{\sum_{r\in\mathcal{R}} \|M^{\mathcal{H}}_{q,r}\|_1}, \tag{7}$$

with $\tau_q(\mathcal{H}, z) = 0$ for a zero denominator. We call the erased mass *higher-order truncation*, as illustraged in Figure 1. Independently permuted client fragment tables may also admit multiple compatible cross-client matchings. We call this ambiguity *hyperedge aliasing*. These two effects can alter global higher-order statistics. We therefore consider a class of linear co-membership targets $\Phi_{q,\psi}$. For any finite type-indexed test function $\psi_{q,r}$, define

$$\Phi_{q,\psi}(\mathcal{H}) = \sum_{e\in\mathcal{E}} w_e \sum_{S\in\binom{A_e}{q}} \psi_{q,r_e}(S, X_S). \tag{8}$$

Here, $\psi_{q,r}(S, X_S)$ specifies the contribution of each q-node subset S, allowing $\Phi_{q,\psi}$ to cover both total order-q co-membership mass and more general feature-dependent higher-order statistics.

Equation (7) quantifies the co-membership mass loss under a fixed partition. Appendices B.2, C.5–C.6 give the Boolean-zeta inversion and show how the co-membership quantities can be estimated under a randomized partition using Horvitz-Thompson estimation (Horvitz & Thompson, 1952).

3.2 Exact two-client compatibility sets

Two clients form the smallest nontrivial setting with an exact correspondence analysis. We use the richer analyst release with observed status indicators $\upsilon_{e,k}$, and exact type r and weight ω, together with $c = 1$ and equal numbers of retained fragment rows across the two clients. Thus, each cross-client hyperedge leaves one nonempty row per client. Let m_h denote the common row count in stratum $h = (r, \omega)$. Within stratum h, fragments $F^h_{1i} \subseteq V_1$ and $F^h_{2j} \subseteq V_2$, indexed by $i, j \in [m_h]$, are matched by an unobserved $\varpi_h \in \mathrm{Sym}(m_h)$, the symmetric group on $[m_h]$. Let Φ_{obs} collect the client-local and within-fragment contributions that are invariant across all compatible matchings. For a candidate cross-client pair (F^h_{1i}, F^h_{2j}), the mixed matching-dependent contribution is

$$C^h_{ij} = \omega \sum_{\substack{S\in\binom{F^h_{1i}\cup F^h_{2j}}{q}: \\ S\cap F^h_{1i}\neq\emptyset,\ S\cap F^h_{2j}\neq\emptyset}} \psi_{q,r}(S, X_S). \tag{9}$$

The cost matrix C^h exposes exactly which part of the target $\Phi_{q,\psi}$ depends on the missing correspondence. Define the identified set of $\Phi_{q,\psi}$ under the analyst release as its attainable values over all compatible latent hypergraphs: $\mathcal{I}_{q,\psi}(y_{\mathrm{A}}) = \{\Phi_{q,\psi}(\mathcal{H}') : \mathcal{H}' \in \mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}}, \Gamma)\}$.

Theorem 1 (Compatible target values under an unobserved permutation)**.** *Assume $K = 2$, $c = 1$, observed status indicators, exactly one nonempty cross fragment at each client for every cross-client hyperedge, exact type-and-weight strata, and equal row counts within every stratum. Then the compatible values and their sharp endpoints are*

$$\mathcal{I}_{q,\psi}(y_{\mathrm{A}}) = \left\{ \Phi_{\mathrm{obs}} + \sum_h \sum_{i=1}^{m_h} C^h_{i,\varpi_h(i)} : \varpi_h \in \mathrm{Sym}(m_h) \right\}, \tag{10}$$

$$\Phi^-_{q,\psi} = \Phi_{\mathrm{obs}} + \sum_h \min_{\Lambda\in\mathbb{B}_{m_h}} \langle C^h, \Lambda\rangle_F, \quad \Phi^+_{q,\psi} = \Phi_{\mathrm{obs}} + \sum_h \max_{\Lambda\in\mathbb{B}_{m_h}} \langle C^h, \Lambda\rangle_F, \tag{11}$$

where $\mathbb{B}_m = \{\Lambda \in \mathbb{R}_+^{m\times m} : \Lambda\mathbf{1} = \mathbf{1}, \Lambda^\top\mathbf{1} = \mathbf{1}\}$ *is the Birkhoff polytope and* $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ *is the Frobenius inner product. Each condition below characterizes point identification within a stratum, independently of the probability distribution over matchings compatible with the observed release:*

$$J_{m_h} C^h J_{m_h} = 0, \tag{12}$$

$$C^h_{ij} = \alpha^h_i + \beta^h_j, \tag{13}$$

$$C^h_{ij} + C^h_{i'j'} - C^h_{ij'} - C^h_{i'j} = 0 \quad \textit{for all } i, i', j, j' \in [m_h], \tag{14}$$

with centering matrix $J_m = I_m - \mathbf{1}\mathbf{1}^\top/m$ *and vectors* $\alpha^h, \beta^h \in \mathbb{R}^{m_h}$ *in Equation* (13).

Each compatible assignment selects one entry from every row and column of C^h, thereby forming a permutation matrix. By Birkhoff-von Neumann integrality Birkhoff (1946), both endpoints are attainable. Double centering in Equation (12) removes additive row and column effects, leaving only the interaction that can vary with the latent correspondence, in a subspace of dimension at most $(m_h - 1)^2$. Equation (14) checks whether any such interaction remains. Appendix C.1 gives the proof and dual certificates. Appendix B.3 also further discusses weaker observation assumptions.

Coarsening the analyst release enlarges the compatible class. Hence, if ϕ_- and ϕ_+ are attainable endpoints under the richer release, any randomized scalar estimator using the coarser learner release has worst-case absolute risk at least $(\phi_+ - \phi_-)/2$ (Appendix C.7). This diagnostic uses two clients on the same native sources, while training uses $K \in \{3, 6, 9\}$ and path invariance holds for any K. Prediction can still be accurate when the local label signal survives unresolved global structure.

Graph matching versus hypergraph matching. Total order-q mass admits sharp sorting bounds in $O(\sum_h m_h \log m_h)$ time (Appendices C.2–C.3). To isolate relation size, consider K clients with m distinct singleton rows each and explicitly listed rewarded tuples $\mathcal{E}_+ \subseteq [m]^K$. Each listed tuple has reward one, all others zero. A completion with at least b rewarded relations exists exactly when these candidate K-node hyperedges contain a matching of size b. For $K = 2$, this is polynomial bipartite graph matching. For fixed $K \geq 3$, it is complete for nondeterministic polynomial time (NP-complete) but fixed-parameter tractable (FPT) in Kb, meaning time $f(Kb)\,\mathrm{poly}(L_{\mathrm{in}})$ for input length L_{in}. With unbounded K, it is complete for W[1], the first level of the W hierarchy, in b (Appendix C.4). This separates pairwise from multiway correspondence under the same observation and reward rules, independently of the optimization difficulty of a neural classifier.

4 Benefit Bounds of Additional Information

Given these fundamental identifiability and computational challenges under hyperedge correspondence, a natural thought is to exploit more information beyond local observations for better prediction. Yet a subtle tradeoff arises here: additional information can reveal label signal, while correspondence uncertainty can induce prediction variation and fitting error. In other words, its benefit is bounded. In this section, we characterize these benefit bounds through cross-entropy decomposition.

Information value and fitting cost. For class $y \in [C] = \{1, \ldots, C\}$, $C \geq 2$, predictor F uses $\mathcal{S}_0$: client attributes, anonymous incidence, and the parameter-only transcript through validation, excluding held-out labels. Predictor G additionally uses A_{aux} and its derived training states, giving $\mathcal{S}_1 = \mathcal{S}_0 \vee \sigma(A_{\mathrm{aux}})$. Let $p_i = \Pr(y = \cdot \mid \mathcal{S}_i)$, $i = 0, 1$, and q_F, q_G be their predictions. Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(p\|q) = \sum_c p_c \log(p_c/q_c)$ yields the information and fitting terms

$$\mathcal{I}_{\mathrm{add}} = \mathbb{E}D_{\mathrm{KL}}(p_1\|p_0), \qquad \mathcal{E}_F = \mathbb{E}D_{\mathrm{KL}}(p_0\|q_F), \qquad \mathcal{E}_G = \mathbb{E}D_{\mathrm{KL}}(p_1\|q_G), \tag{15}$$

Then $\mathcal{I}_{\mathrm{add}}$ can measure new label information and $\mathcal{E}_F, \mathcal{E}_G$ measure fitting error to the respective optima. For measurable predictions with finite log risk $\mathcal{R}_{\log}(q) = \mathbb{E}[-\log q_y]$, conditional entropy yields the exact risk decomposition below, whose derivation is given in Appendix C.10:

$$\mathcal{R}_{\log}(q_G) - \mathcal{R}_{\log}(q_F) = \underbrace{\mathcal{E}_G - \mathcal{E}_F}_{\text{difference in fitting error}} - \underbrace{\mathcal{I}_{\mathrm{add}}}_{\text{additional label information}}. \tag{16}$$

Thus F has strictly lower risk exactly when $\mathcal{E}_G - \mathcal{E}_F > \mathcal{I}_{\mathrm{add}}$, with the reverse strict inequality favoring G. Additional information weakly improves ideal risk because it can be ignored (Blackwell, 1953). For fitted models, however, its gain competes with approximation, estimation, and optimization errors. When $p_1 = p_0$, fitting quality alone determines which predictor has lower risk.

How correspondence creates a prediction cost. To quantify this cost, remove the class mean from enriched logits $s_G \in \mathbb{R}^C$ and set $\bar{s}_G = \mathbb{E}[s_G \mid \mathcal{S}_0]$. For cross-entropy risk $\mathcal{R}_{\mathrm{CE}}(s) = \mathcal{R}_{\log}(\mathrm{softmax}(s))$ and finite-risk local logits s_F, define $V_{\mathrm{aux}} = \mathbb{E}\|s_G - \bar{s}_G\|_2^2$, $\Delta_{\mathrm{label}} = \mathbb{E}\|p_1 - p_0\|_2^2$, and $\delta_F = \mathcal{R}_{\mathrm{CE}}(s_F) - \mathcal{R}_{\mathrm{CE}}(\bar{s}_G)$. Under $\|s_G\|_\infty \le B_{\mathrm{logit}} < \infty$ almost surely, Theorem 6 gives

$$\boxed{\frac{e^{-2B_{\mathrm{logit}}}}{2C} V_{\mathrm{aux}} > \sqrt{\Delta_{\mathrm{label}} V_{\mathrm{aux}}} + \delta_F \quad \Longrightarrow \quad \mathcal{R}_{\mathrm{CE}}(s_G) > \mathcal{R}_{\mathrm{CE}}(s_F).} \tag{17}$$

Hence local prediction has lower risk when variation cost exceeds the label-alignment bound plus the signed fitting gap. The criterion is sufficient, since Cauchy–Schwarz bounds label alignment. When $\Delta_{\mathrm{label}} = 0$ and $\delta_F \le 0$, any $V_{\mathrm{aux}} > 0$ certifies a strict local advantage. Moreover, Proposition 5 computes this variation from the double-centered matching costs of Theorem 1 for affine logits under independent uniform matchings, linking the identification result to this prediction-risk comparison.

Communication and privacy boundary. Additional exchanges incur communication costs and can expose sensitive information (Shi et al., 2025). Parameter-only exchange avoids these auxiliary disclosures, although model updates still require protection against information leakage.

5 FedASTRA: Learning from Local Structure

Section 4 yields a design principle: local prediction can outperform a fitted richer-information model when its fitting cost exceeds the label-information gain. Guided by this principle, FedASTRA stays within the observable local release: it learns diffusion scales from anonymous incidence, predicts masked representations, and communicates parameters only, eliminating the need for cross-client pairing. Under probabilistic recovery and approximation conditions, FedASTRA is asymptotically optimal relative to permitted local information. Under representation and task-alignment conditions, its actual updates achieve global population-risk convergence with an $O(1/T)$ optimization term.

Learning diffusion scales. For binary incidence B_k with nonempty columns, let $D_{v,k} = \mathrm{diag}(B_k\mathbf{1})$, $D_{e,k} = \mathrm{diag}(B_k^\top\mathbf{1})$, and $I_{0,k} = \mathrm{diag}(\mathbf{1}\{(D_{v,k})_{vv} = 0\})$. Using the Moore–Penrose inverse $\dagger$, define the node-to-node transition on these observed local memberships by

$$P_k = D_{v,k}^{\dagger} B_k D_{e,k}^{-1} B_k^\top + I_{0,k}. \tag{18}$$

This walk averages over incident relations and their members, leaving isolated nodes fixed (Zhou et al., 2006). For latent width d, encode $U_k = E_{\theta_{\mathrm{enc}}}(X_k) \in \mathbb{R}^{|V_k|\times d}$ and use $R \ge 0$ positive-time channels with increments $\eta \in \mathbb{R}^{R\times d}$. The context parameters $\vartheta = \mathrm{ctx}(\theta) = (\theta_{\mathrm{enc}}, \eta)$ belong to the online state θ. The positive floor $\epsilon_{\mathrm{ord}} > 0$ orders each coordinate's diffusion times:

$$\mathsf{K}_k(t) = \exp\{t(P_k - I)\} = \sum_{\ell=0}^{\infty} e^{-t}\frac{t^\ell}{\ell!}P_k^\ell, \qquad t \ge 0, \tag{19}$$

$$t_{bj} = \sum_{a=1}^{b}\{\mathrm{softplus}(\eta_{aj}) + \epsilon_{\mathrm{ord}}\}, \qquad b \in [R], \quad j \in [d], \tag{20}$$

$$[\mathsf{K}_k(\boldsymbol{t}_b)U_k]_{:j} := \mathsf{K}_k(t_{bj})U_{k,:j}, \qquad W_k(X_k \mid \vartheta) = [\,U_k, \mathsf{K}_k(\boldsymbol{t}_1)U_k, \ldots, \mathsf{K}_k(\boldsymbol{t}_R)U_k\,]. \tag{21}$$

Each channel mixes walk lengths with Poisson weights (Kondor & Lafferty, 2002), while $R = 0$ retains only U_k. Each t_{bj} is the expected walk length for coordinate j. Learning these times adapts the extent of local mixing, while the undiffused channel retains the original encoded attributes. Thus R controls scales, independently of observation order q. Write $W_{k,\theta}(X) = W_k(X \mid \mathrm{ctx}(\theta))$. Column relabeling preserves both features and coupled training updates. Appendices C.8–C.9 prove stability and give a shared recurrence with Poisson-tail error control for each learned diffusion time.

Local prediction and federation. Two masked views predict a full-view exponential moving average (EMA) target through $\mathcal{L}_{\mathrm{pred,k}}$, combining cosine alignment, variance–covariance regularization, and view agreement (Appendix D.1). For classifier $c_{\theta_k} : \mathbb{R}^{(R+1)d} \to \mathbb{R}^C$, context $\vartheta_k = \mathrm{ctx}(\theta_k)$, labeled set $\mathcal{V}_k^{\mathrm{tr}}$ of size $n_k^{\mathrm{tr}} > 0$, and predictive weight $\lambda_{\mathrm{pred}} \ge 0$, the local training objective is

$$\mathcal{L}_k = (n_k^{\mathrm{tr}})^{-1}\sum\nolimits_{v\in\mathcal{V}_k^{\mathrm{tr}}} \mathrm{CE}(c_{\theta_k}(W_k(X_k \mid \vartheta_k)_v), y_v) + \lambda_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred},k}. \tag{22}$$

Algorithm 1 FedASTRA with an ordered client-local diffusion basis

Require: Rounds T, steps S_{loc}, clients $C_u \subseteq [K]$, counts n_k^{tr}, weight λ_{pred}, momentum μ.

1: Initialize θ^0 and client-local operators P_k.
2: **for** $u = 0, \dots, T-1$ **do**
3: **for** $k \in C_u$ in parallel **do**
4: Set $\theta_k \leftarrow \theta^u$ and $\bar{\vartheta}_k \leftarrow \text{ctx}(\theta_k)$.
5: Repeat S_{loc} times: update θ_k on Equation (22), then set $\bar{\vartheta}_k \leftarrow \mu\bar{\vartheta}_k + (1-\mu)\,\text{ctx}(\theta_k)$.
6: Return $\theta_k^{u+1} \leftarrow \theta_k$ and discard the target.
7: **end for**
8: $\theta^{u+1} \leftarrow \frac{\sum_{k\in C_u} n_k^{\text{tr}} \theta_k^{u+1}}{\sum_{k\in C_u} n_k^{\text{tr}}}$.
9: **end for**
10: Select one validation checkpoint and evaluate test data once.

Here CE is cross-entropy for $y_v \in [C]$. Algorithm 1 updates target $\bar{\vartheta}_k$ with momentum $\mu \in [0, 1)$ and transmits only online parameters, weighted by each client's number of labeled training nodes.

Local information and optimal prediction. Under the rank, attribute, and incidence-recovery conditions of Theorem 7 in Appendix C.11, the optimal cross-entropy gap between the complete heat representation and local information is $O(\varepsilon_n \log(1/\varepsilon_n))$, where $\varepsilon_n \to 0$ bounds joint failure on n nodes. On successful recovery, the complete heat representation determines the permitted local release. The residual risk bound accounts for decoder failure through its binary entropy and the finite label alphabet. Thus any local prediction rule can be reproduced on the event of successful recovery.

From client risks to federated risk. Let w_k^{pop} be test-client probabilities and $\mathcal{R}_{\text{fed}}$ their weighted cross-entropy. For the model $\widehat{\theta}$ selected after T rounds on $N = \sum_k n_k^{\text{tr}}$ labels, Theorem 8 gives

$$\mathbb{E}\left[\mathcal{R}_{\text{fed}}(\widehat{\theta}) - \sum_k w_k^{\text{pop}} \mathcal{R}^*_{\text{loc},k}\right] \le \mathbb{E}\left[\sum_k w_k^{\text{pop}}(d_k + a_k) + a_{\text{share}}\right] + 2u_N + e_{N,T}. \tag{23}$$

Here $\mathcal{R}^*_{\text{loc},k}$ is client k's optimal local-information risk. The terms $d_k, a_k, a_{\text{share}}$ measure information loss, local approximation error, and shared-parameter cost. Finally, u_N bounds expected uniform training-to-population discrepancy, and $e_{N,T}$ bounds empirical optimization error, addressed next.

Global convergence along the implemented updates. Let J_N be evaluation-mode training cross-entropy on domain Θ, convex in the final affine head at fixed features. After round u_0, assume each representation admits a head within distance $D_{\text{cls}} > 0$ of the current head and loss at most $\inf_\Theta J_N + a_{\text{ctx}}$, with $a_{\text{ctx}} \ge 0$. If the actual-update decrease bound (Proposition 9) is at least $\gamma > 0$ times the squared head-gradient norm, Theorem 9 bounds $e_{N,T} = \mathbb{E}[J_N(\widehat{\theta}) - \inf_\Theta J_N]$ by

$$e_{N,T} \le \mathbb{E}\left[a_{\text{ctx}} + \frac{D_{\text{cls}}^2}{\gamma(T-u_0)} + s^{\text{sel}}_{N,T}\right], \qquad T > u_0. \tag{24}$$

Here $s^{\text{sel}}_{N,T} = [J_N(\widehat{\theta}) - J_N(\theta^T)]_+$ accounts for validation selection, with $[x]_+ = \max\{x, 0\}$. Substitution into Equation (23) gives the population-risk bound. For fixed u_0 and uniformly bounded $\mathbb{E}[D_{\text{cls}}^2/\gamma]$, the optimization term is $O(1/T)$. Vanishing residual terms then give local-information Bayes consistency. Appendix C.12 also covers imperfect update alignment.

6 Experiments

We evaluate FedASTRA through four research questions (RQ1-RQ4) concerning predictive performance, mechanism validation, information access, and scalability:

- *RQ1: How effective is FedASTRA on identifier-free native hypergraphs?*
- *RQ2: Do the empirical results support the proposed analysis and FedASTRA's design?*
- *RQ3: How does FedASTRA compare under different information-access settings?*
- *RQ4: Can FedASTRA scale to large real-world hypergraphs without sampling?*

Additional experimental settings and results are provided in Appendices D and E, respectively.

Datasets. We evaluate four native hypergraph datasets: CoraCA coauthorship, DBLP4k bibliographic, IMDB4k movie, and Trivago hotel-session data (Yadati et al., 2019; Sun et al., 2011; Fu et al., 2020; Feng et al., 2019; Gao et al., 2023; Li et al., 2026), together with the Cora and CiteSeer citation networks, the Facebook Page-Page social network (Yang et al., 2016; Rozemberczki et al., 2021), and large ogbn-papers100M dataset (Hu et al., 2020). Appendix D.3 details the datasets.

Baselines. We consider two groups of baselines.(1) Federated adaptations (FedAvg (McMahan et al., 2017)) of native-hypergraph methods, including Fed-HGNN (HGNN (Feng et al., 2019)), Fed-HNHN (HNHN (Dong et al., 2020)), Fed-HyperGCN (HyperGCN (Yadati et al., 2019)), Fed-AllDeepSets (DeepSets (Zaheer et al., 2017)), Fed-AllSetTransformer (AllSet (Chien et al., 2022)), Fed-UniGCNII (UniGCNII (Huang & Yang, 2021)) and Fed-ED-HNN (ED-HNN (Wang et al., 2023)). (2) Federated graph learning methods with different levels of cross-client information access, including FedSage with no access and FedSage+ with generated information (Zhang et al., 2021), FedGCN (Yao et al., 2023) and FedCog (Lei et al., 2023) with direct access to correspondence. Details of baselines are presented in Appendix D.4.

Implementation Details. Fed-HNHN uses the configuration reported by FedHGL (Luo et al., 2025), and Fed-HyperGCN uses the local HyperGCN projection. FedASTRA uses two masks, a local EMA target, and parameter-only FedAvg (McMahan et al., 2017; Kingma & Ba, 2015). Results are averaged over five runs, except three for Fed-HNHN and all Trivago-related experiments. We report node classification accuracy (Acc.) and Macro-F1. The nodes are partitioned across $K \in \{3, 6, 9\}$ clients using class-conditioned Dirichlet sampling (Luo et al., 2025). All methods follow the same data splits and training and evaluation protocols. Appendix D details the experimental protocols.

RQ1: performance on identifier-free native hypergraphs. As shown in Table 1, FedASTRA achieves the best Accuracy and Macro-F1 across all four hypergraph datasets from distinct domains, especially on CoraCA, outperforming the baseline by 6.5% and 8.5%, respectively. The consistent cross-domain gains support FedASTRA's correspondence-invariant design, which avoids global structure recovery while approaching the optimum permitted by local information under the stated conditions. In contrast, direct FedAvg adaptations of traditional hypergraph models may underexploit local information when global structure is fragment. More detailed results under different number of clients are reported in Appendix E.1.

Table 1: Performance on identifier-free native hypergraphs. Results are averaged over $K \in \{3, 6, 9\}$.

Method	CoraCA		DBLP4k		IMDB4k		Trivago	
	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Fed-HGNN	68.64 ± 1.45	64.35 ± 1.71	77.63 ± 1.35	76.98 ± 1.18	48.85 ± 1.29	48.59 ± 1.23	21.85 ± 0.19	4.27 ± 0.20
Fed-HNHN	54.73 ± 1.17	50.42 ± 1.76	69.58 ± 1.18	68.79 ± 1.12	44.96 ± 1.50	43.05 ± 1.24	25.68 ± 1.02	7.72 ± 0.24
Fed-HyperGCN	64.02 ± 1.64	58.07 ± 1.95	72.69 ± 0.77	71.46 ± 0.86	44.14 ± 1.16	40.86 ± 2.38	9.17 ± 0.20	0.13 ± 0.02
Fed-AllDeepSets	39.63 ± 1.83	27.08 ± 2.04	39.35 ± 0.55	32.45 ± 1.71	40.32 ± 1.15	34.12 ± 2.20	10.88 ± 0.15	0.27 ± 0.01
Fed-AllSetTransformer	30.22 ± 0.00	6.63 ± 0.00	30.99 ± 1.29	21.81 ± 0.86	37.04 ± 0.05	18.10 ± 0.07	8.99 ± 0.00	0.10 ± 0.00
Fed-UniGCNII	59.54 ± 1.31	50.11 ± 3.86	77.35 ± 2.26	75.98 ± 2.60	41.62 ± 0.74	30.57 ± 3.32	9.15 ± 0.13	0.15 ± 0.03
Fed-ED-HNN	59.62 ± 1.65	53.28 ± 2.30	72.10 ± 2.54	71.36 ± 2.55	43.46 ± 0.89	36.58 ± 1.05	9.19 ± 0.29	0.13 ± 0.03
FedASTRA	**75.17 ± 0.42**	**72.88 ± 0.60**	**79.22 ± 0.98**	**78.61 ± 0.80**	**51.13 ± 0.97**	**51.06 ± 0.77**	**29.09 ± 1.19**	**7.96 ± 0.10**

RQ2: validation on proposed analysis and mechanism. We first validate the proposed identification analysis through two-client audits. Among 120 evaluated cells, no higher-order co-membership target is point identified, and the median endpoint-hull width increases from 36.37% at $q = 2$ to 93.71% at $q = 5$. Type coarsening preserves endpoint containment in all 40 audited cells, while 32 column permutations leave P_k, heat features, and logits unchanged. These results confirm the predicted ambiguity under latent correspondence and the correspondence-invariant property of the learner representation. Besides, higher-order truncation exhibits different predictive trends across datasets, indicating that its impact depends on how the lost higher-order structure aligns with task-relevant signal. Complete audit configurations and results are reported in Appendix E.3.

For mechanism ablations, we evaluate heat-basis width $R \in \{0, 1, 2, 4, 8\}$ and the masked prediction objective. Figure 2 show that removing diffusion ($R = 0$) degrades performance, while moderate heat widths improve accuracy. Removing the predictive objective ($\lambda_{pred} = 0$) also consistently reduces performance, demonstrating its complementary role in FedASTRA. Complete results are reported in Appendix E.3.

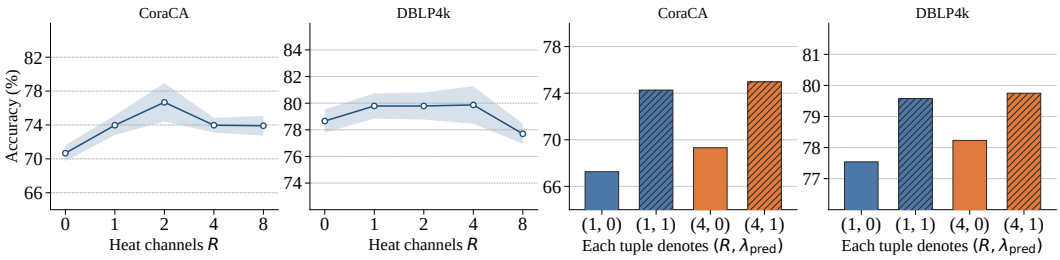


Figure 2: Ablation study on the adaptive diffusion (left)and the masked prediction (right).

RQ3: performance against baselines with varying information access. We empirically contrast the information-access regimes discussed in Section 4, as depicted in Table 2. FedASTRA outperforms FedSage and FedGCN (0-hop), which use only local information, and also surpasses FedSage+, despite the latter using additional communication to generate cross-client information. Although FedGCN (2-hop) and FedCog are stronger under a different setting with known correspondence, the gap is substantially narrowed, suggesting that the remaining gap is not solely missing cross-client information, but also how effectively the local structure is represented and exploited. The stable performance of FedASTRA as the client number grows further indicates that stronger local modeling can partially offset increasing structural information loss. Appendix E.2 provides complete metrics, information-access details, and traffic accounting.

Table 2: Performance againest federated graph learning methods. Communication traffic (MiB) is averaged across datasets and client counts, with model traffic per round and auxiliary traffic per run.

Method	Cora			CiteSeer			Facebook			Average communication traffic	
	$K=3$	$K=6$	$K=9$	$K=3$	$K=6$	$K=9$	$K=3$	$K=6$	$K=9$	Model/round	Auxiliary/run
FedSage	70.35 ± 1.96	63.18 ± 0.97	58.65 ± 1.85	65.08 ± 0.96	63.73 ± 0.74	62.93 ± 1.72	72.63 ± 0.73	65.64 ± 1.81	60.41 ± 1.61	4.82	0
FedGCN (0-hop)	74.07 ± 1.57	65.90 ± 2.05	62.07 ± 1.79	68.68 ± 1.50	65.83 ± 0.56	65.14 ± 1.80	75.14 ± 0.95	67.87 ± 1.71	64.30 ± 1.22	2.41	0
FedSage+	69.67 ± 0.71	63.11 ± 1.66	61.02 ± 1.40	63.16 ± 1.17	62.53 ± 1.66	62.95 ± 1.62	68.17 ± 0.33	61.85 ± 0.95	58.56 ± 1.33	4.82	17314.45
FedGCN (2-hop)	84.05 ± 1.17	84.05 ± 1.16	84.12 ± 0.92	72.20 ± 1.44	72.27 ± 1.53	72.38 ± 1.50	83.21 ± 0.76	82.84 ± 0.50	82.73 ± 0.85	2.41	673.79
FedCog	84.29 ± 1.50	84.34 ± 1.06	83.92 ± 0.78	72.21 ± 1.49	72.38 ± 1.04	72.51 ± 1.42	82.46 ± 0.69	82.50 ± 0.84	82.15 ± 0.94	2.41	752.27
FedASTRA	77.54 ± 1.08	74.45 ± 1.60	73.33 ± 1.22	67.61 ± 2.32	68.00 ± 1.86	67.92 ± 1.88	76.88 ± 1.44	75.57 ± 0.94	75.12 ± 1.49	13.53	0

RQ4: scalability on large-scale data. We evaluate FedASTRA on ogbn-papers100M after local one-hop lifting, as summarized in Table 3. We also include Fed-MLP, Fed-GCN (Kipf & Welling, 2017), and Fed-SIGN (Frasca et al., 2020) as baselines. FedASTRA outperforms the strongest baseline by up to 9.69% under comparable communication overhead. Its scalability stems from the correspondence-free local diffusion design and. the strong performance further suggests that adaptive diffusion can better exploit local structures. Appendix D.4 reports detailed results.

Table 3: Performance on complete ogbn-papers100M dataset. Nodes are partitioned across 8 clients. Traffic counts bidirectional trainable-parameter per round (MiB). BatchNorm buffers stay local.

Method	Acc. (%)	Macro-F1 (%)	#Parameters	Parameter trafffic/round
Fed-MLP	33.61	13.74	983,380	60.02
Fed-SIGN	38.42	16.56	984,442	60.09
Fed-HGNN	31.04	9.97	982,672	59.98
Fed-HNHN	35.93	13.46	34,672	2.12
Fed-HyperGCN	40.38	21.44	984,442	60.09
Fed-GCN	39.70	17.40	983,380	60.02
FedASTRA	**50.07**	**29.73**	976,504	59.60

7 Conclusion

We establish identification and computational boundaries for latent hyperedge correspondence and quantify when fitting costs outweigh the benefit of additional information. FedASTRA combines invariant local diffusion, masked prediction, and parameter-only communication. Under stated assumptions, information-preservation and global risk bounds connect observable memberships to prediction. Results on native hypergraphs, federated subgraphs, and large-scale ogbn-papers100M data support this approach under explicit information-access and communication constraints.

REFERENCES

Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *J. ACM*, 42(4):844–856, 1995.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Jinheon Baek, Wonyong Jeong, Jiongdao Jin, Jaehong Yoon, and Sung Ju Hwang. Personalized subgraph federated learning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1396–1415. PMLR, 2023.

Yikun Bai, Rocio Diaz Martin, Abihith Kothapalli, Hengrong Du, Xinran Liu, and Soheil Kolouri. Partial gromov-wasserstein metric. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3322a9a72a1707de14badd5e552ff466-Abstract-Conference.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iWpcWZ8phD.

Garrett Birkhoff. Tres observaciones sobre el álgebra lineal. *Universidad Nacional de Tucumán, Revista A*, 5:147–151, 1946.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Guy Bresler, Chenghao Guo, and Yury Polyanskiy. Thresholds for reconstruction of random hypergraphs from graph projections. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 632–647. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/bresler24a.html.

Guy Bresler, Chenghao Guo, Yury Polyanskiy, and Andrew Yao. Partial and exact recovery of a random hypergraph from its graph projection. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 534–593. PMLR, 2025. URL https://proceedings.mlr.press/v291/bresler25a.html.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

Yang Chen, Cong Fang, Zhouchen Lin, and Bing Liu. Relational learning in pre-trained models: A theory from hypergraph recovery perspective. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6666–6698. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24l.html.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Samir Chowdhury, Tom Needham, Ethan Semrad, Bei Wang, and Youjia Zhou. Hypergraph cooptimal transport: Metric and categorical properties. *Journal of Applied and Computational Topology*, 8(5):1171–1230, 2024. doi: 10.1007/s41468-023-00142-9. URL https://doi.org/10.1007/s41468-023-00142-9.

Yihe Dong, Will Sawin, and Yoshua Bengio. HNHN: Hypergraph networks with hyperedge neurons. *arXiv preprint arXiv:2006.12278*, 2020. URL https://arxiv.org/abs/2006.12278.

Rodney G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness II: on completeness for W[1]. *Theor. Comput. Sci.*, 141(1&2):109–131, 1995.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049. URL https://doi.org/10.1080/01621459.1969.10501049.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Benjamin Chamberlain, Michael Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020. URL https://arxiv.org/abs/2004.11198.

Lele Fu, Bowen Deng, Sheng Huang, Tianchi Liao, Shirui Pan, and Chuan Chen. Less is more: Federated graph learning with alleviating topology heterogeneity from A causal perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17727–17745. PMLR, 2025. URL https://proceedings.mlr.press/v267/fu25a.html.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. MAGNN: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*, pp. 2331–2341. ACM, 2020. doi: 10.1145/3366423.3380297. URL https://doi.org/10.1145/3366423.3380297.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. HGNN+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2023. doi: 10.1109/TPAMI.2022.3182052. URL https://doi.org/10.1109/TPAMI.2022.3182052.

Yue Gao, Yifan Feng, Shiquan Liu, Xiangmin Han, Shaoyi Du, Zongze Wu, and Han Hu. Hypergraph foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(4):4063–4080, 2026.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

Jing Huang and Jie Yang. UniGNN: A unified framework for graph and hypergraph neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2563–2569, 2021. doi: 10.24963/ijcai.2021/353. URL https://www.ijcai.org/proceedings/2021/0353.pdf.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher (eds.), *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9. URL https://doi.org/10.1007/978-1-4684-2001-2_9.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

Sunwoo Kim, Shinhwan Kang, Fanchen Bu, Soo Yong Lee, Jaemin Yoo, and Kijung Shin. Hypeboy: Generative self-supervised representation learning on hypergraphs. In *ICLR*. OpenReview.net, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 315–322, 2002. URL https://people.cs.uchicago.edu/~risi/papers/diffusion-kernels.pdf.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://doi.org/10.1002/nav.3800020109.

Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2, position paper.

Dongjin Lee and Kijung Shin. I'm me, we're us, and i'm us: Tri-directional contrastive learning on hypergraphs. In *AAAI*, pp. 8456–8464. AAAI Press, 2023.

Kyuhan Lee, Geon Lee, and Kijung Shin. MARIOH: Multiplicity-aware hypergraph reconstruction. In *2025 IEEE 41st International Conference on Data Engineering*, pp. 3113–3125, 2025. doi: 10.1109/ICDE65448.2025.00233. URL https://doi.org/10.1109/ICDE65448.2025.00233.

Runze Lei, Pinghui Wang, Junzhou Zhao, Lin Lan, Jing Tao, Chao Deng, Junlan Feng, Xidian Wang, and Xiaohong Guan. Federated learning over coupled graphs. *IEEE Transactions on Parallel and Distributed Systems*, 34(4):1159–1172, 2023. doi: 10.1109/TPDS.2023.3240527.

Fan Li, Xiaoyang Wang, Wenjie Zhang, Ying Zhang, and Xuemin Lin. DHG-Bench: A comprehensive benchmark for deep hypergraph learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lhsb1ChUDF.

Xunkai Li, Zhengyu Wu, Wentao Zhang, Yinlin Zhu, Ronghua Li, and Guoren Wang. Fedgta: Topology-aware averaging for federated graph learning. *Proc. VLDB Endow.*, 17(1):41–50, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Linfeng Luo, Zhiqi Guo, Fengxiao Tang, Zihao Qiu, and Ming Zhao. Federated hypergraph learning with local differential privacy: Toward privacy-aware hypergraph structure completion. In *2025 IEEE International Conference on Data Mining*, pp. 1400–1409, 2025. doi: 10.1109/ICDM65498.2025.00149. URL https://doi.org/10.1109/ICDM65498.2025.00149.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478. URL https://doi.org/10.1007/b97478.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672. PMLR, 2016. URL https://proceedings.mlr.press/v48/peyre16.html.

William P. Pierskalla. Letter to the editor—the multidimensional assignment problem. *Operations Research*, 16(2):422–431, 1968. doi: 10.1287/opre.16.2.422. URL https://doi.org/10.1287/opre.16.2.422.

Gathika Ratnayaka, James Nichols, and Qing Wang. Learning partial graph matching via optimal partial transport. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83575–83595, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/cfa8440d500a6a6867157dfd4eaff66e-Paper-Conference.pdf.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*. OpenReview.net, 2018.

Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021. doi: 10.1093/comnet/cnab014. URL https://doi.org/10.1093/comnet/cnab014.

Shanghao Shi, Ning Wang, Yang Xiao, Chaoyu Zhang, Yi Shi, Y. Thomas Hou, and Wenjing Lou. Scale-MIA: A scalable model inversion attack against secure federated learning via latent space reconstruction. In *Network and Distributed System Security Symposium*, 2025. doi: 10.14722/ndss.2025.240644. URL https://www.ndss-symposium.org/ndss-paper/scale-mia-a-scalable-model-inversion-attack-against-secure-federated-learning-via-

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011. doi: 10.14778/3402707.3402736. URL https://www.vldb.org/pvldb/vol4/p992-sun.pdf.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 2019. URL https://proceedings.mlr.press/v97/titouan19a.html.

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RiTjKoscnNd.

Yanbang Wang and Jon Kleinberg. From graphs to hypergraphs: Hypergraph projection and its reconstruction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e92b8441b702d142435f243806f9ec5a-Abstract-Conference.html.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1efa39bcaec6f3900149160693694536-Abstract.html.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48. PMLR, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Yuhang Yao, Weizhao Jin, Srivatsan Ravi, and Carlee Joe-Wong. FedGCN: Convergence-communication tradeoffs in federated training of graph convolutional networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79748–79760, 2023. doi: 10.52202/

075280-3491. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fc07feae9af49dd3f1a1e049b77f4e17-Abstract-Conference.html.

Jean-Gabriel Young, Giovanni Petri, and Tiago P. Peixoto. Hypergraph reconstruction from network data. *Communications Physics*, 4:135, 2021. doi: 10.1038/s42005-021-00637-w. URL https://doi.org/10.1038/s42005-021-00637-w.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/34adeb8e3242824038aa65460a47c29e-Abstract.html.

Stephen Y. Zhang, Fangfei Lan, Youjia Zhou, Agnese Barbensi, Michael P. H. Stumpf, Bei Wang, and Tom Needham. Geometry of the space of partitioned networks: A unified theoretical and computational framework. *Information and Inference: A Journal of the IMA*, 15(3):iaag022, 2026. doi: 10.1093/imaiai/iaag022. URL https://doi.org/10.1093/imaiai/iaag022.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, volume 19, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/hash/dff8e9c2ac33381546d96deea9922999-Abstract.html.

Appendix

Contents

A More Related Work . . . 16
B Formal Results and Identification Boundaries . . . 16
B.1 Important Notation . . . 17
B.2 Observation Channel and Design-based Recovery . . . 17
B.3 Two-client Assumptions and Weight-stratum Boundary . . . 18
C Proofs . . . 18
C.1 Two-client Completion and Point Identification . . . 18
C.2 Order-q Mass Identification . . . 19
C.3 Structured Tractable Class: Anti-Monge Bounds . . . 20
C.4 Multi-client Computational Boundary . . . 20
C.5 Boolean-zeta Inversion and Partition Censoring . . . 22
C.6 Horvitz–Thompson Estimation . . . 22
C.7 Observational Equivalence and a Two-point Bound . . . 23
C.8 Released-data Factorization and Observable Spectral Approximation . . . 24
C.9 Observable Diffusion and Predictive Dynamics . . . 25
C.10 Downstream Risk with Additional Information . . . 27
C.11 Randomized Recovery from Diffusion Features and Local Prediction Risk . . . 29
C.12 Stochastic Federated Optimization and Risk Transfer . . . 31
D Experimental Protocols and Implementation Details . . . 35
D.1 FedASTRA Implementation . . . 35
D.2 Graph Partitions and Lifts . . . 37
D.3 Dataset Details . . . 38
D.4 Federated Protocols and Prespecified Configurations . . . 38
D.5 Metrics and Additional Measurements . . . 41
E More Experiment Results . . . 42
E.1 Performance on Native Hypergraphs . . . 42
E.2 Performance Against Baselines with Varying Information Access . . . 42
E.3 Validation on Proposed Analysis and Mechanism . . . 44
E.4 Scalability on Large-scale Data . . . 46
F AI Use Statement . . . 47
G Reproducibility Statement . . . 47
H Ethics Statement . . . 47

A More Related Work

Latent correspondence and partial identification. Unknown correspondence is classical in record linkage (Fellegi & Sunter, 1969), where latent pairings between records must be inferred from incomplete identifying information. Partial identification instead retains the set of target values compatible with the observations when point identification is unavailable (Manski, 2003). Our setting combines these perspectives at the level of higher-order relations: client-local hyperedge fragments are observed, but their cross-client instance correspondence is not. Rather than selecting or estimating one global linkage, we characterize the full compatibility class induced by the released fragments and derive sharp target bounds over all compatible alignments.

Alignment under latent correspondence. When object identities are not shared, relational alignment is often formulated through latent couplings. Gromov–Wasserstein and fused Gromov–Wasserstein formulations compare structured objects without requiring a common coordinate system (Peyré et al., 2016; Vayer et al., 2019), while recent partial formulations relax full-mass or full-matching requirements (Bai et al., 2025; Ratnayaka et al., 2025). For higher-order structures, HyperCOT develops transport-based comparisons of hypergraphs (Chowdhury et al., 2024), and recent partitioned-network geometry unifies graph and hypergraph alignment within a common metric framework (Zhang et al., 2026). These approaches optimize an alignment or distance under a specified geometry. But more specifically here, we are more concerned with which target values remain invariant to latent cross-client correspondence and which vary within the compatibility class induced by a fixed federated release.

Hypergraph reconstruction from coarsened observations. Some existing papers study whether higher-order structure can be reconstructed from lower-order or otherwise coarsened observations. Early work develops statistical reconstruction from network data (Young et al., 2021). Wang and Kleinberg characterize information loss under hypergraph projection and study reconstruction with additional structural knowledge (Wang & Kleinberg, 2024). For random uniform hypergraphs, recent work gives information-theoretic and algorithmic thresholds for exact recovery (Bresler et al., 2024), followed by a characterization of both partial and exact recovery across hyperedge sizes and multiplicity regimes (Bresler et al., 2025). Multiplicity-aware supervised reconstruction provides a complementary algorithmic approach (Lee et al., 2025), while recent learning theory has also formulated relational learning itself through a hypergraph-recovery model (Chen et al., 2024). These works primarily ask when a latent source hypergraph is recoverable under projection or a specified generative model. Our observation is however different: clients retain higher-order fragments themselves, but their cross-client instance identities are suppressed. We therefore retain all compatible global hypergraphs and characterize what is identifiable without directly assuming that one global completion can be recovered.

Assignment and multiway matching complexity. Two-way correspondence is closely related to classical linear assignment and bipartite matching (Birkhoff, 1946; Kuhn, 1955), whereas multidimensional assignment and three-dimensional matching exhibit fundamentally different computational behavior (Pierskalla, 1968; Karp, 1972). Parameterized algorithms for set packing and matching draw on tools such as color coding, while the W-hierarchy distinguishes fixed-parameter tractability from parameterized intractability (Alon et al., 1995; Downey & Fellows, 1995). Recent partial graph matching still exposes tractable assignment structure in important pairwise settings (Ratnayaka et al., 2025). Our result places these phenomena within one latent-correspondence problem: under the same observation and reward rules, two-client correspondence is polynomial, fixed-client multiway matching is NP-complete but fixed-parameter tractable in the stated joint parameter, and unbounded client count is W[1]-complete in matching size.

B Formal Results and Identification Boundaries

This section records the observation assumptions used in the proofs: which fragment attributes are retained, when partition censoring admits unbiased correction, and when unknown correspondence is exactly a bipartite assignment.

B.1 Important Notation

Table 4: Core notation for the observation model, identification results, algorithm, and experiments.

Symbol	Meaning
$\mathcal{H} = (V, \mathcal{E}, X, r, w)$	Finite node-attributed, typed, weighted multihypergraph.
A_e, r_e, w_e	Support, type in $\mathcal{R}$, and positive weight of relation instance $e \in \mathcal{E}$.
$w_r(A)$	Aggregate weight of type-r instances with support A.
$M^{\mathcal{H}}_{q,r}(S)$	Weighted type-r order-q co-membership count for $S \in \binom{V}{q}$.
Z, z, V_k	Random node-to-client assignment, one realization, and the nodes assigned to client k.
c, $\upsilon_{e,k}$	Minimum visible fragment size and the analyst's local/cross status mark.
Γ, $\mathcal{M}_\Gamma$	Public observation assumptions and its allowed latent-hypergraph class.
$Y_{\mathrm{A}}, y_{\mathrm{A}}$	Analyst release as a random element and one realized value.
$Y_{\mathrm{L}}, y_{\mathrm{L}}$	Coarsened learner release $Y_{\mathrm{L}} = \rho(Y_{\mathrm{A}})$ and one realized value.
$\mathcal{F}_{\mathrm{A}}, \mathcal{F}_{\mathrm{L}}$	Analyst and learner compatibility classes defined in Equation (4) and Corollary 2.
$\widetilde{M}_{q,r}(\cdot, z)$, τ_q	Retained type-r coordinates and the type-aggregated erased-mass fraction.
π_S, π_q	Inclusion probability of a q-set and its common value under an exchangeable randomized partition design.
$\mathcal{I}_{q,\psi}(y_{\mathrm{A}})$	Values of $\Phi_{q,\psi}$ over $\mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}}, \Gamma)$.
$h = (r, \omega)$, C^h, $\mathbb{B}_{m_h}$	Exact type-and-weight stratum, its assignment cost, and the Birkhoff polytope.
$B_k, D_{v,k}, D_{e,k}, I_{0,k}$	Client incidence matrix, node and relation degrees, and isolated-node indicator.
P_k, P_k^{sym}	Observable Markov operator and its symmetric representative on positive-degree nodes.
$\mathsf{K}_k(t)$	Continuous-time local diffusion $\exp\{t(P_k - I)\}$ at time $t \geq 0$.
R, t_{bj}, ϵ_{ord}	Number of positive heat channels, channel-b time for feature j, and its positive increment floor.
$W_k(X_k \mid \vartheta)$, $W_{k,\theta}(X_k)$	Heat-basis representation and its shorthand with $\vartheta = \mathrm{ctx}(\theta)$.
L_{tail}, $\varepsilon_{\mathrm{tail}}$	Shared recurrence degree and its deterministic Poisson-tail tolerance.
θ, ϑ, $\bar{\vartheta}$	Complete online state, its context coordinates $(\theta_{\mathrm{enc}}, \eta)$, and the client-local EMA target context.
p_θ, c_θ, y_v	Rowwise predictor, rowwise classifier, and class label of node v.
γ_{view}, λ_{pred}, μ	View-agreement weight, predictive-loss weight, and target momentum.
$\mathcal{P}$, ξ, $\mathcal{T}_T$	Fixed training protocol, algorithmic randomness, and the transcript through round T.

Probability and conditioning convention. Unless a statement says otherwise, $\mathcal{H}$, z, Γ, and the realized release are fixed. Expectations and variances in Theorem 3 are taken over the declared random partition Z. Corollary 3 explicitly averages over a uniform latent matching. Algorithmic laws average over $\xi \sim \nu$, where the fixed law ν is independent of $\mathcal{H}$. The auxiliary variable N_t^{aux} below only expresses deterministic Poisson coefficients and tail bounds and is absent from implementation sampling. Appendix C.10 separately defines its downstream evaluation law.

B.2 Observation Channel and Design-based Recovery

B.2.1 Aggregate support–type weights and Boolean-zeta coordinates

The analyst release retains each fragment occurrence with its exact type, weight, and local/cross status. Assignment strata use these individual rows, including duplicates. Co-membership coordinates instead depend on $w_r(A)$, the total weight of rows with support A and type r. Aggregation loses the allocation of weight among duplicate occurrences, so it cannot replace the row-level table in the matching analysis. Erasing types further reduces the observation to $M_q = \sum_r M_{q,r}$. The following inclusion describes the resulting uncertainty on the same latent class $\mathcal{M}_\Gamma$.

Corollary 1 (Coarsened or missing type marks enlarge identified sets)**.** *Let $\chi : \mathcal{R} \to \mathcal{R}_\chi$ merge or suppress type labels, and let $\chi_\#$ apply this map to the release while preserving its other fields. For an exact-type release y_{ex}, set $y_\chi = \chi_\# y_{\mathrm{ex}}$ and $O_\chi = \chi_\# \circ O_{z,\Gamma}$, with public type restrictions pushed forward accordingly. On the common class $\mathcal{M}_\Gamma$, define*

$$\mathcal{F}_{\mathrm{ex}}(y_{\mathrm{ex}}, \Gamma) = \{\mathcal{H} \in \mathcal{M}_\Gamma : O_{z,\Gamma}(\mathcal{H}) = y_{\mathrm{ex}}\}, \qquad \mathcal{F}_\chi(y_\chi, \Gamma) = \{\mathcal{H} \in \mathcal{M}_\Gamma : O_\chi(\mathcal{H}) = y_\chi\}. \tag{25}$$

For a scalar target $\phi : \mathcal{M}_\Gamma \to \mathbb{R}$, write $\mathcal{I}_{\phi,\mathrm{ex}} = \phi(\mathcal{F}_{\mathrm{ex}}(y_{\mathrm{ex}}, \Gamma))$ and $\mathcal{I}_{\phi,\chi} = \phi(\mathcal{F}_\chi(y_\chi, \Gamma))$. Then

$$\mathcal{F}_{\mathrm{ex}}(y_{\mathrm{ex}}, \Gamma) \subseteq \mathcal{F}_\chi(y_\chi, \Gamma), \qquad \mathcal{I}_{\phi,\mathrm{ex}} \subseteq \mathcal{I}_{\phi,\chi}. \tag{26}$$

When the scalar endpoints are finite, this inclusion gives

$$\inf \mathcal{I}_{\phi,\chi} \le \inf \mathcal{I}_{\phi,\mathrm{ex}} \le \sup \mathcal{I}_{\phi,\mathrm{ex}} \le \sup \mathcal{I}_{\phi,\chi}. \tag{27}$$

Thus type coarsening can only preserve or widen the sharp endpoint interval.

Proof. Every hypergraph reproducing the exact marked table also reproduces its χ-coarsening, which proves the first inclusion. Taking the image under ϕ proves the second, and nested scalar sets give the endpoint inequalities. □

B.2.2 Independent and fixed-capacity designs

For fixed $\mathcal{H}$ and $q \ge c$, a coordinate $S \in \binom{V}{q}$ survives when all its nodes share a client. Independent assignments with $p_{vk} = \Pr\{\mathsf{Z}(v) = k\}$ give

$$\pi_S = \Pr\{\mathsf{Z}|_S \text{ is constant}\} = \sum_{k=1}^{K} \prod_{v \in S} p_{vk}. \tag{28}$$

For a uniformly random partition with fixed capacities $(n_1, \dots, n_K)$, exchangeability gives

$$\pi_q = \sum_{k=1}^{K} \frac{\binom{n_k}{q}}{\binom{|V|}{q}}. \tag{29}$$

Positive inclusion probabilities permit coordinatewise Horvitz–Thompson weighting (Horvitz & Thompson, 1952). Its unbiasedness concerns repeated randomized partitions. Proportions from a fixed institutional partition alone actually fail to specify the required assignment law.

B.2.3 Conditions for coordinate-wise partition censoring

Equation (6) requires retained multiplicity, exact weight, type, and $q \ge c$. Deduplication would change the count, so the data interface preserves duplicate and singleton columns. For $c > 1$, some nonempty fragments disappear. Completion analysis then requires constraints on original hyperedge counts or orders. The exact two-client result uses $c = 1$.

B.3 Two-client Assumptions and Weight-stratum Boundary

Theorem 1 assumes two clients, $c = 1$, observed local/cross status, one nonempty row at each client per cross-client hyperedge, and exact type-and-weight strata $h = (r, \omega)$ with $m_{1h} = m_{2h} =: m_h$. Each compatible completion is then a bijection within every stratum. Missing fragments, uncertain counts, or further clients change this feasible set. Weight bins also enlarge it: the public model must specify the allowed underlying weights, and endpoints must be optimized over that range rather than evaluated at an assumed bin midpoint.

Under the exact assumptions, general two-client targets admit assignment bounds, while total order-q mass admits sorting bounds. General three-client targets are NP-hard. For any client count, anchor-additive costs remain tractable because the objective separates into $K - 1$ independent bipartite assignments (Proposition 1).

C Proofs

We prove the structural bounds first, then connect observable heat features to prediction risk and the shared-model federated objective.

C.1 Two-client Completion and Point Identification

Fix a stratum $h = (r, \omega)$. Under the exact two-client assumptions, every $\varpi \in \mathrm{Sym}(m_h)$ is compatible. Each q-set in $F^h_{1i} \cup F^h_{2,\varpi(i)}$ lies within one fragment or intersects both. Summing the within-fragment terms over i is matching-invariant, while the mixed terms give $C^h_{i,\varpi(i)}$. Adding

known local contributions proves Equation (10). By Birkhoff–von Neumann integrality, the vertices of $\mathbb{B}_m$ are exactly the permutation matrices (Birkhoff, 1946). Both linear-program endpoints therefore attain compatible matchings and are sharp. Each costs $O(m_h^3)$ by Hungarian assignment (Kuhn, 1955).

The assignment dual also provides a certificate for the computed bound. For each stratum, the lower-endpoint dual is

$$\max_{u,v} \sum_i u_i + \sum_j v_j \quad \text{s.t.} \quad u_i + v_j \le C^h_{ij}. \tag{30}$$

The upper endpoint has a minimization dual with $u_i + v_j \ge C^h_{ij}$. Dual feasibility and equality of primal and dual values certify an endpoint, with equality on every pair in an optimal matching.

If $C_{ij} = \alpha_i + \beta_j$, every permutation sums to $\sum_i \alpha_i + \sum_j \beta_j$. Conversely, equality of all permutation sums implies equality after exchanging any two assigned columns, hence

$$C_{ij} + C_{i'j'} - C_{ij'} - C_{i'j} = 0. \tag{31}$$

When all such rectangular cross-differences vanish, set $\alpha_i = C_{i1}$ and $\beta_j = C_{1j} - C_{11}$. The cycle using row 1 and column 1 gives $C_{ij} = \alpha_i + \beta_j$. For $m = 1$ this representation is immediate. Finally, $J_m C J_m = 0$ if and only if C lies in the kernel of double centering, which is exactly the row–column-additive subspace. Thus the quotient of cost matrices by the kernel of assignment variation is represented uniquely by $J_m C J_m$. Since J_m has rank $m-1$, this image has dimension $(m-1)^2$. The quotient maps injectively into the following $(m-1)^2$ base four-cycle coordinates:

$$\left(C_{ij} - C_{i1} - C_{1j} + C_{11}\right)_{i,j=2}^{m} \tag{32}$$

Zero base coordinates imply row–column additivity, so these $(m-1)^2$ coordinates are independent on the quotient. The dimension is attainable with $q = 2$ and unit-weight singleton fragments by setting $\psi_{2,r}(\{a_i, b_j\}, X_{\{a_i,b_j\}}) = C_{ij}$. Repeated fragments or restricted test functions may reduce the realized dimension, without changing the endpoint argument. Since strata vary independently, point identification requires additivity in every stratum. □

C.2 Order-q Mass Identification

For a general target, the preceding proof requires an assignment solver. Total order-q mass has more structure: with $q \ge 2$, its mixed cost depends only on fragment sizes. This allows both endpoints to be computed by sorting. Within stratum $h = (r, \omega)$, set $a^h_i = |F^h_{1i}|$, $b^h_j = |F^h_{2j}|$, and $\omega_h = \omega$. Vandermonde's identity and its discrete mixed difference give

$$\kappa_q(a,b) = \binom{a+b}{q} - \binom{a}{q} - \binom{b}{q} = \sum_{\ell=1}^{q-1} \binom{a}{\ell}\binom{b}{q-\ell}, \tag{33}$$

$$\kappa_q(a,b) + \kappa_q(a',b') - \kappa_q(a,b') - \kappa_q(a',b) = \sum_{x=a}^{a'-1} \sum_{y=b}^{b'-1} \binom{x+y}{q-2} \ge 0 \tag{34}$$

for $a \le a'$ and $b \le b'$. Same-order pairing maximizes the total mass at T^+_q, while opposite-order pairing minimizes it at T^-_q. Writing the released local mass as M^{obs}_q gives sharp erased-mass bounds:

$$\tau^-_q = 1 - M^{\text{obs}}_q / T^-_q, \qquad \tau^+_q = 1 - M^{\text{obs}}_q / T^+_q. \tag{35}$$

The zero-total convention and explicit formulas for $T^{\pm}_q$ appear in Appendix C.3. Point identification holds exactly when the mixed difference vanishes throughout every stratum. Sorting computes both attainable endpoints in $O(\sum_h m_h \log m_h)$ time by sorting fragment sizes within each stratum.

To determine when these bounds coincide, take the constant-one test. Its stratum cost is $C^h_{ij} = \omega_h \kappa_q(a^h_i, b^h_j)$. Equation (34) yields the exact rectangular 2×2 cross-difference within each fixed type-and-weight stratum:

$$C^h_{ij} + C^h_{i'j'} - C^h_{ij'} - C^h_{i'j} = \omega_h \sum_{x=a^h_i}^{a^h_{i'}-1} \sum_{y=b^h_j}^{b^h_{j'}-1} \binom{x+y}{q-2}. \tag{36}$$

Theorem 1 makes vanishing of all these quantities necessary and sufficient for point identification. If both lists vary, take $a^h_{i'} = \max_i a^h_i$ and $b^h_{j'} = \max_j b^h_j$ with strictly smaller a^h_i, b^h_j. When $a^h_{i'}+b^h_{j'} \geq q$, the summand at $(x, y) = (a^h_{i'} - 1, b^h_{j'} - 1)$ is positive, so point identification fails. For $q = 2$ every nonempty summand equals one. Thus, at $q = 2$, all cross-differences vanish exactly when at least one size list is constant in each stratum. For general q, if one list is constant then all rectangles are empty in that direction. If both vary but $\max_i a^h_i + \max_j b^h_j < q$, every mixed q-node-subset count is zero. Together with the preceding strict argument, these are the only point-identified cases. □

C.3 Structured Tractable Class: Anti-Monge Bounds

Vandermonde's identity gives Equation (33) by removing the two within-fragment terms. For $b_q(n) = \binom{n}{q}$, Pascal's identity gives $b_q(n + 2) - 2b_q(n + 1) + b_q(n) = \binom{n}{q-2}$. Summing this difference over $x = a, \ldots, a' - 1$ and $y = b, \ldots, b' - 1$ proves Equation (34), since row-only and column-only terms cancel.

Within each stratum, sort $a^h_1 \leq \cdots \leq a^h_{m_h}$. If a proposed matching pairs $a^h_i \leq a^h_{i'}$ with $b^h_j > b^h_{j'}$, the anti-Monge inequality shows that swapping the two targets weakly increases the total mixed count. Repeating adjacent swaps proves that equal-order pairing maximizes it. Applying the same argument to the reversed b order proves the minimum. Non-negative retained weights preserve the inequality within each fixed stratum, and strata add. With invariant local mass M^{obs}_q, the completed mass under $\boldsymbol{\varpi} = (\varpi_h)_h$ is $T_q(\boldsymbol{\varpi}) = M^{\text{obs}}_q + \sum_h \omega_h \sum_i \kappa_q(a^h_i, b^h_{\varpi_h(i)})$. Since $1 - M^{\text{obs}}_q/T_q$ is nondecreasing in T_q for $T_q > 0$, the claimed truncation bounds follow. If $T^-_q = 0$, then $M^{\text{obs}}_q = 0$: the zero-total convention gives the lower endpoint zero, while the upper endpoint is one if $T^+_q > 0$ and zero otherwise, exactly as in Equation (35). Sorting the fragment-size lists computes both sharp endpoints in $O(m \log m)$ time. □

C.4 Multi-client Computational Boundary

For three clients, anchoring the first fragment index leaves two permutations $\varpi, \varsigma \in \text{Sym}(m)$ and the objective

$$\Phi_{\text{obs}} + \sum_{i=1}^{m} C_{i,\varpi(i),\varsigma(i)}. \tag{37}$$

For any finite tensor T, create unit-weight singleton fragments $\{a_i\}, \{b_j\}, \{c_k\}$ in one type/weight stratum, set $q = 3$, fix one type r, define $\psi_{3,r}(\{a_i, b_j, c_k\}, X_{\{a_i,b_j,c_k\}}) = T_{ijk}$, and set the test function to zero on every other triple. Each completed hyperedge has exactly one q-node subset, so its cost tensor is exactly T. The m^3 test-function table is an explicit polynomial-size part of the endpoint instance. For a three-dimensional-matching instance, choose $T_{ijk} = 0$ on an allowed triple and $T_{ijk} = 1$ otherwise. A completion with value zero exists if and only if the three-dimensional-matching instance has a perfect matching. The latter is NP-complete (Karp, 1972). Hence the lower endpoint is NP-hard for three clients. More clients contain this construction by adding singleton fragment lists and defining the $q = 3$ test to be zero on every triple involving an added client. Their permutations leave the objective unchanged. Together, all these prove hardness for general linear co-membership targets.

For a sparse binary-reward target, correspondence reduces to hypergraph matching. The following result separates the effects of client count and desired matching size.

Theorem 2 (Parameterized complexity of hyperedge matching)**.** *In one exact all-client stratum, let each of $K \geq 2$ clients have m distinct singleton fragments, with one common type and unit weight. Set $q = K$. The input lists favorable tuples $\mathcal{E}_+ \subseteq [m]^K$ and assigns the order-K query value one to them and zero to all other tuples. This sparse default-zero encoding includes all client-row lists, so Km is bounded by the input length. For permutations $\varpi_2, \ldots, \varpi_K$ of $[m]$, define*

$$\Phi_+(\mathcal{E}_+) = \max_{\varpi_2,\ldots,\varpi_K} \sum_{i=1}^{m} \mathbf{1}\{(i, \varpi_2(i), \ldots, \varpi_K(i)) \in \mathcal{E}_+\}. \tag{38}$$

For $1 \leq b \leq m$, deciding $\Phi_+ \geq b$ is polynomial-time solvable when $K = 2$ and NP-complete for every fixed $K \geq 3$. It is fixed-parameter tractable in Kb, and hence in b for fixed K. When K is

unbounded it is W[1]-complete parameterized by b. A direct randomized algorithm has one-sided error at most $\delta_{\mathrm{cc}} \in (0,1)$ and time $O((2e)^{Kb}\,\mathrm{poly}(L_{\mathrm{in}})\log(2/\delta_{\mathrm{cc}}))$, where L_{in} is the input length.

Proof. Treat client-row pairs (k,i) as vertices in K disjoint parts. A favorable tuple is a K-uniform hyperedge containing one vertex from each part. Rewarded tuples in a completion are pairwise disjoint. Conversely, any b disjoint favorable tuples extend to a full completion: each part has $m-b$ unused rows, which can be paired arbitrarily. Thus the endpoint decision is exactly the existence of a b-edge matching in the explicitly supplied K-partite hypergraph. For $K=2$ this is bipartite matching. For $K=3, b=m$ it is perfect three-dimensional matching (Karp, 1972). For larger fixed K, copy the first coordinate into each new coordinate. This preserves disjointness and proves NP-hardness. A list of b tuples is a polynomially checkable witness, establishing NP membership.

For fixed-parameter tractability, put $s = Kb$ and color every client-row independently and uniformly with s colors. The s distinct rows of any fixed witness all receive different colors with probability $s!/s^s \geq e^{-s}$. Discard tuples whose vertices repeat a color. Use a Boolean dynamic program over color subsets, starting from the empty subset and adjoining the K colors of any tuple disjoint from the current subset. Reaching all s colors certifies exactly b disjoint tuples. Disjoint color masks cannot share a row, so every reported witness is valid. One trial costs $O(2^s|\mathcal{E}_+|K + Km)$. Repeating $\lceil e^s \log(2/\delta_{\mathrm{cc}})\rceil$ times yields the claimed one-sided error and running time. Perfect-hash families derandomize color-coding to give deterministic fixed-parameter tractability (Alon et al., 1995).

For W[1] membership, form the conflict graph with one vertex per favorable tuple and an edge between tuples sharing any client-row. The desired matching is an independent set of size b, a parameter-preserving reduction to Independent Set (Downey & Fellows, 1995). For hardness, take an Independent Set instance $H = (U, F_H)$ with $N = |U|$ vertices and target b. Create one private client and one client for each edge of H, adding a harmless client if needed to make $K \geq 2$. Every client has $m = N$ rows. For each $v \in U$, list one favorable tuple t_v. Its private-client row is unique to v. At the client for edge $\{u,v\}$, tuples t_u, t_v share a row, and all other tuples use distinct rows different from that shared row. Exactly $N-1$ rows are used there, which is feasible. A harmless client assigns distinct rows to all tuples. Two favorable tuples share a row precisely when their vertices in H are adjacent. The construction is polynomial and preserves b. Therefore an independent set of size b exists precisely when the constructed endpoint is at least b, proving W[1]-hardness and completeness. □

Here candidate hyperedges have rank K. The fixed-rank and variable-rank classifications concern this sparse reward problem, including its supplied query function. They do not classify arbitrary dense costs or neural-network training. For any client count, separation across a designated anchor gives the following tractable exception.

Proposition 1 (Additive multi-client costs with one designated anchor)**.** *Consider an exact all-client stratum with m rows per client, one row from each compatible original hyperedge at every client, and independently feasible one-to-one matchings anchored at client 1. A fully additive tensor $C_{i_1,\dots,i_K} = \sum_k a^{(k)}_{i_k}$ is point identified. When the cost adds across the anchor–client pairs, $C_{i_1,\dots,i_K} = a_{i_1} + \sum_{k=2}^{K} C^{(k)}_{i_1 i_k}$, each sharp endpoint is the constant $\sum_i a_i$ plus the corresponding optima of $K-1$ independent bipartite assignments. Thus each endpoint costs $O(\sum_h (K-1)m_h^3)$ by solving one independent bipartite assignment for each anchor-client pair within every stratum.*

Proof of Proposition 1. Anchoring client 1 makes the feasible set $\mathrm{Sym}(m)^{K-1}$. For a fully additive cost,

$$\begin{aligned}\sum_{i=1}^{m} C_{i,\varpi_2(i),\dots,\varpi_K(i)} &= \sum_{i=1}^{m} a^{(1)}_i + \sum_{k=2}^{K}\sum_{i=1}^{m} a^{(k)}_{\varpi_k(i)} \\ &= \sum_{k=1}^{K}\sum_{j=1}^{m} a^{(k)}_j, \end{aligned} \tag{39}$$

because each ϖ_k is a permutation, proving identification. Anchor-additive costs instead give

$$\sum_{i=1}^{m} a_i + \sum_{k=2}^{K}\sum_{i=1}^{m} C^{(k)}_{i,\varpi_k(i)}. \tag{40}$$

Each summand has its own feasible permutation, so the total optimum is the sum of independent assignment optima. Their integrality proves sharpness and gives $O(\sum_h (K-1)m_h^3)$ time. Interactions among non-anchor indices would break this separation. □

C.5 Boolean-zeta Inversion and Partition Censoring

The assignment results describe uncertainty about how fragments fit together. We next justify the co-membership coordinates used to measure the information removed by partitioning. Before partitioning, the full collection determines aggregate support weights. After partitioning, each coordinate is either retained exactly or erased, under the observation assumptions stated in Appendix B.2.

Proposition 2 (Invertible Boolean-zeta representation and censoring)**.** *For every type $r \in \mathcal{R}$, the complete nonempty collection $M^{\mathcal{H}}_{|S|,r}(S)$ determines the aggregate support–type weights through*

$$w_r(A) = \sum_{B \supseteq A} (-1)^{|B|-|A|} M^{\mathcal{H}}_{|B|,r}(B). \tag{41}$$

If multiplicity, exact retained weight, and type are observed, then Equation (6) holds for every $q \geq c$.

Proof. Fix a type r and regard w_r as a function on the Boolean lattice 2^V, with zero mass on the empty relation. Equation (2) is its upper zeta transform: $M^{\mathcal{H}}_{|S|,r}(S) = \sum_{A \supseteq S} w_r(A)$. The Möbius function on the interval $[S, A]$ is $(-1)^{|A|-|S|}$, so Boolean-lattice Möbius inversion gives Equation (41). This inversion applies separately to each observed relation type.

For the mask, a copy of S contributed by relation instance e appears in a local fragment if and only if all nodes in S have the same client label. If they do, exactly one client contains that copy. Otherwise none does. Summing over all $A_e \supseteq S$ proves Equation (6). The complete and retained order-q masses therefore satisfy the two identities below for every observed type r and order q:

$$\begin{aligned} \|M^{\mathcal{H}}_{q,r}\|_1 &= \sum_{e: r_e = r} w_e \binom{|A_e|}{q}, \\ \|\widetilde{M}_{q,r}(\cdot, z)\|_1 &= \sum_{e: r_e = r} w_e \sum_{k=1}^{K} \binom{|A_e \cap V_k|}{q}. \end{aligned} \tag{42}$$

Their ratio is exactly the retained order-q co-membership mass in (7) for type r. □

C.6 Horvitz–Thompson Estimation

The masking identity immediately suggests inverse-probability weighting under a known random partition design. It also reveals the cost of this correction: coordinates that rarely remain within one client acquire large variance. The theorem below gives both statements for each coordinate and then sums their squared errors.

Theorem 3 (Known randomized assignment)**.** *Fix a latent hypergraph $\mathcal{H}$, an order $c \leq q \leq |V|$, a type $r \in \mathcal{R}$, and a node subset $S \in \binom{V}{q}$. Let $\mathsf{Z} : V \to [K]$ follow the declared randomized partition design and assume $\pi_S = \Pr\{\mathsf{Z}|_S \text{ is constant}\} > 0$. The coordinate estimator satisfies*

$$\widehat{M}_{q,r}(S) = \widetilde{M}_{q,r}(S, \mathsf{Z})/\pi_S, \qquad \mathbb{E}_{\mathsf{Z}} \widehat{M}_{q,r}(S) = M^{\mathcal{H}}_{q,r}(S), \tag{43}$$

$$\mathrm{Var}_{\mathsf{Z}}[\widehat{M}_{q,r}(S)] = M^{\mathcal{H}}_{q,r}(S)^2(\pi_S^{-1} - 1). \tag{44}$$

If every q-set has positive inclusion probability, the mean squared Frobenius error is

$$\mathbb{E}_{\mathsf{Z}} \|\widehat{M}_{q,r} - M^{\mathcal{H}}_{q,r}\|_F^2 = \sum_{S \in \binom{V}{q}} M^{\mathcal{H}}_{q,r}(S)^2(\pi_S^{-1} - 1). \tag{45}$$

Under uniform independent assignment, the relative coordinate variance of a nonzero coordinate is $K^{q-1} - 1$. If $\pi_S = 0$, no estimator measurable from the censored order-q co-membership array is design-unbiased for that coordinate over all type-r hypergraphs with size-q relation supports.

Proof. Let $I_S(\mathsf{Z}) = \mathbf{1}\{\mathsf{Z}|_S \text{ is constant}\}$. Then $I_S(\mathsf{Z}) \sim \text{Bernoulli}(\pi_S)$ and $\widetilde{M}_{q,r}(S,\mathsf{Z}) = I_S(\mathsf{Z})M^{\mathcal{H}}_{q,r}(S)$, which gives (43)–(44). Linearity of expectation applied to the sum of squared coordinate errors gives (45). Dependence among the indicators is allowed. Uniform assignment has $\pi_S = K(1/K)^q = K^{1-q}$. For $\pi_S = 0$, compare a q-uniform hypergraph with and without a weight-a edge on S. Their censored order-q co-membership arrays agree almost surely but the target coordinate differs by a. □

C.7 Observational Equivalence and a Two-point Bound

Randomized-design unbiasedness concerns repeated releases of a fixed structure. With one release held fixed, the question is different: can two compatible structures have different target values? If they can, any estimator based on that release has the same output law under both structures and must incur error on at least one. We first state the resulting bound for the learner's coarsened release, then prove the underlying two-point argument.

Corollary 2 (Identification-to-learning bridge under release coarsening)**.** *Let $y_{\mathrm{L}} = \rho(y_{\mathrm{A}})$ and hold the released training labels Y^{tr} fixed across all candidate structures. For the same fixed z and Γ, define $\mathcal{F}_{\mathrm{L}}(y_{\mathrm{L}},\Gamma) = \{\mathcal{H} \in \mathcal{M}_\Gamma : \rho(O_{z,\Gamma}(\mathcal{H})) = y_{\mathrm{L}}\}$. Deterministic coarsening gives $\mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma) \subseteq \mathcal{F}_{\mathrm{L}}(y_{\mathrm{L}},\Gamma)$. Let $\phi : \mathcal{M}_\Gamma \to \mathbb{R}$ be a scalar structural functional whose attainable endpoints on $\mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma)$ are ϕ_- and ϕ_+. The minimax absolute risk over randomized rules $\widehat{\phi} = a(y_{\mathrm{L}}, Y^{\mathrm{tr}}, \xi)$, with algorithmic randomness ξ independent of $\mathcal{H}$, satisfies this bound over the compatible class:*

$$\inf_a \sup_{\mathcal{H}\in\mathcal{F}_{\mathrm{L}}(y_{\mathrm{L}},\Gamma)} \mathbb{E}_\xi|\widehat{\phi} - \phi(\mathcal{H})| \geq (\phi_+ - \phi_-)/2. \tag{46}$$

The proof uses the two endpoint structures, which remain indistinguishable after coarsening and therefore induce the same conditional output law.

Proposition 3 (Fixed-release observational equivalence)**.** *Fix z, Γ, and a metric space (Θ, d). A functional $\phi : \mathcal{M}_\Gamma \to \Theta$ is point identified from the deterministic analyst release if and only if it is constant on every nonempty class $\mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma)$. If $\mathcal{H}_0, \mathcal{H}_1 \in \mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma)$ and $d(\phi(\mathcal{H}_0), \phi(\mathcal{H}_1)) = \Delta$, every randomized estimator $\widehat{\phi} = a(y_{\mathrm{A}}, \xi)$ with $\xi \sim \nu$ independent of the latent hypergraph obeys*

$$\max_{b\in\{0,1\}} \mathbb{E}_\xi d(\widehat{\phi}, \phi(\mathcal{H}_b)) \geq \Delta/2. \tag{47}$$

Proof. If $\phi = g \circ O_{z,\Gamma}$, equal analyst releases imply equal values of ϕ, so ϕ is constant on each compatible class. Conversely, if ϕ is constant on every class, define $g(y_{\mathrm{A}})$ as that common value for any $\mathcal{H} \in \mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma)$. Constancy makes this definition independent of the representative.

For the lower bound, $\widehat{\phi} = a(y_{\mathrm{A}}, \xi)$ has the same distribution at both compatible structures. Pointwise, the triangle inequality yields the following deterministic relation between the two endpoint errors:

$$\Delta \leq d(\widehat{\phi}, \phi(\mathcal{H}_0)) + d(\widehat{\phi}, \phi(\mathcal{H}_1)). \tag{48}$$

Taking expectation over the common law ν of ξ shows that one endpoint risk reaches at least $\Delta/2$. □

Proof of Corollary 2. If $\mathcal{H} \in \mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}},\Gamma)$, applying ρ to its analyst release gives y_{L}, hence $\mathcal{H} \in \mathcal{F}_{\mathrm{L}}(y_{\mathrm{L}},\Gamma)$. Theorem 1 supplies compatible structures attaining ϕ_- and ϕ_+. A rule measurable from $\sigma(Y_{\mathrm{L}}, Y^{\mathrm{tr}}, \xi)$ has one conditional output law under both endpoint structures because their training labels are held fixed. Their two absolute errors sum to at least $\phi_+ - \phi_-$, so one endpoint risk reaches $(\phi_+ - \phi_-)/2$. This holds for every measurable rule, and taking the infimum gives Equation (46). □

Corollary 3 (Matching under hyperedge aliasing)**.** *In the two-client construction of Section 3.2, assume each side's m cross-status fragment rows are pairwise distinct and nonempty, share uninformative type and weight marks, and have no shared cross-client hyperedge identifier. A latent matching $\varpi \sim \text{Unif}(\text{Sym}(m))$ is recovered completely with probability at most $1/m!$. An estimator returning s proposed matches has at most s/m correct matches in expectation.*

Proof. The observation is invariant under all $m!$ permutations. Conditional on it, each full permutation is equiprobable. A fixed proposed pair is correct with probability $1/m$. Linearity of expectation yields s/m without an independence assumption across all returned proposals. □

C.8 Released-data Factorization and Observable Spectral Approximation

We establish that FedASTRA's complete trajectory factors through the learner release, then characterize the stability and expressiveness of its observable heat basis.

Proposition 4 (Permutation invariance and heat stability)**.** *For every relation-column permutation* Π*,* $P_k(B_k\Pi) = P_k(B_k)$ *and* $W_{k,\theta}$ *is unchanged. On the nonisolated block,* $P_k^{\mathrm{sym}} = D_{v,k}^{1/2} P_k D_{v,k}^{-1/2}$ *has spectrum in* $[0,1]$*. Hence* $\|e^{t(P_k^{\mathrm{sym}}-I)}\|_2 \le 1$ *for* $t \ge 0$*, and two such symmetric representatives on the same nodes satisfy*

$$\|e^{t(P^{\mathrm{sym}}-I)} - e^{t(\widetilde{P}^{\mathrm{sym}}-I)}\|_2 \le t\|P^{\mathrm{sym}} - \widetilde{P}^{\mathrm{sym}}\|_2. \tag{49}$$

Proof. For $B' = B\Pi$, $D'_e = \Pi^\top D_e \Pi$, while D_v and I_0 are fixed. Thus

$$B'(D'_e)^{-1}(B')^\top = B\Pi(\Pi^\top D_e \Pi)^{-1}\Pi^\top B^\top = BD_e^{-1}B^\top, \tag{50}$$

proving invariance of P and all heat features. On the nonisolated block,

$$P^{\mathrm{sym}} = D_v^{-1/2} B D_e^{-1} B^\top D_v^{-1/2} = GG^\top, \qquad G = D_v^{-1/2} B D_e^{-1/2}. \tag{51}$$

Positive semidefiniteness and similarity to the row-stochastic P give spectrum in $[0,1]$ and contraction. Isolated rows are absorbing. Duhamel's identity gives

$$e^{t(P^{\mathrm{sym}}-I)} - e^{t(\widetilde{P}^{\mathrm{sym}}-I)} = \int_0^t e^{(t-s)(P^{\mathrm{sym}}-I)}(P^{\mathrm{sym}} - \widetilde{P}^{\mathrm{sym}}) e^{s(\widetilde{P}^{\mathrm{sym}}-I)}\, ds, \tag{52}$$

whose norm is bounded by Equation (49). □

For one release, contraction in symmetric coordinates is equivalent to contraction in $\|x\|_{D_v} = \|D_v^{1/2}x\|_2$; the perturbation bound compares symmetric representatives. In the two-client model, joining F_{1i}^h to $F_{2,\varpi_h(i)}^h$ gives

$$\{\mathcal{H}_{\boldsymbol{\varpi}} : \boldsymbol{\varpi} \in \textstyle\prod_h \mathrm{Sym}(m_h)\} \subseteq \mathcal{F}_{\mathrm{A}}(y_{\mathrm{A}}, \Gamma) \subseteq \mathcal{F}_{\mathrm{L}}(\rho(y_{\mathrm{A}}), \Gamma). \tag{53}$$

These matchings range over Equation (10) while preserving all local fragment multisets and hence the learner release.

Theorem 4 (Factorization of the complete federated learning trajectory)**.** *Let* $\mathcal{P}$ *fix the task split, training and validation labels, participation schedule, training-count aggregation weights, checkpoint rule, and hyperparameters; let* ξ *collect initialization, masking, minibatch, and optimizer randomness. Under exact arithmetic, for any client count and horizon* T*, the transcript* $\mathcal{T}_T$ *of all broadcasts, online, EMA and optimizer states, uploads, selected checkpoint, and final logits admits a measurable map*

$$\mathcal{T}_T(\mathcal{H}, \xi) = F_T\big(\rho(\mathcal{O}_{z,\Gamma}(\mathcal{H})), \mathcal{P}, \xi\big)\,. \tag{54}$$

Thus compatible completions have identical transcripts under coupled ξ*, and identical conditional laws under a common* $\mathcal{H}$*-independent distribution for* ξ*.*

Proof. Equation (50) makes P_k a measurable function of $[B_k]_{\mathrm{col}}$. Initialization, masked attributes, heat features, losses and gradients therefore depend only on $(Y_{\mathrm{L}}, \mathcal{P}, \xi)$ and current states. An optimizer step and the EMA recursion preserve this dependence. Induction over local steps gives the uploads; training-count-weighted aggregation gives the next broadcast. Induction over rounds, followed by the fixed validation rule and final forward pass, proves the factorization. Coupling, then integration over ξ, gives the two equalities. □

Corollary 4 (Downstream invariance on learner compatibility classes)**.** *Under the preceding assumptions, equal learner releases imply, throughout training,*

$$\theta^u(\mathcal{H}) = \theta^u(\mathcal{H}'), \qquad \mathrm{logit}_{k,v}^u(\mathcal{H}) = \mathrm{logit}_{k,v}^u(\mathcal{H}'). \tag{55}$$

This includes intermediate local and server states, validation selection, and final predictions, with equal conditional laws after averaging over ξ*.*

Proof. Apply Theorem 4; fixed evaluation labels also give identical Accuracy and Macro-F1. □

Corollary 5 (Density within observable spectral filters)**.** *For symmetric $P^{\mathrm{sym}}(Y_{\mathrm{L}})$ with spectrum in $[0,1]$, every $f \in C([0,1])$, $\delta > 0$, and $\Delta > \epsilon_{\mathrm{ord}}$ admit a finite M_f and coefficients depending only on (f, δ, Δ) such that*

$$\left\| f(P^{\mathrm{sym}}) - a_0 I - \sum_{b=1}^{M_f} a_b e^{b\Delta(P^{\mathrm{sym}} - I)} \right\|_2 < \delta. \tag{56}$$

The grid $t_b = b\Delta$ is feasible in Equation (20).

Proof. The change of variable $x = e^{\Delta(\lambda-1)}$ maps $[0,1]$ homeomorphically to $[e^{-\Delta}, 1]$. Polynomial approximation of $f(1+\log(x)/\Delta)$ gives $p(x) = \sum_{b=0}^{M_f} a_b x^b$ with uniform error below δ. The spectral theorem yields Equation (56); choosing each softplus increment as $\Delta - \epsilon_{\mathrm{ord}} > 0$ realizes the grid. □

Corollary 6 (Information-boundary compliance and observable spectral approximation)**.** *On a positive-degree client block, define $f(P) = D_v^{-1/2} f(P^{\mathrm{sym}}) D_v^{1/2}$ and $G_{M_f}(P) = a_0 I + \sum_{b=1}^{M_f} a_b e^{b\Delta(P-I)}$ using the preceding coefficients. For every conformable U,*

$$\|(f(P) - G_{M_f}(P))U\|_{D_v,F} \le \delta \|U\|_{D_v,F}, \qquad \|U\|_{D_v,F} = \|D_v^{1/2} U\|_F. \tag{57}$$

Both these transforms and the complete training transcript are shared across the learner compatibility class, for fixed $\mathcal{P}$ and ξ.

Proof. Conjugate Equation (56) by $D_v^{1/2}$ and use $\|AU\|_F \le \|A\|_2 \|U\|_F$. Invariance follows from Theorem 4. □

Density concerns the union of finite basis widths. It gives neither fixed-R universality nor a prediction or optimization guarantee for a fitted classifier.

Certified evaluation. All times share the recurrence $U, PU, P^2U, \ldots$. For the auxiliary analytic variable $N_t^{\mathrm{aux}} \sim \mathrm{Poisson}(t)$, deterministic tail-collapsed evaluation is

$$\widehat{\mathsf{K}}_{t,L} U = \sum_{\ell=0}^{L-1} e^{-t} \frac{t^\ell}{\ell!} P^\ell U + \left(1 - \sum_{\ell=0}^{L-1} e^{-t} \frac{t^\ell}{\ell!}\right) P^L U. \tag{58}$$

For active times $\mathcal{T}_{\mathrm{heat}} = \{t_{bj} : b \in [R], j \in [d]\}$, choose the smallest integer L_{tail} satisfying

$$\max_{t \in \mathcal{T}_{\mathrm{heat}}} \Pr(N_t^{\mathrm{aux}} > L_{\mathrm{tail}}) \le \varepsilon_{\mathrm{tail}}, \qquad \max_{t \in \mathcal{T}_{\mathrm{heat}}} \{\Pr(N_t^{\mathrm{aux}} = L_{\mathrm{tail}}) + 2\Pr(N_t^{\mathrm{aux}} > L_{\mathrm{tail}})\} \le \varepsilon_{\mathrm{tail}}. \tag{59}$$

These bound forward and time-derivative coefficient masses. We use $\varepsilon_{\mathrm{tail}} = 10^{-6}$, float64 log-space probabilities, and record both bounds and L_{tail} for each validation-selected native checkpoint. Online and EMA branches use their own times. At $R = 0$, bypass the recurrence and set $L_{\mathrm{tail}} = 0$; configurations requiring $L_{\mathrm{tail}} > 4096$ are refused. Floating-point propagation and accumulation errors are separate from this coefficient certificate.

C.9 Observable Diffusion and Predictive Dynamics

Diffusion time evolves features at fixed incidence and encoder; local training steps evolve the encoder and its EMA target. The following results distinguish these two processes.

Theorem 5 (Continuous-time diffusion under observational equivalence)**.** *Fix y_{L}, client k, and $U_k^0 \in \mathbb{R}^{|V_k| \times d}$. Put $d_{k,v} = (D_{v,k})_{vv}$ and $V_{k,+} = \{v : d_{k,v} > 0\}$. Let $D_{k,+}$ be $D_{v,k}$ restricted to these nodes, and use $+$ likewise for other restrictions. For $t \ge 0$, the trajectory*

$$U_k(t) = e^{t(P_k - I)} U_k^0 \tag{60}$$

uniquely solves $\dot{U}_k = (P_k - I)U_k$, $U_k(0) = U_k^0$. For the incidence-induced components $\mathcal{G}_k$ on $V_{k,+}$, define

$$(\mathsf{Q}_{k,+} U_{k,+}^0)_v = \frac{\sum_{u \in C} d_{k,u} U_{k,u}^0}{\sum_{u \in C} d_{k,u}}, \qquad v \in C \in \mathcal{G}_k. \tag{61}$$

Writing $A = U_{k,+}(t) - \mathsf{Q}_{k,+}U^0_{k,+}$ *and* $\|A\|^2_{D_{k,+},F} = \operatorname{tr}(A^\top D_{k,+}A)$ *gives*

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|A\|^2_{D_{k,+},F} &= -\operatorname{tr}[A^\top D_{k,+}(I - P_{k,+})A] \\ &= -\frac{1}{2}\sum_e \frac{1}{d_{e,k}} \sum_{u,v\in e} \|U_{k,u}(t) - U_{k,v}(t)\|_2^2 \le 0,\end{aligned} \tag{62}$$

where e *ranges over retained columns and* $d_{e,k} = (D_{e,k})_{ee}$. *If the nonstationary subspace is nonzero, let*

$$\lambda_{\mathrm{gap},k} = \min\{1 - \lambda : \lambda \in \operatorname{spec}(P^{\mathrm{sym}}_{k,+}),\ \lambda < 1\} > 0, \quad P^{\mathrm{sym}}_{k,+} = D^{1/2}_{k,+}P_{k,+}D^{-1/2}_{k,+}. \tag{63}$$

Then

$$\|U_{k,+}(t) - \mathsf{Q}_{k,+}U^0_{k,+}\|_{D_{k,+},F} \le e^{-\lambda_{\mathrm{gap},k}t}\|U^0_{k,+} - \mathsf{Q}_{k,+}U^0_{k,+}\|_{D_{k,+},F}. \tag{64}$$

A zero nonstationary subspace gives zero residual. Isolated nodes stay fixed, and compatible completions with equal initial features have identical trajectories.

Proof. The matrix exponential uniquely solves the finite linear system, and column invariance gives the common trajectory. On $V_{k,+}$, $D_{k,+}P_{k,+} = B_{k,+}D^{-1}_{e,k}B^\top_{k,+}$ is symmetric and $P_{k,+}\mathsf{Q}_{k,+} = \mathsf{Q}_{k,+}P_{k,+} = \mathsf{Q}_{k,+}$. Differentiating the energy and expanding the quadratic form gives

$$\operatorname{tr}[A^\top D_{k,+}(I - P_{k,+})A] = \frac{1}{2}\sum_e \frac{1}{d_{e,k}} \sum_{u,v\in e} \|A_u - A_v\|_2^2. \tag{65}$$

Its null space consists of componentwise constant matrices. In symmetric coordinates, $\mathsf{Q}_{k,+}$ is the orthogonal eigenvalue-one projection; every other heat multiplier is at most $e^{-\lambda_{\mathrm{gap},k}t}$. The spectral theorem proves convergence. The isolated block is the identity. □

Corollary 7 (Ordered sampling of the diffusion trajectory)**.** *Equation* (20) *gives, for each coordinate* j,

$$0 = t_{0j} < t_{1j} < \cdots < t_{Rj}, \qquad W^{(j)}_{k,\theta} = [U_{k,:j}(t_{0j}), \ldots, U_{k,:j}(t_{Rj})], \tag{66}$$

where $W^{(j)}_{k,\theta}$ *selects columns* $j, d + j, \ldots, Rd + j$ *and* $U_{k,:j}(t) = \mathsf{K}_k(t)U^0_{k,:j}$. *The feasible grid* $t_b = b\Delta$, $\Delta > \epsilon_{\mathrm{ord}}$, *yields the dense algebra*

$$\operatorname{span}\{1, e^{b\Delta(\lambda-1)} : b \ge 1\} \subset C([0,1]). \tag{67}$$

Density over increasing finite widths follows from Corollary 5.

Masked prediction and target dynamics. Within round u, let s index local steps, o denote optimizer state, $Y_{\mathrm{L},k} = (X_k, [B_k]_{\mathrm{col}})$, and $\xi_{k,u,s}$ contain masks, minibatches, and optimizer randomness. The update has the form

$$(\theta_{k,u,s}, o_{k,u,s}) = \Psi_{k,u,s}\big(\theta_{k,u,s-1}, o_{k,u,s-1}, \bar\vartheta_{k,u,s-1}, Y_{\mathrm{L},k}, \mathcal{P}, \xi_{k,u,s}\big)\,. \tag{68}$$

With $\vartheta_{k,u,s} = \mathrm{ctx}(\theta_{k,u,s})$ containing the encoder and raw time increments, the target satisfies

$$\bar\vartheta_{k,u,s} = \mu\bar\vartheta_{k,u,s-1} + (1-\mu)\vartheta_{k,u,s} = \mu^s\bar\vartheta_{k,u,0} + (1-\mu)\sum_{a=1}^{s}\mu^{s-a}\vartheta_{k,u,a}. \tag{69}$$

For $e_{k,u,s} = \vartheta_{k,u,s} - \bar\vartheta_{k,u,s}$ and $\Delta_{k,u,s} = \vartheta_{k,u,s} - \vartheta_{k,u,s-1}$, subtraction and iteration give

$$e_{k,u,s} = \mu e_{k,u,s-1} + \mu\Delta_{k,u,s}, \qquad \|e_{k,u,s}\| \le \mu^s\|e_{k,u,0}\| + \mu\sum_{a=1}^{s}\mu^{s-a}\|\Delta_{k,u,a}\|. \tag{70}$$

Thus a frozen online context has lag contraction factor μ. The input masking moments in Equation (110) do not imply unbiased nonlinear encodings or gradients.

Lemma 1 (Observation invariance throughout training)**.** *For fixed* $\mathcal{P}$ *and coupled* ξ, *the masked loss, online, optimizer and EMA states, uploads, broadcasts, selected checkpoint, and final logits are measurable in* $(Y_{\mathrm{L}}, \mathcal{P}, \xi)$ *and agree across its compatibility class.*

Proof. This is Theorem 4 applied to the local recursions above and the aggregation

$$\theta_k^{u+1} = \mathsf{U}_{k,u}(\theta^u, Y_{\mathrm{L},k}, \mathcal{P}, \xi_{k,u}), \qquad \theta^{u+1} = \frac{\sum_{k\in C_u} n_k^{\mathrm{tr}} \theta_k^{u+1}}{\sum_{k\in C_u} n_k^{\mathrm{tr}}}. \tag{71}$$

□

Diffusion dissipation and EMA lag concern distinct dynamics; neither is a Lyapunov function for joint nonlinear training. Observation invariance also supplies no formal privacy guarantee for communicated updates.

C.10 Downstream Risk with Additional Information

We compare fitted predictors under one joint law of data, training randomness, and held-out query nodes. Conditional averaging and information comparison are classical (Blackwell, 1953; Chen et al., 2020); here they connect prediction variation to the assignment geometry of Theorem 1.

Information and fitted risk. For a queried node with class $y \in [C]$, $C \geq 2$, let O_k contain client features and anonymous incidence, the query, the parameter-only transcript through validation selection, and exogenous randomness. Held-out labels, test outcomes, auxiliary messages, and states trained using those messages are excluded. Set $\mathcal{S}_0 = \sigma(O_k)$; the FedASTRA predictor F is $\mathcal{S}_0$-measurable. This permits federated parameter exchange while keeping structural computation local. Let A_{aux} include arbitrary additional information and any training state needed for an enriched predictor G, measurable in $\mathcal{S}_1 = \mathcal{S}_0 \vee \sigma(A_{\mathrm{aux}})$. The predictors may have different learned parameters. A conditional structural audit may fix analyst releases and cost arrays in the base, but such conditioning does not grant deployment access to them.

Define, with one-hot label $\boldsymbol{e}_y$,

$$p_0 = \mathbb{E}[\boldsymbol{e}_y \mid \mathcal{S}_0], \qquad p_1 = \mathbb{E}[\boldsymbol{e}_y \mid \mathcal{S}_1], \qquad J_C = I_C - \mathbf{1}\mathbf{1}^\top/C, \tag{72}$$

$$s_F = J_C \operatorname{logit}_F, \qquad s_G = J_C \operatorname{logit}_G, \qquad \bar{s}_G = \mathbb{E}[s_G \mid \mathcal{S}_0]. \tag{73}$$

Assume square-integrable logits. Centering preserves softmax. Write $\operatorname{lse}(s) = \log \sum_c e^{s_c}$, $\mathcal{R}_{\mathrm{CE}}(s) = \mathbb{E}[\operatorname{lse}(s) - s_y]$, and $\mathcal{R}_{\log}(q) = \mathbb{E}[-\log q_y]$ with the usual convention at zero. Conditional log risk equals conditional entropy plus expected KL divergence. Consequently,

$$\inf_{q\ \mathcal{S}_0\text{-measurable}} \mathcal{R}_{\log}(q) - \inf_{q\ \mathcal{S}_1\text{-measurable}} \mathcal{R}_{\log}(q) = \mathcal{I}_{\mathrm{add}} := \mathbb{E} D_{\mathrm{KL}}(p_1 \| p_0) \geq 0. \tag{74}$$

Indeed, the minimizers are p_0, p_1 and the tower property identifies the entropy difference. For fitted probability rules of finite log risk, put $\mathcal{E}_F = \mathbb{E} D_{\mathrm{KL}}(p_0 \| q_F)$ and $\mathcal{E}_G = \mathbb{E} D_{\mathrm{KL}}(p_1 \| q_G)$. Subtracting $\mathcal{R}_{\log}(q_F) = H(y \mid \mathcal{S}_0) + \mathcal{E}_F$ from $\mathcal{R}_{\log}(q_G) = H(y \mid \mathcal{S}_1) + \mathcal{E}_G$ proves Equation (16). Thus F strictly improves on G exactly when $\mathcal{E}_G - \mathcal{E}_F > \mathcal{I}_{\mathrm{add}}$.

Theorem 6 (Cross-entropy comparison with arbitrary additional information)**.** *In the model above, define*

$$V_{\mathrm{aux}} = \mathbb{E}\|s_G - \bar{s}_G\|_2^2, \qquad \Delta_{\mathrm{label}} = \mathbb{E}\|p_1 - p_0\|_2^2, \tag{75}$$

$$\delta_F = \mathcal{R}_{\mathrm{CE}}(s_F) - \mathcal{R}_{\mathrm{CE}}(\bar{s}_G), \qquad C_{\mathrm{label}} = \mathbb{E}\langle p_1 - p_0, s_G - \bar{s}_G \rangle. \tag{76}$$

For $D_{\mathrm{lse}}(a, b) = \operatorname{lse}(a) - \operatorname{lse}(b) - \langle \nabla \operatorname{lse}(b), a - b \rangle$,

$$\mathcal{R}_{\mathrm{CE}}(s_G) - \mathcal{R}_{\mathrm{CE}}(s_F) = \mathbb{E} D_{\mathrm{lse}}(s_G, \bar{s}_G) - C_{\mathrm{label}} - \delta_F. \tag{77}$$

If $\|s_G\|_\infty \leq B_{\mathrm{logit}} < \infty$ *almost surely, then*

$$\mathcal{R}_{\mathrm{CE}}(s_G) - \mathcal{R}_{\mathrm{CE}}(s_F) \geq \frac{e^{-2B_{\mathrm{logit}}}}{2C} V_{\mathrm{aux}} - \sqrt{\Delta_{\mathrm{label}} V_{\mathrm{aux}}} - \delta_F. \tag{78}$$

A positive right side gives strictly smaller risk for F*. Under label sufficiency* $p_1 = p_0$*, the alignment term* C_{label} *vanishes.*

Proof. Put $r = s_G - \bar{s}_G$, so $\mathbb{E}[r \mid \mathcal{S}_0] = 0$. The Bregman linear term has zero expectation, and

$$\mathbb{E}\langle \boldsymbol{e}_y, r\rangle = \mathbb{E}\langle p_1, r\rangle = \mathbb{E}\langle p_1 - p_0, r\rangle = C_{\text{label}}. \tag{79}$$

Subtracting risks proves Equation (77); in particular, taking $s_F = \bar{s}_G$ (so $\delta_F = 0$) gives

$$\mathcal{R}_{\text{CE}}(s_G) - \mathcal{R}_{\text{CE}}(\bar{s}_G) = \mathbb{E}D_{\text{lse}}(s_G, \bar{s}_G) - C_{\text{label}}.$$

For $p = \text{softmax}(s)$ with $\|s\|_\infty \le B_{\text{logit}}$, $p_c \ge e^{-2B_{\text{logit}}}/C$. For $v \perp \mathbf{1}$ and $a = p^\top v$,

$$v^\top \nabla^2 \,\text{lse}(s) v = \sum_c p_c (v_c - a)^2 \ge \frac{e^{-2B_{\text{logit}}}}{C} \sum_c (v_c - a)^2 \ge \frac{e^{-2B_{\text{logit}}}}{C} \|v\|_2^2. \tag{80}$$

The segment from $\bar{s}_G$ to s_G remains bounded and centered. Integral Taylor remainder gives $D_{\text{lse}}(s_G, \bar{s}_G) \ge e^{-2B_{\text{logit}}} \|r\|_2^2/(2C)$, proving the bound and Equation (17). □

The Bregman term charges variation around the base-conditional mean; C_{label} credits alignment with new label signal, and the signed δ_F retains fitting quality. If A_{aux} is base-measurable, both V_{aux} and Δ_{label} vanish.

Proposition 5 (Assignment geometry of auxiliary prediction variation)**.** *Condition on base information fixing the cost arrays and intercepts. In the exact strata of Theorem 1, declare independent uniform candidate permutations $\varpi_h \in \text{Sym}(m_h)$ and affine logits*

$$[\text{logit}_G]_c = b_c + \sum_h \sum_{i=1}^{m_h} [C_c^h]_{i,\varpi_h(i)}, \qquad c \in [C]. \tag{81}$$

Set $\bar{C}^h = C^{-1} \sum_c C_c^h$ and $\widetilde{C}_c^h = J_{m_h}(C_c^h - \bar{C}^h)J_{m_h}$. Then

$$\mathbb{E}[\|s_G - \bar{s}_G\|_2^2 \mid \mathcal{S}_0] = \sum_{h:m_h \ge 2} \frac{1}{m_h - 1} \sum_{c=1}^{C} \|\widetilde{C}_c^h\|_F^2. \tag{82}$$

Averaging gives V_{aux}. A scalar cost C^h has variance $\|J_{m_h} C^h J_{m_h}\|_F^2/(m_h - 1)$ for $m_h \ge 2$; singleton strata have zero variance.

Proof. Double centering a fixed cost C_* changes assignment sums by a constant. For $D_* = J_m C_* J_m$ and uniform ϖ, $S = \sum_i [D_*]_{i,\varpi(i)}$ has zero mean. Equal-row terms in $\mathbb{E}S^2$ contribute $\|D_*\|_F^2/m$; distinct-row terms give

$$\frac{1}{m(m-1)} \sum_{i \ne i'} \sum_{j \ne j'} [D_*]_{ij}[D_*]_{i'j'} = \frac{\|D_*\|_F^2}{m(m-1)}. \tag{83}$$

The equality uses $\sum_{i' \ne i, j' \ne j} [D_*]_{i'j'} = [D_*]_{ij}$, from zero row and column sums. Adding gives $\text{Var}(S) = \|D_*\|_F^2/(m-1)$. Class centering replaces C_c^h by $C_c^h - \bar{C}^h$; summing the scalar identities over classes and independent strata proves the claim. □

Thus the same interaction detects nonidentification and measures variation under the declared matching law. A scalar assignment interval of width w_* gives variance at most $w_*^2/4$. Uniform endpoint bounds for $C_c^h - \bar{C}^h$, together with uniformly bounded centered intercepts, can supply B_{logit} across base realizations. Class centering is essential because common logit shifts do not affect predictions. Nonlinear auxiliary rules remain covered by Theorem 6.

Corollary 8 (A conditional advantage for the fitted FedASTRA predictor)**.** *Under the bounded-logit assumptions above, let F_R be the fitted parameter-only FedASTRA predictor with R positive heat channels. If $\delta_{F_R} \le \varepsilon_R$ for a stated $\varepsilon_R \ge 0$, then*

$$\frac{e^{-2B_{\text{logit}}}}{2C} V_{\text{aux}} - \sqrt{\Delta_{\text{label}} V_{\text{aux}}} > \varepsilon_R \quad \Longrightarrow \quad \mathcal{R}_{\text{CE}}(s_{F_R}) < \mathcal{R}_{\text{CE}}(s_G). \tag{84}$$

For a conditionally label-independent candidate channel, the square-root term is zero; affine assignment logits have the variance in Proposition 5.

Proof. Transcript factorization makes F_R base-measurable. Substitute the residual bound into Equation (78); conditional label independence gives $p_1 = p_0$. □

Under label sufficiency $p_1 = p_0$, equivalently $y \perp A_{\text{aux}} \mid \mathcal{S}_0$, the two information levels have the same conditional entropy. Hence, for $p_{F_R} = \text{softmax}(s_{F_R})$ and $p_G = \text{softmax}(s_G)$,

$$\mathcal{R}_{\log}(p_G) - \mathcal{R}_{\log}(p_{F_R}) = \mathbb{E}D_{\text{KL}}(p_0 \| p_G) - \mathbb{E}D_{\text{KL}}(p_0 \| p_{F_R}). \tag{85}$$

If the fixed-R class approximates p_0 arbitrarily well in expected KL, its infimum equals both information-specific Bayes risks. A fitted error at most ε_R strictly improves on a specified G with larger KL error. Label sufficiency, fixed-model approximation, and fitting are separate assumptions; spectral density and correspondence invariance imply none of them.

Proposition 6 (Target direction of the masked alignment component)**.** *Freeze the EMA context $\bar{\vartheta}_k$ for one local update. Let O_{mask} be the predictor's masked representation and U_{tar} its full-view, layer-normalized and unit-normalized target, with $\|U_{\text{tar}}\|_2 = 1$ almost surely. Then*

$$\inf_{u:\, \|u(O_{\text{mask}})\|_2 = 1} \mathbb{E}[1 - \langle u(O_{\text{mask}}), U_{\text{tar}} \rangle] = 1 - \mathbb{E}\|\mathbb{E}[U_{\text{tar}} \mid O_{\text{mask}}]\|_2. \tag{86}$$

For $m = \mathbb{E}[U_{\text{tar}} \mid O_{\text{mask}}] \neq 0$, the minimizing direction is $m/\|m\|_2$; when $m = 0$, every unit direction is optimal.

Proof. Conditional expected alignment is $\langle u, m \rangle \leq \|m\|_2$ by Cauchy–Schwarz, with equality in the stated directions. Integrate the optimum. □

This concerns the frozen-target alignment component with unrestricted directions. Anti-collapse regularization, two-view agreement, moving targets, and finite predictors remain in the full objective (112) and its optimization analysis.

Classification metrics. For base-measurable $p_F = \text{softmax}(s_F)$, let $h_F = \arg\max_c p_{F,c}$ and $h_0 = \arg\max_c p_{0,c}$ with a fixed tie rule. Comparing their selected coordinates and applying Pinsker and Cauchy–Schwarz bounds classification regret by $\mathbb{E}\|p_0 - p_F\|_1 \leq \sqrt{2\mathbb{E}D_{\text{KL}}(p_0 \| p_F)}$. This is regret against the base Bayes rule; a surrogate-risk ordering alone need not order two fitted classifiers' accuracies. If additionally $p_1 = p_0$, the top-two conditional probability gap is at least $\kappa_{\text{cls}} > 0$, and $h_{F_R} = h_0$ almost surely, then $\text{Acc}(F_R) - \text{Acc}(G) \geq \kappa_{\text{cls}} \Pr(h_G \neq h_0)$ by conditioning on $\mathcal{S}_1$. Macro-F1 also depends on classwise confusion probabilities: increased true positives and decreased false positives and false negatives suffice, whereas cross-entropy ordering alone does not determine it. Both metrics use the same held-out evaluation protocol.

C.11 Randomized Recovery from Diffusion Features and Local Prediction Risk

Heat features retain permitted local information when they recover the diffusion operator and its pairwise projection, and that projection determines the observed incidence. These are separate conditions, independent of unknown correspondence across clients. Incidence is binary; column multiplicities are retained unless simplicity is assumed.

Lemma 2 (Projection and recovery from diffusion features)**.** *Let B have n rows and nonempty columns, and let P be Equation (18), including absorbing isolated nodes. For $i \neq j$, $P_{ij} > 0$ exactly when a column contains both nodes, so its off-diagonal support determines the pairwise projection $G(B)$. Fix $\Delta > 0$, $d, \ell \geq 1$, $U \in \mathbb{R}^{n \times d}$, and $H_\Delta = e^{\Delta(P-I)}$. Define*

$$Q_- = [U, H_\Delta U, \ldots, H_\Delta^{\ell-1} U], \qquad Q_+ = [H_\Delta U, \ldots, H_\Delta^{\ell} U]. \tag{87}$$

If $\text{rank}(Q_-) = n$, then in exact arithmetic

$$H_\Delta = Q_+ Q_-^{\dagger}, \qquad P = I + \Delta^{-1} \log H_\Delta, \tag{88}$$

using the Moore–Penrose inverse and principal matrix logarithm. Times $b\Delta$ are permitted common-time ordered heat channels when $\Delta > \epsilon_{\text{ord}}$.

Proof. For positive-degree i, $P_{ij} = d_i^{-1} \sum_e B_{ie} B_{je} / |e|$; nonnegativity proves the support claim, and isolated rows add no pairs. Since $Q_+ = H_\Delta Q_-$, full row rank proves the first inverse. On positive-degree nodes, P is similar to the positive semidefinite symmetric matrix $D_v^{-1/2} B D_e^{-1} B^\top D_v^{-1/2}$ and is stochastic; its spectrum lies in $[0, 1]$. The isolated block is the identity. Thus P is diagonalizable with real spectrum and $\log e^{\Delta(P-I)} = \Delta(P - I)$. □

Randomized full rank. Retain $U = [X, Z]$, where $Z \in \mathbb{R}^{n \times d_{\mathrm{pr}}}$ has independent standard Gaussian entries independent of the hypergraph, attributes, and labels. Let n_{eig} count the distinct eigenvalues of P, with largest multiplicity $g_{\max}$. Then $\ell \geq n_{\mathrm{eig}}$ and $d_{\mathrm{pr}} \geq g_{\max}$ give $\operatorname{rank}(Q_-) = n$ almost surely. Indeed, in $P = S\Lambda S^{-1}$, each eigenspace block of $S^{-1}Z$ has full row rank almost surely. A vector annihilating the Krylov matrix yields a Vandermonde system in the distinct eigenvalues of H_Δ; invertibility forces every eigenspace coefficient, and hence the vector, to vanish. In particular, $d_{\mathrm{pr}} \geq n, \ell = 1$ suffices for every P. These probes remain local and are absent from reported runs; the finite attribute encoder needs its own rank check.

A reconstructible random family. Let $\mathcal{E}_n \subseteq \mathcal{A}_n$ be any random simple hypergraph, where candidate supports $e \in \mathcal{A}_n$ satisfy $|e| \geq 2$. Let Ω_2 contain unordered distinct edge pairs with intersection size at least two, and Ω_3 unordered triples supporting a Berge triangle: distinct u, v, w and an ordering e, f, g with $\{u, v\} \subset e$, $\{v, w\} \subset f$, $\{w, u\} \subset g$. Define

$$\varepsilon_n^{\mathrm{all}} = \sum_{\{e,f\} \in \Omega_2} \Pr(e, f \in \mathcal{E}_n) + \sum_{\{e,f,g\} \in \Omega_3} \Pr(e, f, g \in \mathcal{E}_n). \tag{89}$$

These are actual joint inclusion probabilities; edges need not be independent, uniform in size, or bounded in degree. The condition concerns the local incidence law, since partitioning can create singletons or duplicates.

Proposition 7 (Nonuniform and dependent reconstruction family)**.** *A deterministic decoder recovers* $[B]_{\mathrm{col}}$ *from* $G(B)$ *with probability at least* $1 - \varepsilon_n^{\mathrm{all}}$*. For every projected edge* uv*, it forms* $\{u, v\} \cup (N_G(u) \cap N_G(v))$ *and removes repeated sets. An adjacency-matrix implementation costs* $O(n^3)$ *operations.*

Proof. The union bound excludes both configurations with probability at least $1 - \varepsilon_n^{\mathrm{all}}$. Then each projected pair uv belongs to a unique edge e. Its other members are common neighbors. An external common neighbor w would give edges containing uw and vw: if equal they overlap e twice, and if distinct they form a Berge triangle with e. Thus each candidate equals e. Every edge contains a pair; scanning at most n^2 pairs and n neighbors proves recovery and the cost. □

Incidence forests have $\varepsilon_n^{\mathrm{all}} = 0$, permitting dependent, nonuniform edges and order-n edge counts. More generally, $\varepsilon_n^{\mathrm{all}} \to 0$ suffices. The following transfer also accepts other decoders with proved failure bounds, including reconstruction results for independent uniform hypergraphs (Bresler et al., 2025).

Theorem 7 (Heat recovery and asymptotic preservation of local information)**.** *Let* $O_n = (X_n, [B_n]_{\mathrm{col}})$ *have any joint law with* $y_n \in [C]$*, where* $C \geq 2$ *is fixed. Suppose a decoder from* $G(B_n)$ *recovers* $[B_n]_{\mathrm{col}}$ *with failure probability at most* $\varepsilon_n^{\mathrm{rec}}$*. Let* $T_n = \Psi_n(O_n, \xi_n^{\mathrm{pr}})$*, where* Ψ_n *is fixed and measurable and* ξ_n^{pr} *is independent of* (O_n, y_n)*. Suppose* T_n *retains the features in Lemma 2 and enough information to recover* X_n*, with joint rank/attribute-recovery failure probability at most* $\varepsilon_n^{\mathrm{obs}}$*. A decoder from* $(T_n, \xi_n^{\mathrm{pr}})$ *then recovers* O_n *with error at most* $\varepsilon_n = \varepsilon_n^{\mathrm{rec}} + \varepsilon_n^{\mathrm{obs}}$*. For* $\varepsilon_n \leq 1/2$*,*

$$0 \leq \mathcal{R}^*_{\mathrm{heat,n}} - \mathcal{R}^*_{\mathrm{loc,n}} \leq h_2(\varepsilon_n) + \varepsilon_n \log C =: d_n, \tag{90}$$

where $\mathcal{R}^*_{\mathrm{heat,n}} = H(y_n \mid T_n, \xi_n^{\mathrm{pr}})$*,* $\mathcal{R}^*_{\mathrm{loc,n}} = H(y_n \mid O_n)$*, and* $h_2(u) = -u \log u - (1-u)\log(1-u)$*, with* $0 \log 0 = 0$*. Hence* $d_n = O(\varepsilon_n \log(1/\varepsilon_n)) \to 0$ *when* $\varepsilon_n \to 0$*.*

Proof. On simultaneous success, Equation (88) recovers P, its positive off-diagonal support gives $G(B_n)$, and the incidence and attribute decoders give O_n. The union bound proves the error claim. Data processing gives the nonnegative entropy gap. If E indicates decoder failure, O_n is determined by $(T_n, \xi_n^{\mathrm{pr}}, E = 0)$, so

$$I(y_n; O_n \mid T_n, \xi_n^{\mathrm{pr}}) \leq H(E) + \Pr(E = 1) \log C \leq h_2(\varepsilon_n) + \varepsilon_n \log C.$$

Independence of the seed and deterministic generation of T_n identify this mutual information with the entropy gap; Bayes cross-entropy equals conditional entropy. □

Why incidence recovery needs an assumption. On six vertices, $\mathcal{E}_1 = \{123, 145, 246, 356\}$ and $\mathcal{E}_2 = \{124, 135, 236, 456\}$ have degree two and identical pair co-incidences. Thus $B_1 B_1^\top = B_2 B_2^\top$ and $P_1 = P_2 = B_1 B_1^\top/6$, despite different column-permutation classes. Identical attributes give identical heat features at every time. A uniform mixture, padded with common isolated nodes, has reconstruction error at least $1/2$. Duplicating all columns also preserves P for multihypergraphs. Additional heat channels cannot remove either ambiguity.

Recovery at a queried node. For a simple hypergraph $\mathcal{E}$ with edge sizes at least two, let $V_{\mathrm{amb}}(\mathcal{E})$ contain endpoints of pairs lying in two distinct edges and the three cycle vertices of each Berge triangle. Put $p_{n,v}^{\mathrm{query}} = \Pr(v_* = v \mid \mathcal{E}_n)$ for the queried vertex v_*.

Proposition 8 (Query-weighted recovery and prediction risk)**.** *For $v \notin V_{\mathrm{amb}}(\mathcal{E})$, its incident edge set satisfies*

$$\widehat{\mathcal{E}}(v) = \mathrm{unique}\{\{v, u\} \cup (N_G(v) \cap N_G(u)) : u \in N_G(v)\} = \mathcal{E}(v), \tag{91}$$

where $\mathcal{E}(v) = \{e \in \mathcal{E} : v \in e\}$ and unique removes repeats. Under any joint edge law, failure at v_* has probability at most

$$\varepsilon_n^{\mathrm{query}} = \mathbb{E} \sum_{v=1}^{n} p_{n,v}^{\mathrm{query}} \mathbf{1}\{v \in V_{\mathrm{amb}}(\mathcal{E}_n)\}, \tag{92}$$

equal to $\mathbb{E}|V_{\mathrm{amb}}(\mathcal{E}_n)|/n$ for an independent uniform query. More generally, let O_n include all permitted local information and the query, and let S_n^{task} be measurable from O_n. If $T_n = \Psi_n(O_n, \xi_n^{\mathrm{pr}})$ for fixed measurable Ψ_n and a seed independent of (O_n, y_n), $y_n \in [C]$, and $(T_n, \xi_n^{\mathrm{pr}})$ recovers S_n^{task} with error at most $\delta_n \le 1/2$, then

$$0 \le H(y_n \mid T_n, \xi_n^{\mathrm{pr}}) - H(y_n \mid O_n) \le I(y_n; O_n \mid S_n^{\mathrm{task}}) + h_2(\delta_n) + \delta_n \log C. \tag{93}$$

Proof. For $v \notin V_{\mathrm{amb}}$ each pair vu has a unique source edge e. An external common neighbor would create either a repeated pair through v or a Berge triangle with cycle vertex v, exactly as in Proposition 7. Thus the decoder returns precisely all incident edges, including the empty set at an isolated vertex. Conditioning on $\mathcal{E}_n$ proves the failure bound. The failure-indicator argument above gives $I(y_n; S_n^{\mathrm{task}} \mid T_n, \xi_n^{\mathrm{pr}}) \le h_2(\delta_n) + \delta_n \log C$. Use $H(y_n \mid T_n, \xi_n^{\mathrm{pr}}) \le H(y_n \mid S_n^{\mathrm{task}}) + I(y_n; S_n^{\mathrm{task}} \mid T_n, \xi_n^{\mathrm{pr}})$, subtract $H(y_n \mid O_n)$, and apply measurability and data processing. □

If S_n^{task} contains the query, its incident edges and their member attributes (including its own when isolated), the rank and attribute assumptions give $\delta_n \le \varepsilon_n^{\mathrm{query}} + \varepsilon_n^{\mathrm{obs}}$. Only the probability of querying an affected vertex must vanish; $I(y_n; O_n \mid S_n^{\mathrm{task}})$ separately charges label information beyond this neighborhood. No locality of the dense heat operator is assumed. For known two-node simple graphs, projection recovers every edge, including those in cycles.

For a concrete separation, put $m_n \to \infty$, $m_n/n \to 0$ disjoint six-node examples among n vertices, choose their incidences independently and uniformly, and put a fixed incidence forest elsewhere. With fixed noninformative attributes, an encoder independent of these choices, and a retained query index, all 2^{m_n} incidences have identical heat features, even with independent probes. Full reconstruction succeeds with probability at most 2^{-m_n}, yet $\varepsilon_n^{\mathrm{query}} \le 6m_n/n$. Label each block vertex by its block choice and forest vertices by zero. Full incidence determines the labels; for a uniform query the optimal full-matrix heat cross-entropy is $(6m_n/n)\log 2 \to 0$.

These are information bounds for the complete local heat matrix. Its decoder need not be expressible by the implemented finite encoder and rowwise classifier. The next bound retains that approximation error, shared-parameter cost, and training error; neither private probes nor a reconstruction stage is added to the implementation or its payload.

C.12 Stochastic Federated Optimization and Risk Transfer

Population risk and shared parameters. Let a random test node belong to client k with probability w_k^{pop}, and write $\mathcal{R}_k(\theta) = \mathbb{E}[-\log q_{\theta,k}(y)]$, $\mathcal{R}_{\mathrm{fed}}(\theta) = \sum_k w_k^{\mathrm{pop}} \mathcal{R}_k(\theta)$ for the actual prediction rule. Evaluation weights may differ from FedAvg weights $w_k^{\mathrm{tr}} = n_k^{\mathrm{tr}}/\sum_j n_j^{\mathrm{tr}}$. All posterior risks include the same permitted training-side information and exclude held-out labels. If conditioning on a

training transcript, include the realized probe seed in both local and heat information: the representation is then deterministic given these data, and recovery bounds must hold under the conditional law. Unconditional probe independence need not survive this conditioning. Expectations also average over training and algorithmic randomness.

Theorem 8 (From local information to shared-model population risk)**.** *Let Θ be a nonempty shared parameter domain with finite risks and $\widehat{\theta} \in \Theta$. For each client suppose the optimal local and heat risks satisfy $0 \le \mathcal{R}^*_{\mathrm{heat},k} - \mathcal{R}^*_{\mathrm{loc},k} \le d_k$, using Theorem 7 or Proposition 8. Define*

$$a_k = \left[\inf_{\theta\in\Theta} \mathcal{R}_k(\theta) - \mathcal{R}^*_{\mathrm{heat},k}\right]_+,$$
$$a_{\mathrm{share}} = \inf_{\theta\in\Theta} \sum_k w_k^{\mathrm{pop}} \mathcal{R}_k(\theta) - \sum_k w_k^{\mathrm{pop}} \inf_{\theta\in\Theta} \mathcal{R}_k(\theta) \ge 0. \tag{94}$$

Here $[x]_+ = \max\{x, 0\}$; a_k includes finite-encoder and rowwise-readout restrictions, and a_{share} charges shared parameters. For fixed empirical training cross-entropy J_N on N labels, suppose

$$\mathbb{E} \sup_{\theta\in\Theta} |J_N(\theta) - \mathcal{R}_{\mathrm{fed}}(\theta)| \le u_N, \qquad \mathbb{E}[J_N(\widehat{\theta}) - \inf_{\theta\in\Theta} J_N(\theta)] \le e_{N,T}. \tag{95}$$

Then, without independence across clients,

$$\mathbb{E}\left[\mathcal{R}_{\mathrm{fed}}(\widehat{\theta}) - \sum_k w_k^{\mathrm{pop}} \mathcal{R}^*_{\mathrm{loc},k}\right] \le \mathbb{E}\left[\sum_k w_k^{\mathrm{pop}}(d_k + a_k) + a_{\mathrm{share}}\right] + 2u_N + e_{N,T}. \tag{96}$$

Expectations include any conditioning information on which the risks depend.

Proof. Two uniform-discrepancy bounds give $\mathbb{E}\mathcal{R}_{\mathrm{fed}}(\widehat{\theta}) \le \mathbb{E}\inf_\Theta \mathcal{R}_{\mathrm{fed}} + 2u_N + e_{N,T}$. Substitute the definition of a_{share} and $\inf_\Theta \mathcal{R}_k - \mathcal{R}^*_{\mathrm{loc},k} \le a_k + d_k$. For transcript-conditioned risks apply these inequalities conditionally and average. □

The random-test-node objective needs weighted recovery errors, not simultaneous recovery at every client. A single parameter sequence approaching all positively weighted client optima suffices for $a_{\mathrm{share}} = 0$; separate local approximation does not. If cross-entropies lie in $[0, M]$, u_N is at most the training-weighted uniform sampling deviation plus $M\|w^{\mathrm{tr}} - w^{\mathrm{pop}}\|_1/2$. Dependent hypergraphs supply no automatic independent-node concentration rate.

Actual displacements and task progress. Let $d_{k,u,s} = \theta_{k,u,s} - \theta_{k,u,s+1}$ be the realized displacement at local step s of round u, with $S_{k,u}$ steps. Full-participation FedAvg gives

$$D_u = \theta^u - \theta^{u+1} = \sum_k w_k^{\mathrm{tr}} \sum_{s=0}^{S_{k,u}-1} d_{k,u,s}. \tag{97}$$

Masks, minibatches, dropout, clipping, optimizer moments, weight decay, numerical heat evaluation, and target motion enter through these displacements. Different step counts are allowed; subsampling uses normalized participating-client weights.

Proposition 9 (Task-risk bound along the implemented trajectory)**.** *Let J_N be evaluation-mode empirical cross-entropy with exact heat, and let $L_u \ge 0$ bound its largest Hessian eigenvalue on $\{\theta^u - aD_u : 0 \le a \le 1\}$. For every prefix of every finite realized trajectory,*

$$J_N(\theta^t) \le J_N(\theta^0) - \sum_{u<t} \left\{\langle \nabla J_N(\theta^u), D_u\rangle - \tfrac{1}{2} L_u \|D_u\|_2^2\right\}. \tag{98}$$

A positive bracket guarantees strict reduction that round. The inequality holds simultaneously for all prefixes, including any validation-selected checkpoint; its right side minus $\inf_\Theta J_N$, followed by expectation, bounds $e_{N,T}$.

Proof. Taylor's integral remainder is $\int_0^1 (1-a) D_u^\top \nabla^2 J_N(\theta^u - aD_u) D_u \, da \le L_u \|D_u\|_2^2/2$; sum over rounds. This is pathwise and requires neither unbiased gradients, a fixed teacher, nor gradient domination. The native evaluation network's linear maps, GELU, softplus, matrix exponential,

positive-epsilon LayerNorm, and cross-entropy are smooth, giving finite curvature on each finite segment. The scale model also uses LayerNorm and a smooth closed-form heat operator. Predictive anti-collapse losses affect D_u but are absent from J_N; baseline BatchNorm buffers are outside this proof. □

If implemented logits differ from exact logits by at most $\epsilon_u^{\text{logit}}$ in infinity norm at checkpoint u, the 2-Lipschitz property of cross-entropy adds only $2\epsilon_0^{\text{logit}} + 2\epsilon_t^{\text{logit}}$ to the implemented-loss bound. Gradients and curvatures refer to exact heat; intermediate numerical errors enter D_u.

Classifier, Adam, and target identities. Split parameters into context ϑ, classifier θ_{cls}, and masked predictor θ_{pred}. The predictive loss does not use the classifier, so

$$\nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{pred,k}} = 0, \qquad \nabla_{\theta_{\text{cls}}} \mathcal{L}_k = \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{CE,k}}. \tag{99}$$

Joint clipping can rescale this gradient, and context updates change its inputs. Suppress client/round indices, let g_s be the realized stochastic gradient, $c_s^{\text{clip}} \in [0,1]$ its clipping multiplier, $\ell_s > 0$ the learning rate, and $\lambda_{\text{wd}} \geq 0$ weight decay. Coupled Adam uses $z_s = c_s^{\text{clip}} g_s + \lambda_{\text{wd}} \theta_s$. From zero moments,

$$\widehat{m}_s = z_s + \sum_{j<s} \frac{(1-\beta_1)\beta_1^{s-j}}{1-\beta_1^{s+1}} (z_j - z_s), \qquad d_s = \ell_s \operatorname{diag}(\sqrt{\widehat{v}_s} + \epsilon_{\text{Adam}})^{-1} \widehat{m}_s, \tag{100}$$

where $\widehat{v}_s = (1-\beta_2) \sum_{j \leq s} \beta_2^{s-j} z_j^{\odot 2} / (1-\beta_2^{s+1})$, $\beta_1, \beta_2 \in [0,1)$, $\epsilon_{\text{Adam}} > 0$, and $\odot 2$ is coordinatewise squaring. AdamW instead sets $z_s = c_s^{\text{clip}} g_s$ and adds $\ell_s \lambda_{\text{wd}} \theta_s$ to d_s. The native optimizer resets each round. The target also resets to $\bar{\vartheta}_0 = \vartheta_0$, and post-step EMA gives

$$\bar{\vartheta}_s = \vartheta_0 - \sum_{j<s} (1-\mu^{s-j}) d_j^{\text{ctx}}, \qquad \vartheta_s - \bar{\vartheta}_s = -\sum_{j<s} \mu^{s-j} d_j^{\text{ctx}}. \tag{101}$$

These exact identities enter Equation (98), even when moving-target updates are not gradients of one fixed loss. EMA controls lag to the current context, not distance to a fixed reference teacher. Neither these identities nor the architecture guarantees task alignment; general Adam can fail even on convex problems (Reddi et al., 2018).

Conditional convergence through the affine output layer. Write $\theta = (\theta_{\text{rest}}, \theta_{\text{out}})$, where θ_{out} consists of the final affine classification weights and bias. With θ_{rest} fixed, evaluation cross-entropy is convex in this block, which also satisfies Equation (99). This applies to both FedASTRA implementations and the optional deeper classifier, whose preceding layers belong to θ_{rest}.

Theorem 9 (Global task-risk convergence along actual updates)**.** *Use J_N, D_u, L_u from Proposition 9, let $J_N^* = \inf_{\theta \in \Theta} J_N(\theta) > -\infty$ and $E_u = J_N(\theta^u) - J_N^*$. Fix $0 \leq u_0 < T$. For every $u = u_0, \ldots, T$, assume a comparator with actual and comparison parameters in Θ satisfying*

$$J_N(\theta_{\text{rest}}^u, \theta_{\text{out}}^{u,\circ}) \leq J_N^* + a_{\text{ctx}}, \qquad \|\theta_{\text{out}}^u - \theta_{\text{out}}^{u,\circ}\|_2 \leq D_{\text{cls}}, \tag{102}$$

with $a_{\text{ctx}} \geq 0$ and $0 < D_{\text{cls}} < \infty$. This assumption concerns affine-head adequacy on each realized representation, separately from full-model approximation over Θ. Define $\Delta_u = \langle \nabla J_N(\theta^u), D_u \rangle - L_u \|D_u\|_2^2/2$. Suppose $\gamma_u > 0$, $\zeta_u \geq 0$ satisfy

$$\Delta_u \geq \gamma_u \|\nabla_{\theta_{\text{out}}} J_N(\theta^u)\|_2^2 - \zeta_u, \qquad u = u_0, \ldots, T-1. \tag{103}$$

Then, pathwise,

$$E_{u+1} \leq E_u - \frac{\gamma_u}{D_{\text{cls}}^2} (E_u - a_{\text{ctx}})_+^2 + \zeta_u. \tag{104}$$

If all deficits ζ_u vanish, then

$$E_T \leq a_{\text{ctx}} + r_{u_0,T}, \qquad r_{u_0,T} := \left[(E_{u_0} - a_{\text{ctx}})_+^{-1} + D_{\text{cls}}^{-2} \sum_{u=u_0}^{T-1} \gamma_u \right]^{-1}, \tag{105}$$

with $r_{u_0,T} = 0$ if $(E_{u_0} - a_{\text{ctx}})_+ = 0$. In particular, $\gamma_u \geq \gamma > 0$ gives $r_{u_0,T} \leq D_{\text{cls}}^2/[\gamma(T-u_0)]$. For general deficits,

$$\min_{u_0 \leq u < T} (E_u - a_{\text{ctx}})_+ \leq D_{\text{cls}} \sqrt{\frac{E_{u_0} + \sum_{u=u_0}^{T-1} \zeta_u}{\sum_{u=u_0}^{T-1} \gamma_u}}. \tag{106}$$

Proof. Affine-head convexity and Equation (102) give

$$E_u - a_{\mathrm{ctx}} \le \langle \nabla_{\theta_{\mathrm{out}}} J_N(\theta^u), \theta^u_{\mathrm{out}} - \theta^{u,\circ}_{\mathrm{out}} \rangle \le D_{\mathrm{cls}} \|\nabla_{\theta_{\mathrm{out}}} J_N(\theta^u)\|_2.$$

Combining with $E_{u+1} \le E_u - \Delta_u$ proves the recurrence. For zero deficits, E_u is nonincreasing. Put $\delta_u = (E_u - a_{\mathrm{ctx}})_+$ and $\kappa_u = \gamma_u / D^2_{\mathrm{cls}}$. Whenever $\delta_{u+1} > 0$, $\delta_{u+1} \le \delta_u(1 - \kappa_u \delta_u)$ with $\kappa_u \delta_u < 1$, hence $\delta^{-1}_{u+1} \ge \delta^{-1}_u + \kappa_u$. A zero δ_u stays zero. Summing proves the reciprocal rate. For arbitrary deficits, summing the recurrence and using $E_T \ge 0$ gives $D^{-2}_{\mathrm{cls}} \sum_u \gamma_u \delta^2_u \le E_{u_0} + \sum_u \zeta_u$; bounding the minimum by the weighted mean proves the last claim. □

The coefficients γ_u measure realized task progress and are not Adam learning rates. For any chosen $\gamma_u > 0$, the deficit $\zeta_u = [\gamma_u \|\nabla_{\theta_{\mathrm{out}}} J_N(\theta^u)\|^2_2 - \Delta_u]_+$ makes the deficit bound applicable. The $O((T - u_0)^{-1})$ conclusion additionally requires adequate realized contexts, zero deficits, and progress coefficients bounded away from zero. The architecture supplies affine-head convexity; it does not establish these trajectory conditions. No unbiased-gradient or full-network Polyak–Lojasiewicz condition is used.

Corollary 9 (Global population convergence of the returned model)**.** *Assume Theorems 8 and 9, with $\zeta_u = 0$ on the stated interval. For the actual validation-selected $\widehat{\theta}_{N,T}$ define $s^{\mathrm{sel}}_{N,T} = [J_N(\widehat{\theta}_{N,T}) - J_N(\theta^T)]_+$, and put $A_N = \mathbb{E}[\sum_k w^{\mathrm{pop}}_k (d_k + a_k) + a_{\mathrm{share}}] + 2u_N$. Then*

$$\mathbb{E}\left[\mathcal{R}_{\mathrm{fed}}(\widehat{\theta}_{N,T}) - \sum_k w^{\mathrm{pop}}_k \mathcal{R}^*_{\mathrm{loc},k}\right] \le A_N + \mathbb{E}[a_{\mathrm{ctx}} + r_{u_0,T} + s^{\mathrm{sel}}_{N,T}]. \tag{107}$$

Expectations include all path-dependent quantities. If $\gamma_u \ge \gamma > 0$ and $\mathbb{E}D^2_{\mathrm{cls}}$ is uniformly bounded, the reciprocal term is $O((T - u_0)^{-1})$ in expectation. Vanishing A_N, $\mathbb{E}a_{\mathrm{ctx}}$, $\mathbb{E}r_{u_0,T}$, and $\mathbb{E}s^{\mathrm{sel}}_{N,T}$ along a joint data-size and training-horizon limit give local-information Bayes consistency.

Proof. By definition, $J_N(\widehat{\theta}_{N,T}) - J^*_N \le E_T + s^{\mathrm{sel}}_{N,T}$. Insert Equation (105) into Theorem 8; the stated rate and limit follow. □

The selection term covers the implementation's validation Accuracy/Macro-F1 rule. Nonzero alignment deficits retain only Equation (106)'s best-iterate guarantee, unless selection is controlled separately. These conditional rates do not follow from finite experiments or from the choice of Adam, EMA, and FedAvg alone.

Predictive-surrogate mismatch. The same decomposition for one predictor class makes objective mismatch explicit.

Corollary 10 (Conditional local Bayes consistency)**.** *Let d_n be from Theorem 7, and let $\mathcal{Q}_N$ be an available predictor class measurable from retained heat information, with population cross-entropy $R(q)$ and approximation error $a_N = \inf_{q \in \mathcal{Q}_N} R(q) - \mathcal{R}^*_{\mathrm{heat,n}} \ge 0$. Suppose $\mathbb{E} \sup_{q \in \mathcal{Q}_N} |\widehat{R}_N(q) - R(q)| \le u_N$ and the returned $\widehat{q}_{N,T} \in \mathcal{Q}_N$ satisfies*

$$\mathbb{E}[\widehat{R}_N(\widehat{q}_{N,T}) - \inf_{q \in \mathcal{Q}_N} \widehat{R}_N(q)] \le \kappa_N + e_{N,T}, \tag{108}$$

where $\kappa_N, e_{N,T} \ge 0$ bound surrogate/readout mismatch and optimization. Then

$$\mathbb{E}R(\widehat{q}_{N,T}) - \mathcal{R}^*_{\mathrm{loc,n}} \le d_n + a_N + 2u_N + \kappa_N + e_{N,T}. \tag{109}$$

Thus $\varepsilon_n, a_N, u_N, \kappa_N, e_{N,T} \to 0$ imply local Bayes consistency.

Proof. Two uniform-deviation bounds yield $\mathbb{E}R(\widehat{q}_{N,T}) \le \inf_{\mathcal{Q}_N} R + 2u_N + \kappa_N + e_{N,T}$; substitute a_N and Equation (90). For a fixed-teacher joint objective $J_N = \widehat{R}_N + \lambda_{\mathrm{pred}} \widehat{L}_{\mathrm{pred,N}}$, let q index the complete model, including auxiliary heads. If a comparator q°_N has CE gap at most κ^{CE}_N and predictive-loss gap at most κ^{pred}_N, both relative to their infima over $\mathcal{Q}_N$, an $e_{N,T}$-optimal solution of J_N satisfies Equation (108) with $\kappa_N = \kappa^{\mathrm{CE}}_N + \lambda_{\mathrm{pred}} \kappa^{\mathrm{pred}}_N$. This follows by subtracting the predictive-loss lower bound. It requires a fixed surrogate; moving-teacher updates do not automatically optimize that objective. □

Information recovery controls d_n; the finite representation/readout controls a_N; the data law and training procedure must control $u_N, \kappa_N, e_{N,T}$. Under label sufficiency in Section 4, the local Bayes limit equals the enriched-information limit, and Equation (17) gives the criterion for beating a particular fitted enriched predictor.

D Experimental Protocols and Implementation Details

D.1 FedASTRA Implementation

Network and heat parameterization. The context encoder applies a linear map, layer normalization, and GELU to the raw attributes, followed by encoder dropout. Let its width be d. The training forward pass uses the type-marginalized binary incidence matrix from Equation (18). Observed types remain outside the model forward, and repeated relation instances remain separate columns. The established finite tasks use $R = 4$, while Trivago uses $R = 8$. Every unconstrained increment in Equation (20) starts at zero, so the initial time is close to $b \log 2$ for $b = 1, \ldots, R$. The increment floor is $\epsilon_{\text{ord}} = 10^{-4}$, and the strict cumulative parameterization is preserved throughout optimization. The classifier receives $(R + 1)d$ coordinates, applies layer normalization and dropout, and uses one linear readout. The predictive head maps $(R+1)d$ to $2d$, applies layer normalization and GELU, and maps back to $(R + 1)d$. For each local update, two independent Bernoulli masks remove attribute coordinates. The removal probability is 0.40 on the established tasks and 0.30 on Trivago, with retained values divided by 0.60 and 0.70, respectively. Incidence indices, relation columns, and node membership remain fixed across both online views and the target view. The target context encoder is copied from the round's broadcast context encoder. This copy contains both the attribute map and the raw ordered-time increments. All of these coordinates are updated with $\mu = 0.99$ on the established tasks and $\mu = 0.996$ on Trivago after each local step, then discarded before upload. The copy remains outside the optimizer, saved checkpoints, and the transmitted online state dictionary.

Mask variance and target memory. Write the coordinate-removal probability as $p < 1$ and let $\delta_j \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - p)$. The inverted mask $\mathcal{M}_p(x) = \delta \odot x/(1 - p)$ obeys exactly

$$\mathbb{E}_\delta \mathcal{M}_p(x) = x, \qquad \mathbb{E}_\delta \|\mathcal{M}_p(x) - x\|_2^2 = \frac{p}{1-p}\|x\|_2^2. \tag{110}$$

Thus p controls the exact variance of each local prediction view. Within one client round, write $\vartheta_s = \text{ctx}(\theta_s)$ and $\bar{\vartheta}_s$ for the online and target context states after local step s, with the round reset $\bar{\vartheta}_0 = \vartheta_0$. The update $\bar{\vartheta}_s = \mu\bar{\vartheta}_{s-1} + (1 - \mu)\vartheta_s$ expands over the complete local context history as

$$\bar{\vartheta}_s = \mu^s \vartheta_0 + (1 - \mu) \sum_{j=1}^{s} \mu^{s-j} \vartheta_j. \tag{111}$$

The momentum therefore assigns a geometric age profile with effective scale $(1 - \mu)^{-1}$ to the attribute-map and time-increment coordinates together. Both operations remain client-local. For every fixed $p < 1$ and $\mu \in [0, 1)$, coupling their randomness preserves the complete observation invariance in Corollary 4. The mechanism table evaluates the complete predictive path, keeping p and μ fixed across all paired conditions to isolate the effect of enabling this component.

Exact local objective. Each local update encodes two masked views online. At client k, let $\vartheta_k = \text{ctx}(\theta_k)$ and let $\bar{\vartheta}_k = (\bar{\theta}_{k,\text{enc}}, \bar{\eta}_k)$ denote the round-local EMA copy of the complete context encoder, including its learned time increments. The target branch forms $Z_k^{\text{tar}} = W_k(X_k \mid \bar{\vartheta}_k)$, while $\widehat{Z}_k^{(a)} = p_{\theta_k}(W_k(\widetilde{X}_k^{(a)} \mid \vartheta_k))$ predicts it for $a \in \{1, 2\}$. Here the rowwise predictor is $p_{\theta_k} : \mathbb{R}^{(R+1)d} \to \mathbb{R}^{(R+1)d}$ and the rowwise classifier for C classes is $c_{\theta_k} : \mathbb{R}^{(R+1)d} \to \mathbb{R}^C$. We write $[C] = \{1, \ldots, C\}$ and take $y_v \in [C]$. Define $n_k = |V_k|$ and let LN denote rowwise layer normalization (LN), let sg stop gradients, and let cos denote cosine similarity. Set $\mathcal{A}_k^{(a)} = n_k^{-1} \sum_{v \in V_k} [1 - \cos(\text{LN}(\widehat{Z}_{k,v}^{(a)}), \text{sg}(\text{LN}(Z_{k,v}^{\text{tar}})))]$. With an anti-collapse term inspired by variance–invariance–covariance regularization (VICReg) $\mathcal{R}_{\text{ac}}$ (Bardes et al., 2022) and a fixed coefficient $\gamma_{\text{view}} \geq 0$,

$$\mathcal{L}_{\text{pred},k} = \frac{1}{2} \sum_{a=1}^{2} [\mathcal{A}_k^{(a)} + \mathcal{R}_{\text{ac}}(\text{LN}(\widehat{Z}_k^{(a)}))] + \gamma_{\text{view}} \left[1 - \frac{1}{n_k} \sum_{v \in V_k} \cos(\widehat{Z}_{k,v}^{(1)}, \widehat{Z}_{k,v}^{(2)})\right]. \tag{112}$$

For $A \in \mathbb{R}^{n\times h}$, define $\overline{A}_j = n^{-1}\sum_{i=1}^n A_{ij}$ and $\mathrm{Var}_n(A_{\cdot j}) = \sum_{i=1}^n (A_{ij} - \overline{A}_j)^2/\max(n-1,1)$. Thus the implementation uses sample variance for $n > 1$ and the zero variance for $n = 1$. Define

$$\mathcal{V}(A) = \frac{1}{h}\sum_{j=1}^{h}\left[1 - \sqrt{\mathrm{Var}_n(A_{\cdot j}) + 10^{-4}}\right]_+, \tag{113}$$

$$C(A) = \frac{1}{h}\left\|\mathrm{offdiag}\left(\frac{\widetilde{A}^\top\widetilde{A}}{\max(n-1,1)}\right)\right\|_F^2, \qquad \widetilde{A}_{\cdot j} = \frac{A_{\cdot j} - \overline{A}_j}{\sqrt{\mathrm{Var}_n(A_{\cdot j}) + 10^{-4}}}. \tag{114}$$

The anti-collapse term in Equation (112) is $\mathcal{R}_{\mathrm{ac}}(A) = 0.10\mathcal{V}(A) + 0.005C(A)$ and receives $A = \mathrm{LN}(\widehat{Z}_k^{(a)})$, the rowwise layer-normalized predictor matrix defined before Equation (112). The term $\mathcal{A}_k^{(a)}$ averages the nodewise target-alignment cosine losses. The two-view agreement weight is $\gamma_{\mathrm{view}} = 0.10$. The fixed predictive coefficient is $\lambda_{\mathrm{pred}} = 1$. Supervised cross-entropy is evaluated only on the training mask. Prediction uses all client-local nodes, attributes, and incidence. The validation-selected checkpoint uses Accuracy with Macro-F1 tie-breaking and is tested once.

One recurrence for all diffusion times. All learned times share one recurrence. Starting from U_k, the client applies P_k once to obtain each subsequent power and accumulates the corresponding float64 Poisson coefficients into every basis element. Let L_{tail} be the degree chosen by Equation (59), let $\mathcal{E}_k^{\mathrm{rel}}$ be the retained relation columns, and let $\Omega_k = \{(v,e) : B_{k,ve} = 1\}$ be the client incidence list. One application of P_k costs $O(|\Omega_k|d)$ time and $O((|V_k| + |\mathcal{E}_k^{\mathrm{rel}}|)d)$ working storage. The complete basis computation therefore has the following time and storage complexity:

$$O(L_{\mathrm{tail}}|\Omega_k|d + L_{\mathrm{tail}}R|V_k|d) \quad \text{time}, \qquad O\Big((R+1)|V_k|d + |\mathcal{E}_k^{\mathrm{rel}}|d + |\Omega_k|\Big) \quad \text{storage}. \tag{115}$$

The propagation recurrence is shared across basis elements. Increasing R adds coefficient accumulation and representation width without another sequence of incidence traversals. A degree-L_{tail} polynomial filter requires the same L_{tail} sparse traversals, while R heat channels reuse those powers and add only the accumulation term in Equation (115). The current maximum learned time and fixed tolerance determine the certified degree L_{tail}. Although the target heat operator has infinite Poisson support, Equation (58) evaluates it at finite degree with certified forward and time-gradient errors. Table 5 further verifies the certified recurrence depth, numerical accuracy, and local evaluation cost under the finite-task implementation.

Table 5: Evaluator-only shared-recurrence audit on one CoraCA client at $K = 6$. Certified tails use tolerance 10^{-6}, and residuals compare with a 10^{-12} reference under the finite-task evaluator. Central processing unit (CPU) timings are in milliseconds (ms) on identical local inputs.

R	L_{tail}	Forward tail	Time-gradient tail	Relative ℓ_2 residual	CPU median (ms)
2	11	2.95×10^{-8}	2.87×10^{-7}	3.07×10^{-8}	26.30
4	15	4.35×10^{-8}	2.98×10^{-7}	2.26×10^{-8}	33.72
6	18	9.34×10^{-8}	5.26×10^{-7}	1.48×10^{-8}	33.27
8	21	1.06×10^{-7}	5.34×10^{-7}	2.20×10^{-8}	42.99

The audited client has 445 nodes, 485 anonymous relation columns, 774 incidences, and 64 feature channels. Each median uses five CPU timing repeats. Every setting has maximum absolute residual 2.38×10^{-7}. This study audits recurrence depth, numerical error, and local evaluation cost. The paired reported $R \in \{1, 4\}$ mechanism study supplies the accompanying accuracy comparison under matched CoraCA and DBLP4k training runs at $K = 6$. The runs use the shared anonymous release.

Communication protocol FedASTRA communicates only the online model state, while auxiliary training states remain client-local. FedAvg uses each client's fixed labeled-training count n_k^{tr}. Optimizer state is created at the beginning of a client's round and discarded after its local training. The local EMA target, which is reset at each round, learned node states, incidence data, and all numerical heat buffers remain local. For auditability, each finite run records parameter count, state-dictionary bytes, both communication directions, learned times, Poisson mass assigned to walk lengths above two, and both numerical tail bounds before test evaluation of the validation-selected checkpoint and before any test label is read.

Threat model and privacy boundary. Table 6 states the server view. Zero auxiliary structural payload limits the transcript to online model updates and fixed FedAvg train-count weights. End-to-end differential privacy is a separate property. Model updates can reveal attributes or labels, and secure aggregation alone can leave aggregate inversion attacks possible (Shi et al., 2025). Deployments seeking record-level privacy require a declared threat model, calibrated randomization, privacy accounting, and compatible secure aggregation. These additions can be composed with the structural observation model under an explicit deployment-specific privacy guarantee.

Table 6: Server transcript and client-local state for FedASTRA.

Protocol item	Declared scope
Server receives each round	Online model parameters from each participating client.
Server broadcasts	The aggregated online model parameters.
Fixed aggregation metadata	Each client's labeled-training count, established before training and excluded from learned-state byte totals.
Remains client-local	Raw attributes, labels, incidence tables, relation membership, relation identifiers, round-local optimizer state, heat powers, latent states, and the local EMA target context encoder reset at each round.
Auxiliary learned-state payload	Zero bytes. Embeddings, relation summaries, and target parameters remain client-local.
Structural access claim	Direct cross-client node, edge, and source-hyperedge identifiers are absent from the prescribed transcript.
Privacy extension	Secure aggregation and differential privacy require their own mechanism, parameters, and accountant.

D.2 Graph Partitions and Lifts

Main graph partition. For the main experiments, following (Yao et al., 2023; Luo et al., 2025), we fix the task split before client assignment and partition nodes across $K \in 3, 6, 9$ clients using class-conditioned Dirichlet sampling. Client allocation probabilities are drawn from a symmetric Dirichlet distribution with concentration 10^4, subject to the declared load cap and minimum client size. The resulting assignment is precomputed from the full class labels, fixed before model training, and shared by all compared methods. For graph datasets, this node partition is also fixed before constructing client-local one-hop relations from the induced subgraphs.

Truncation-controlled graph partition. Each paired comparison begins from one balanced label allocation and permits swaps between nodes of the same class. Every proposal therefore preserves each client's load and label histogram. With nonnegative fixed weights α_2, α_3 and $\rho_k = n_k / \sum_j n_j$, the three objectives are

$$
\begin{aligned}
J_{\mathrm{low}}(z) &= \sum_{q\in\{2,3\}} \alpha_q \tau_q(z), \\
J_{\mathrm{ref}}(z) &= \sum_{q\in\{2,3\}} \alpha_q \left(\tau_q(z) - \left[1 - \sum_k \rho_k^q\right]\right)^2, \quad \text{s.t.} \quad |V_k| = n_k, \qquad \mathrm{hist}(Y_{V_k}) = h_k. \qquad (116) \\
J_{\mathrm{high}}(z) &= -\sum_{q\in\{2,3\}} \alpha_q \tau_q(z),
\end{aligned}
$$

We minimize the corresponding objective for the low-truncation, independent-allocation reference, and high-truncation arms. Each pass proposes 2,048 same-class cross-client swaps, accepts strict improvements, and stops after at most 16 passes or after a pass accepts zero swaps. The three arms thus have identical client loads and client-by-label histograms while their truncation coordinates differ. Hyperedges are treated as sets, relation multiplicity and type are retained, and every objective is invariant to stored member order. The prespecified diagnostic assigns equal weight to the pair and triple coordinates, $\alpha_2 = \alpha_3 = 1$. These evaluator coordinates are fixed diagnostics, separate from the heat-basis width and the learned model's transition-length mixture in every reported arm.

Graph lifting. For graph datasets, we adopt a partition-then-lift protocol: each client first receives the induced subgraph $G[V_k]$, from which one-hop relations are constructed without remote endpoints or cross-client edges. The following proposition clarifies its relation to the alternative lift-then-partition construction.

Proposition 10 (Lift-then-partition and partition-then-lift protocols)**.** *Let the rooted global one-hop relation be $e_v = \{v\} \cup N_G(v)$. For a local root $v \in V_k$, the two graph-lift protocols obey the exact*

identity below for every client and every locally assigned root under one-hop lifting:

$$e_v \cap V_k = \{v\} \cup N_{G[V_k]}(v). \tag{117}$$

Thus a local-root one-hop fragment has the same members under both constructions. Lift-then-partition additionally creates fragments on client k for roots $v \notin V_k$. Partition-then-lift contains locally rooted columns. For multi-hop relations, even a local root can differ because a path may leave V_k and return. This identity locates the extra globally constructed information in lift-then-partition and isolates globally supplied fragments from edges removed by the induced client graph.

D.3 Dataset Details

CoraCA, DBLP4k, and IMDB4k are loaded from the public dataset release accompanying HGNN and HGNN+ (Feng et al., 2019; Gao et al., 2023). CoraCA uses coauthorship relations (Yadati et al., 2019). DBLP4k retains paper and term relations, and IMDB4k retains director and actor relations as separate channels (Sun et al., 2011; Fu et al., 2020). Trivago follows the public benchmark for deep hypergraph learning, DHG-Bench (Li et al., 2026). It contains 172,738 hotel nodes, 233,202 browsing-session hyperedges, 726,861 incidences, 160 classes, and the public released 300-dimensional node-feature matrix. Each browsing session remains a distinct relation instance, and every method receives the same unprojected feature matrix. Rows with equal member sets remain separate across channels, and every source relation instance is retained. FedASTRA combines all released incidence columns in one observable operator. Relation marks define analysis strata and add no categorical relation embedding to the online model. The graph-derived track uses the Cora and CiteSeer citation networks (Yang et al., 2016) and Facebook Page–Page (Rozemberczki et al., 2021). Each local root defines a closed one-hop neighborhood after graph partitioning. Scale preprocessing scans every node and citation in ogbn-papers100M under its official temporal split (Hu et al., 2020). For a root $v \in V_k$, the papers100M relation is exactly $\{v\} \cup (N_G(v) \cap V_k)$, so every retained incidence is visible to its assigned client through the common released hyperedge table.

D.4 Federated Protocols and Prespecified Configurations

Training and evaluation protocols. Across comparisons, we use the same data splits, released inputs, communication rounds, local updates, and validation-selection rule. The native evaluation includes CoraCA, DBLP4k, IMDB4k, and Trivago with $K \in \{3, 6, 9\}$. All Trivago experiments use three repetitions. On the other native datasets, Fed-HNHN uses three repetitions, while all remaining methods use five. The graph evaluation contains Cora, CiteSeer, and Facebook under the partition-then-lift construction with the same client counts and five repetitions. The controlled study contains CoraCA and DBLP4k at $K = 6$ under low-truncation, independent-allocation reference, and high-truncation partitions. Its paired construction preserves each client's load and label histogram. The mechanism study crosses basis width one or four with predictive weight zero or one on the same two datasets at $K = 6$. Its full-model condition is shared with the corresponding primary evaluation. The scale study contains one ogbn-papers100M execution over all nodes and citations without node, neighbor, or relation sampling, with eight clients, 20 communication rounds, and the complete all-root streaming schedule below, covering every official split across all seven methods.

Training fractions are 0.10, 0.06, and 0.06 for CoraCA, DBLP4k, and IMDB4k, with validation and test fractions of 0.20 and 0.40. Trivago uses a per-class stratified 0.50/0.25/0.25 split that assigns every node once. Unlabeled nodes retain structural roles. Attributes and released local structure for all nodes remain visible to the transductive encoders. Training labels determine model parameters, validation accuracy selects one checkpoint, and macro-F1 resolves an exact tie. No test label or test metric enters either step. The selected online model is then evaluated once with the test labels.

The established finite conditions train for 150 communication rounds with three local updates per round. Trivago uses 30 rounds with three local updates and three independent repetitions for every native method. Each Fed-HNHN condition on the other native datasets uses three, and the other established finite studies use five. The additional heat-width study uses three paired repetitions on each dataset. All reported means and sample standard deviations use these repetition counts. The evaluator retains fragment multiplicity and type, whereas the learner uses the coarsened anonymous-incidence interface. The sharper two-client assumptions in Theorem 1 apply only to the identification analysis and do not alter the algorithm's input.

Baseline configurations. The native comparison includes the native federates hypergraph neural networks (HGNN, Fed-HGNN), Hypergraph Networks with Hyperedge Neurons (HNHN, Fed-HNHN), and graph convolutional networks (GCNs) on hypergraphs(HyperGCN, Fed-HyperGCN). It also includes the AllSet Deep Sets and Set Transformer variants (Fed-AllDeepSets, Fed-AllSetTransformer), UniGCNII with initial residual connections and identity mapping (Fed-UniGCNII), and equivariant-diffusion hypergraph neural networks (ED-HNN, Fed-ED-HNN) (Feng et al., 2019; Dong et al., 2020; Yadati et al., 2019; Zaheer et al., 2017; Chien et al., 2022; Huang & Yang, 2021; Wang et al., 2023). Every native method consumes identical client-local attributes and anonymous released incidence columns. Their column indices are independent across clients, and only trainable model parameters cross the federation. A baseline result is included only when its dataset, task split, client assignment, observed structure, round and local-update budget, and evaluation protocol match the proposed condition. Aggregation and paired comparisons retain only protocol-matched complete runs. Fed-HNHN follows the public FedHGL equations with bias-free mean maps and hidden BatchNorm (Luo et al., 2025). Its running statistics stay client-local, only trainable parameters enter FedAvg, and its auxiliary payload is zero.

The graph comparison follows the evaluation protocol reported in Luo et al. (2025) and includes FedSage, FedGCN (0-hop), FedSage+, FedGCN (2-hop), and FedCog, followed by FedASTRA. Every method uses the same task split, node assignment, training horizon, and evaluation rule. FedSage, FedGCN (0-hop), and FedASTRA operate on client-induced structure. FedSage+ exchanges generated boundary features and reconstruction gradients without direct cross-client identifiers (Zhang et al., 2021). FedGCN (2-hop) receives cross-client node identifiers, edges, and two-hop raw features (Yao et al., 2023). FedCog coordinates cross-client intermediate representations using the corresponding node and edge links (Lei et al., 2023). These observation assumptions are recorded per row and remain part of the interpretation of every predictive result, keeping the local-input comparison distinct from direct-access reference results in every table.

Table 7 records the common FedASTRA configuration. All entries use raw attributes and set preprocessing powers to zero. The positive-time channels share one power recurrence and an explicit identity channel, with $R = 4$ on established tasks and $R = 8$ on Trivago. The learned times determine the certified degree through Equation (59). The theoretical co-membership order q, the diffusion-basis width R, and the recurrence degree L_{tail} describe different objects throughout the analysis and implementation. FedASTRA has no order-q input and performs no selection over q. The identification statements cover every integer order $q \geq 2$, including $q = 4$ and $q = 5$. These evaluator queries do not alter the trained predictor or its model-selection procedure.

Table 7: Configurations of FedASTRA.

Item	Established finite tasks	Trivago
Federation	FedAvg with train-count weights, 150 rounds, three local updates	FedAvg with train-count weights, 30 rounds, three local updates
Hyperedge input	Type-marginalized binary incidence, unit weights	Type-marginalized binary incidence, unit weights
Online model	Width 64, dropout 0.35	Width 128, dropout 0.20
Optimization	Adam, learning rate 10^{-2}, weight decay 5×10^{-4}, gradient clip 5.0	Adam, learning rate 3×10^{-3}, weight decay 5×10^{-4}, gradient clip 5.0
Heat basis	$R = 4$, increment floor 10^{-4}, tail tolerance 10^{-6}	$R = 8$, increment floor 10^{-4}, tail tolerance 10^{-6}
Predictive path	Two masks at 0.40, predictive weight 1, EMA 0.99	Two masks at 0.30, predictive weight 1, EMA 0.996
Transmitted state	Online model parameters, zero auxiliary payload	Online model parameters, zero auxiliary payload
Selection	Validation checkpoint, one test evaluation	Validation checkpoint, one test evaluation

Ablation settings. We evaluate $R \in \{0, 1, 2, 4, 8\}$ at $K = 6$ with $\lambda_{\text{pred}} = 1$ on CoraCA, DBLP4k, and Trivago. Three paired repetitions share node splits, client assignments, and training-seed keys across all five widths. The $R = 0$ control retains the feature encoder, classifier, masked predictor, and EMA target but bypasses all diffusion. The complete study therefore contains 45 fitted models and reports every prespecified width on all three datasets.

The parameter column includes the online encoder, learned time increments, classifier, and predictive head, excluding the client-local EMA copy. Let d_x be input-feature dimension, d the hidden width, and C the class count. For this study's single-linear-layer classifier, the online parameter count is

$$N_R = d_x d + 4d^2 + (C + 12)d + C + Rd(4d + C + 4). \tag{118}$$

The attribute encoder contributes $d_x d+3d$, the time increments Rd, the classifier $(R+1)d(C+2)+C$, and the predictive head $4(R+1)d^2+(R+1)d+6d$. Their sum gives the formula. Increasing R jointly changes the diffusion basis, downstream input width, and model capacity, so the study measures both predictive utility and the corresponding parameter and computation costs at each representation size.

Each dataset runs on one NVIDIA RTX A6000 graphics processing unit (GPU), with repetitions and widths executed serially. Elapsed time synchronized through the Compute Unified Device Architecture (CUDA) begins before client-batch device transfer and includes model initialization, every training round, validation selection, checkpoint restoration, and final model diagnostics. Dataset loading, split construction, and the terminal test pass are excluded. Consequently, runtimes compare widths within the same dataset and training horizon. Figure 3 and Table 16 report Accuracy, Macro-F1, parameters, and elapsed fitting time using this common measurement boundary.

The three-repetition width sweep uses the current implementation. Table 11 instead reports five internally paired repetitions from an earlier implementation whose exact source snapshot is unavailable. Recorded splits and assignments establish pairing within that study. Both studies are summarized independently, without pooling their measurements or pairing runs across implementation versions.

Communication accounting. Let b_θ be the number of bytes in the online state dictionary and let $K_u = |C_u|$ be the number of participating clients in round u. For FedASTRA, the server holds the fixed labeled-training counts used for aggregation. These scalar metadata are excluded from the learned-state payload, which contains one download and one upload per participating client according to the exact accounting identity

$$b_{\text{round},u} = K_u b_\theta + K_u b_\theta = 2K_u b_\theta, \qquad b_{\text{auxiliary}} = 0. \tag{119}$$

Model communication is the logical number of bytes in one server-to-client download plus one client-to-server upload, aggregated over the participating clients in a round. Serialization, encryption, packet headers, retransmission, and collective implementation overhead are excluded for every method. For the federated graph access study, auxiliary communication counts method-specific tensors outside model state under the implemented protocol: generated features and reconstruction gradients for FedSage+, cross-client identifiers and features for FedGCN, and routed intermediate representations for FedCog. Model traffic is reported per round and auxiliary traffic over the full run separately.

For the large-scale study, the same communication-accounting rule is used for all proposed and baseline runs, with client-local BatchNorm running buffers excluded from the transmitted model state. Equation (119) applies directly to FedASTRA at scale. The reported scale table and associated system audit use this common accounting rule for all methods.

Full-data ogbn-papers100M protocol without sampling. The scale implementation uses the official temporal split (Hu et al., 2020) and preprocesses all 111,059,956 nodes and 1,615,685,872 citations. One explicit node permutation is divided into eight contiguous client blocks whose sizes differ by at most one. Each node in an official split defines one rooted prediction example. The loader verifies that the official train, validation, and test root arrays contain no duplicate roots and are mutually disjoint. For root $v \in V_k$, client k receives $\{v\} \cup (N_G(v) \cap V_k)$. Each round traverses all 1,207,179 nodes in the official training split once. Validation and test cover every node in their official splits. The 200,000-incidence budget controls lossless streaming batches and never truncates a client-local rooted relation. All methods use 20 rounds and one local epoch over each client's training examples. Each streaming minibatch performs one optimizer step. For FedASTRA, it also performs one EMA target update. Its AdamW optimizer (Loshchilov & Hutter, 2019) uses learning rate 2×10^{-3}, weight decay 10^{-4}, encoder and readout dropout 0.2, and gradient-norm clipping at 2.0. The FedASTRA model uses context width 197, four diffusion bases, mask probability 0.40, predictive weight one, and target momentum 0.99 after each streaming optimizer update in every round. This width gives 976,504 trainable parameters. It is the largest integer width that stays below the 984,442-parameter ceiling used by the approximately one-million-parameter controls. Width 198 would contain 985,420 parameters. The study evaluates width 197 once and selects its reported checkpoint by validation under the reported complete-data protocol and budget.

The streaming representation uses disjoint copies of node occurrences across rooted examples. A paper appearing in two relations therefore has one copy in each relation, which prevents batch-

dependent paths between separate prediction examples. Every copied node belongs to exactly one relation column, and distinct columns have disjoint support. These two conditions give $D_v = I$ and $B^\top B = D_e$. The symmetric occurrence-space operator $P_{\mathrm{occ}} = BD_e^{-1}B^\top$ is consequently an orthogonal projection, as the following computation verifies for every streaming minibatch:

$$P_{\mathrm{occ}}^2 = BD_e^{-1}(B^\top B)D_e^{-1}B^\top = P_{\mathrm{occ}}. \tag{120}$$

The exponential series can therefore be summed exactly for every streaming minibatch:

$$\mathsf{K}_{\mathrm{occ}}(t) = e^{-t}I + (1 - e^{-t})P_{\mathrm{occ}}. \tag{121}$$

Corollary 11 (Exact spectral collapse in the occurrence protocol)**.** *For every scalar multiplier f on* $\mathrm{spec}(P_{\mathrm{occ}}) \subseteq \{0, 1\}$,

$$f(P_{\mathrm{occ}}) = f(0)I + [f(1) - f(0)]P_{\mathrm{occ}}. \tag{122}$$

Consequently, for any single $t > 0$, $\mathrm{span}\{I, \mathsf{K}_{\mathrm{occ}}(t)\} = \mathrm{span}\{I, P_{\mathrm{occ}}\}$. *The identity channel and one positive heat channel therefore span every linear spectral filter available in this two-point spectrum exactly on each streamed example in the complete reported ogbn-papers100M protocol.*

The corollary follows from the spectral projectors $I - P_{\mathrm{occ}}$ and P_{occ}, together with $1 - e^{-t} > 0$. Additional learned times in the scale model produce separately parameterized feature channels before nonlinear normalization and prediction, while leaving the linear spectral space of this specific occurrence operator unchanged. This closed form is specific to disjoint occurrence copies. Merging a node copy across rooted relations or placing one copy in multiple relation columns removes the projection condition. The general client-local transition in Equation (18) can therefore satisfy $P_k^2 \neq P_k$ and is evaluated with the certified recurrence in Appendix C.8. A reported float32 audit on the scale implementation gives $\max |P_{\mathrm{occ}}^2 - P_{\mathrm{occ}}| = 0$ and a maximum absolute difference of 1.91×10^{-6} between Equation (121) and a direct matrix exponential. The comparison includes Fed-MLP, Fed-SIGN, Fed-HGNN, Fed-HNHN, Fed-HyperGCN, and Fed-GCN (Frasca et al., 2020; Feng et al., 2019; Dong et al., 2020; Yadati et al., 2019; Kipf & Welling, 2017). Every baseline matches the dataset version, official split, client count, 20-round horizon, validation checkpoint rule, terminal test rule, and full-graph scope. Architectures retain their reported optimization settings. The Fed-HNHN entry uses two mean-aggregation layers, hidden width 16, dropout 0.5, Adam learning rate 0.01, and zero weight decay. Only trainable coefficients are aggregated, with each client's BatchNorm buffers kept out of the transmitted model state throughout every communication round.

The full-scale record uses eight logical clients. Table 17 reports wall time, throughput, peak device allocation, trainable-parameter size, and logical client–server payload for one reported execution of each method. Timing and throughput depend on the recorded hardware and parallelism.

D.5 Metrics and Additional Measurements

Every native method receives independently indexed client-local hyperedge tables, with direct cross-client node, edge, and source-hyperedge identifiers absent. The simple-graph study keeps the task split and node assignment fixed while recording each method's distinct information-access setting and communication. Node classification uses accuracy and macro-F1. Macro-F1 averages per-class F1 over the full declared label set, with zero contribution from a class absent from the test split. The evaluator computes τ_2 and τ_3 from the source hypergraph and the known experimental partition after the training assets have been fixed. These values describe pair and triple censoring and never enter model fitting, validation selection, or inference, remaining evaluator-only diagnostics throughout all reported runs. For independent verification, the native, controlled-truncation, and mechanism runs store learned heat times, Poisson mass assigned to walk lengths above two, the certified recurrence degree, the two evaluator-only truncation coordinates, parameter count, state-dictionary bytes, auxiliary payload, and logical two-direction model-state communication. Graph runs carry their prespecified experimental-condition identifiers and observed-data diagnostics. The scale run carries complete-dataset counts and system accounting. Its exact idempotent closed form on disjoint per-example node copies makes a finite Poisson-tail recurrence field inapplicable. Every reported test metric comes from one evaluation of its restored, validation-selected checkpoint.

The table generator accepts only complete reported panels. Primary finite summaries average the three client counts within each repetition and report sample SD across repetitions. Trivago contributes three repetitions for every method, Fed-HNHN contributes three on the remaining native

datasets, and the heat-width study uses three. Other established finite conditions contribute five. Paired studies retain common repetition keys. The ogbn-papers100M row records one complete execution with full assets and is reported separately from the repeated finite-experiment summaries.

E More Experiment Results

This section provides additional experimental results by reporting extended performance evaluations, controlled analyses of the proposed mechanisms, and detailed scalability measurements under the corresponding experimental protocols.

E.1 Performance on Native Hypergraphs

Table 8 reports the complete native-hypergraph results for each client count $K \in \{3, 6, 9\}$. Experiments on CoraCA, DBLP4k, and IMDB4k use five independent repetitions for all methods except Fed-HNHN, which uses three. Experiments on Trivago use three repetitions. The across-setting summary aggregates the 12 prespecified dataset-by-client-count effects while retaining the client-count factor. On CoraCA, the performance gap between FedASTRA and Fed-HGNN increases as the number of clients grows from 3 to 9, from 4.83% to 7.75% accuracy and from 6.54% to 9.84% Macro-F1. This trend is consistent with the increasing fragmentation of local observations under matched client inputs.

Table 8: Complete native-hypergraph results on different number of clients.

Method	$K = 3$		$K = 6$		$K = 9$	
	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
CoraCA						
FedASTRA	76.27 ± 0.54	74.33 ± 0.72	74.97 ± 0.63	72.79 ± 0.93	74.27 ± 0.73	71.52 ± 0.97
Fed-HGNN	71.44 ± 1.90	67.79 ± 2.12	67.97 ± 1.38	63.57 ± 1.50	66.52 ± 2.32	61.68 ± 3.02
Fed-HNHN	57.30 ± 0.82	54.40 ± 1.21	54.16 ± 2.46	49.08 ± 3.29	52.74 ± 1.13	47.77 ± 2.57
Fed-HyperGCN	67.38 ± 1.01	63.08 ± 1.64	63.84 ± 2.16	57.65 ± 2.56	60.85 ± 2.56	53.47 ± 3.07
Fed-AllDeepSets	45.90 ± 3.43	37.62 ± 4.62	36.54 ± 2.11	24.37 ± 2.86	36.45 ± 1.45	19.26 ± 6.62
Fed-AllSetTransformer	30.22 ± 0.00	6.63 ± 0.00	30.22 ± 0.00	6.63 ± 0.00	30.22 ± 0.00	6.63 ± 0.00
Fed-UniGCNII	63.57 ± 1.89	56.52 ± 5.53	58.89 ± 1.52	49.01 ± 3.67	56.16 ± 2.07	44.80 ± 3.52
Fed-ED-HNN	61.90 ± 1.51	57.27 ± 1.53	59.28 ± 2.38	52.11 ± 3.62	57.69 ± 2.22	50.46 ± 3.45
DBLP4k						
FedASTRA	78.52 ± 1.26	77.84 ± 1.04	79.75 ± 0.82	79.15 ± 0.75	79.40 ± 1.35	78.84 ± 1.24
Fed-HGNN	77.62 ± 1.48	76.98 ± 1.39	77.73 ± 1.62	77.00 ± 1.37	77.55 ± 1.30	76.96 ± 1.16
Fed-HNHN	72.23 ± 2.64	71.44 ± 2.66	68.43 ± 2.55	67.70 ± 2.67	68.08 ± 1.23	67.22 ± 0.97
Fed-HyperGCN	74.74 ± 1.29	73.43 ± 1.63	72.29 ± 0.60	70.88 ± 0.90	71.04 ± 1.85	70.06 ± 1.80
Fed-AllDeepSets	39.25 ± 0.68	31.74 ± 1.79	39.36 ± 0.72	31.96 ± 2.36	39.45 ± 1.18	33.63 ± 1.75
Fed-AllSetTransformer	30.87 ± 0.97	22.50 ± 2.61	30.82 ± 1.20	21.30 ± 1.64	31.28 ± 1.73	21.62 ± 1.93
Fed-UniGCNII	79.03 ± 2.12	77.85 ± 2.47	77.60 ± 2.49	76.18 ± 2.98	75.44 ± 3.10	73.92 ± 3.27
Fed-ED-HNN	71.92 ± 3.52	71.19 ± 3.46	72.30 ± 2.30	71.59 ± 2.40	72.09 ± 2.36	71.30 ± 2.37
IMDB4k						
FedASTRA	51.99 ± 1.85	51.92 ± 1.53	51.05 ± 1.53	50.88 ± 1.34	50.34 ± 0.92	50.37 ± 0.85
Fed-HGNN	50.14 ± 1.38	50.10 ± 1.39	48.36 ± 1.47	48.25 ± 1.46	48.05 ± 1.49	47.41 ± 2.11
Fed-HNHN	46.14 ± 0.62	44.75 ± 0.57	43.81 ± 2.13	41.84 ± 2.95	44.92 ± 1.93	42.54 ± 1.81
Fed-HyperGCN	47.36 ± 1.87	46.33 ± 2.22	43.26 ± 1.50	41.35 ± 2.30	41.81 ± 1.64	34.90 ± 5.71
Fed-AllDeepSets	45.82 ± 1.87	43.96 ± 4.83	37.73 ± 1.67	30.24 ± 1.83	37.42 ± 1.35	28.15 ± 3.67
Fed-AllSetTransformer	37.04 ± 0.03	18.04 ± 0.05	37.06 ± 0.05	18.06 ± 0.10	37.03 ± 0.11	18.21 ± 0.22
Fed-UniGCNII	51.12 ± 2.34	49.88 ± 3.45	37.09 ± 0.87	21.95 ± 5.33	36.66 ± 0.87	19.87 ± 3.96
Fed-ED-HNN	48.93 ± 1.98	48.61 ± 2.06	40.75 ± 0.94	30.57 ± 1.29	40.71 ± 0.93	30.58 ± 1.22
Trivago						
FedASTRA	33.55 ± 1.60	9.64 ± 0.35	27.70 ± 1.14	7.55 ± 0.32	26.01 ± 1.29	6.70 ± 0.15
Fed-HGNN	22.99 ± 0.72	5.00 ± 0.71	22.08 ± 0.77	4.05 ± 0.59	20.48 ± 0.87	3.75 ± 0.67
Fed-HNHN	29.10 ± 1.36	9.19 ± 0.19	24.87 ± 1.06	7.37 ± 0.35	23.08 ± 0.70	6.59 ± 0.24
Fed-HyperGCN	9.40 ± 0.65	0.15 ± 0.06	9.00 ± 0.02	0.11 ± 0.01	9.10 ± 0.21	0.13 ± 0.04
Fed-AllDeepSets	11.37 ± 0.11	0.30 ± 0.01	10.82 ± 0.15	0.28 ± 0.03	10.45 ± 0.22	0.24 ± 0.01
Fed-AllSetTransformer	8.99 ± 0.00	0.10 ± 0.00	8.99 ± 0.00	0.10 ± 0.00	8.99 ± 0.00	0.10 ± 0.00
Fed-UniGCNII	9.25 ± 0.23	0.16 ± 0.05	9.14 ± 0.09	0.16 ± 0.01	9.07 ± 0.10	0.13 ± 0.03
Fed-ED-HNN	9.54 ± 0.86	0.16 ± 0.09	9.05 ± 0.06	0.13 ± 0.03	8.99 ± 0.01	0.10 ± 0.00

E.2 Performance Against Baselines with Varying Information Access

Table 9 reports comparisons against federated graph learning baselines under different cross-client information-access settings, while Table 10 summarizes methods according to their available struc-

tural information and communication requirements. FedASTRA consistently outperforms identifier-free baselines across client counts and also surpasses FedSage+ despite the latter introducing additional exchanged information. In contrast, FedGCN (2-hop) and FedCog access explicit cross-client correspondence and therefore exploit a strictly richer observation setting. These results suggest that additional cross-client information is not universally beneficial: its value depends on whether the introduced information can be reliably exploited under the correspondence assumptions.

Table 9: Complete results under different cross-client information-access settings.

Method	$K = 3$		$K = 6$		$K = 9$	
	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Cora						
FedSage	70.35 ± 1.96	67.77 ± 1.67	63.18 ± 0.97	58.59 ± 0.87	58.65 ± 1.85	52.17 ± 2.81
FedGCN (0-hop)	74.07 ± 1.57	71.24 ± 1.86	65.90 ± 2.05	61.96 ± 3.34	62.07 ± 1.79	57.44 ± 2.91
FedSage+	69.67 ± 0.71	66.87 ± 0.30	63.11 ± 1.66	57.78 ± 1.63	61.02 ± 1.40	55.05 ± 3.65
FedGCN (2-hop)	84.05 ± 1.17	83.01 ± 1.23	84.05 ± 1.16	83.12 ± 1.20	84.12 ± 0.92	82.92 ± 0.78
FedCog	84.29 ± 1.50	83.15 ± 1.64	84.34 ± 1.06	83.35 ± 1.00	83.92 ± 0.78	82.73 ± 0.66
FedASTRA	77.54 ± 1.08	76.06 ± 1.31	74.45 ± 1.60	72.13 ± 1.81	73.33 ± 1.22	71.83 ± 1.31
CiteSeer						
FedSage	65.08 ± 0.96	59.56 ± 1.42	63.73 ± 0.74	57.00 ± 1.51	62.93 ± 1.72	55.67 ± 1.25
FedGCN (0-hop)	68.68 ± 1.50	61.60 ± 1.32	65.83 ± 0.56	59.06 ± 0.52	65.14 ± 1.80	58.10 ± 1.68
FedSage+	63.16 ± 1.17	57.75 ± 1.59	62.53 ± 1.66	56.48 ± 2.61	62.95 ± 1.62	57.69 ± 2.10
FedGCN (2-hop)	72.20 ± 1.44	65.64 ± 2.26	72.27 ± 1.53	64.88 ± 1.44	72.38 ± 1.50	64.85 ± 2.07
FedCog	72.21 ± 1.49	65.84 ± 1.77	72.38 ± 1.04	65.09 ± 1.03	72.51 ± 1.42	65.11 ± 1.94
FedASTRA	67.61 ± 2.32	63.62 ± 1.61	68.00 ± 1.86	63.67 ± 2.07	67.92 ± 1.88	63.65 ± 2.13
Facebook						
FedSage	72.63 ± 0.73	69.00 ± 1.08	65.64 ± 1.81	61.85 ± 2.39	60.41 ± 1.61	53.46 ± 4.16
FedGCN (0-hop)	75.14 ± 0.95	72.62 ± 0.74	67.87 ± 1.71	64.39 ± 2.24	64.30 ± 1.22	59.44 ± 2.83
FedSage+	68.17 ± 0.33	65.41 ± 0.64	61.85 ± 0.95	57.62 ± 1.32	58.56 ± 1.33	54.02 ± 1.99
FedGCN (2-hop)	83.21 ± 0.76	82.02 ± 0.87	82.84 ± 0.50	81.58 ± 0.67	82.73 ± 0.85	81.51 ± 1.04
FedCog	82.46 ± 0.69	81.24 ± 0.73	82.50 ± 0.84	81.20 ± 1.09	82.15 ± 0.94	80.85 ± 1.13
FedASTRA	76.88 ± 1.44	75.58 ± 1.60	75.57 ± 0.94	73.71 ± 0.86	75.12 ± 1.49	73.30 ± 1.76

Table 10: Summarized results of federated graph baselines under different cross-client information-access settings. Results are aggregated across client counts $K \in \{3, 6, 9\}$.

Method	Exchanged or accessed information	Accuracy (%)	Macro-F1 (%)
Cora			
FedASTRA	Model parameters only, with no direct cross-client identifiers	75.11 ± 0.99	73.34 ± 1.13
FedSage+	Locally generated boundary features and cross-client reconstruction gradients, with no direct identifiers	64.60 ± 1.11	59.90 ± 1.71
FedGCN (2-hop)	Cross-client node IDs, edges, and two-hop raw features	84.07 ± 1.03	83.02 ± 1.04
FedCog	Cross-client node IDs, edges, and intermediate embeddings	84.18 ± 1.04	83.08 ± 1.03
CiteSeer			
FedASTRA	Model parameters only, with no direct cross-client identifiers	67.84 ± 1.66	63.64 ± 1.83
FedSage+	Locally generated boundary features and cross-client reconstruction gradients, with no direct identifiers	62.88 ± 1.02	57.31 ± 1.74
FedGCN (2-hop)	Cross-client node IDs, edges, and two-hop raw features	72.28 ± 1.46	65.12 ± 1.68
FedCog	Cross-client node IDs, edges, and intermediate embeddings	72.37 ± 1.26	65.35 ± 1.44
Facebook			
FedASTRA	Model parameters only, with no direct cross-client identifiers	75.86 ± 1.02	74.20 ± 1.07
FedSage+	Locally generated boundary features and cross-client reconstruction gradients, with no direct identifiers	62.86 ± 0.71	59.02 ± 1.06
FedGCN (2-hop)	Cross-client node IDs, edges, and two-hop raw features	82.92 ± 0.61	81.70 ± 0.72
FedCog	Cross-client node IDs, edges, and intermediate embeddings	82.37 ± 0.73	81.10 ± 0.90

E.3 Validation on Proposed Analysis and Mechanism

Table 11 isolates the contributions of the adaptive diffusion basis and the masked prediction objective. Removing masked prediction $\lambda_{\text{pred}} = 0$ consistently reduces both Accuracy and Macro-F1 on CoraCA and DBLP4k, indicating that the predictive alignment component provides complementary representation learning signals. Increasing the heat basis width from $R = 1$ to $R = 4$ further improves performance when the predictive objective is enabled, while its effect is smaller without masked prediction. These results support the complementary roles of adaptive diffusion and masked representation alignment in FedASTRA.

Table 11: Ablation study of adaptive diffusion and masked prediction at $K = 6$. This study is separate from the heat-width sweep in Appendix D.

Dataset	Reported condition	Parameters	Accuracy (%)	Macro-F1 (%)
CoraCA	$R = 1,\ \lambda_{\text{pred}} = 0$	126,407	67.26 ± 1.48	62.94 ± 1.75
CoraCA	$R = 1,\ \lambda_{\text{pred}} = 1$	126,407	74.27 ± 0.56	71.96 ± 0.58
CoraCA	$R = 4,\ \lambda_{\text{pred}} = 0$	177,671	69.32 ± 0.69	65.31 ± 1.15
CoraCA	$R = 4,\ \lambda_{\text{pred}} = 1$	177,671	74.97 ± 0.63	72.79 ± 0.93
DBLP4k	$R = 1,\ \lambda_{\text{pred}} = 0$	55,684	77.54 ± 0.70	76.77 ± 0.78
DBLP4k	$R = 1,\ \lambda_{\text{pred}} = 1$	55,684	79.58 ± 0.67	78.99 ± 0.49
DBLP4k	$R = 4,\ \lambda_{\text{pred}} = 0$	106,372	78.23 ± 0.51	77.60 ± 0.56
DBLP4k	$R = 4,\ \lambda_{\text{pred}} = 1$	106,372	79.75 ± 0.82	79.15 ± 0.75

Table 12 reports the controlled-truncation results. The controlled partitions exhibit the expected ordering of retained higher-order co-membership mass. The observed pair/triple mass is 72.44/65.11%, 16.67/2.78%, and 9.68/1.10% for the low, reference, and high CoraCA arms, respectively. The corresponding DBLP4k values are 27.21/9.24%, 16.67/2.78%, and 16.33/2.70%. These results confirm that the three partition conditions differ in structural truncation while preserving client loads and label histograms across paired comparisons. The ordering applies to co-membership mass only. prediction performance also depends on feature and label alignment, so the truncation level alone does not determine Accuracy or Macro-F1.

Table 12: Controlled-truncation results at $K = 6$.

Dataset	Reported condition	Accuracy (%)	Macro-F1 (%)
CoraCA	Low-truncation	77.99 ± 0.58	76.40 ± 0.74
CoraCA	Reference	75.58 ± 1.33	73.27 ± 1.07
CoraCA	High-truncation	73.97 ± 0.80	71.34 ± 0.80
DBLP4k	Low-truncation	78.23 ± 1.76	77.54 ± 1.59
DBLP4k	Reference	79.46 ± 1.23	78.85 ± 1.13
DBLP4k	High-truncation	79.24 ± 1.49	78.66 ± 1.49

Table 13 and Table 14 summarize the two-client structural identification audit. The audit uses five fixed $K = 2$, $c = 1$ diagnostic partitions per source, exact hyperedge types, and unit weights. Held-out parent tokens mark cross-client rows and verify the oracle $\Phi_q(\mathcal{H}) = \sum_e \binom{|e|}{q}$ after the partition is fixed. Sorting fragment sizes gives both attainable endpoints. Labels are used only by the partition constructor, and no learner is trained. These evaluator partitions are independent of the reported $K \in \{3, 6, 9\}$ training partitions. Across all evaluated cells, no higher-order co-membership target is point identified, and the median relative endpoint width increases from 36.37% at $q = 2$ to 93.71% at $q = 5$. These results empirically support the predicted ambiguity caused by latent hyperedge correspondence, especially for higher-order structural targets.

Table 13: Two-client structural identification audit on four native datasets. For each order q, the audit contains 20 releases and 30 eligible strata. Relative endpoint width is $(\Phi_q^+ - \Phi_q^-)/\Phi_q^+$. By Corollary 2, half of this normalized width gives the corresponding minimax lower bound for observable point estimation of the same order-q mass.

q	Releases	Eligible strata	Point identified	Median relative width	q-mass floor	Oracle coverage
2	20	30	0/30	36.37%	18.18%	20/20
3	20	30	0/30	70.56%	35.28%	20/20
4	20	30	0/30	86.62%	43.31%	20/20
5	20	30	0/30	93.71%	46.85%	20/20

Table 14: Median absolute sharp-endpoint gaps over five two-client diagnostic partitions, measured in order-q tuple units. The values preserve the original dataset scale, complementing the normalized endpoint widths reported in Table 13.

Dataset	$q = 2$	$q = 3$	$q = 4$	$q = 5$
CoraCA	6,714	69,204	469,120	2,631,843
DBLP4k	3,752,114	983,644,998	2.1971×10^{11}	4.2294×10^{13}
IMDB4k	18,587	205,338	1,581,947	10,490,694
Trivago	410,593	3,263,206	23,744,190	209,428,691

To further verify the correspondence-invariant property of the learned representation, we conduct an evaluator-only permutation audit on the released relation columns. A separate CPU float32 audit permutes the 743 released relation columns of CoraCA client 0 at $K = 3$ and protocol seed 0 a total of 32 times. The fixed instance has 892 nodes, 1,499 incidences, a 128-dimensional projected input, hidden width 16, and $R = 4$. It uses neither labels nor training. The maximum differences in P_k, all heat-filtered features, and fixed-model logits are each zero, verifying the algebra used by Corollary 4 under this fixed implementation audit with shared numerical inputs.

Table 15 reports evaluator-only audits under relaxed observation assumptions. Type coarsening removes a matching restriction while preserving the released row inventory, and the numeric containment check confirms that the coarsened endpoint hull contains the exact hull in all audited cells. In contrast, status removal and one-sided deletion alter the released inventory, so finite compatible classes require additional public bounds beyond the exact equal-count endpoint formula. Training and model selection are absent from all 120 audit outcomes in this evaluator-only audit.

Table 15: Evaluator-only audit of relaxed observation assumptions on CoraCA and DBLP4k. The 40 cells consist of five fixed partitions and $q \in \{2, 3, 4, 5\}$ for each dataset.

Release change	Cells	Numeric endpoint	Audited conclusion
Merge observed type strata	40	Available	The coarsened endpoint hull contains the exact hull in 40/40 cells.
Remove local/cross marks	40	Bound required	Finite endpoints require a public cross-row count and support bound.
Delete 10% of one-side fragments	40	Bound required	Finite endpoints require public missing-row count, weight, and support bounds.

Figure 3 and Table 16 report the heat-width sweep over $R \in \{0, 1, 2, 4, 8\}$ on CoraCA, DBLP4k, and Trivago. CoraCA reaches 76.68% Accuracy and 74.35% Macro-F1 at $R = 2$, compared with 70.67% and 67.85% at $R = 0$. DBLP4k has closely grouped means over $R = 1, 2, 4$, within 0.09 Accuracy points and 0.10 Macro-F1 points, followed by lower means at $R = 8$. Trivago reaches its highest Accuracy of 27.66% at $R = 4$ and its highest Macro-F1 of 7.32% at $R = 8$. The latter change from $R = 4$ adds 0.30 Macro-F1 points, with parameters increasing from 472,224 to 818,336 and fitting time from 235.3 to 417.9 seconds.

The sweep identifies a dataset-dependent finite-capacity tradeoff. Moderate widths improve the attribute-only control, while additional channels increase parameters and runtime and can change the

two classification metrics differently. Spectral density describes approximation by a growing family of heat filters, but finite-sample accuracy also depends on capacity and fitting; therefore, density alone does not impose a monotone ordering on these empirical curves. Each width's runtime and parameter count follows the same measurement protocol described in Table 7.

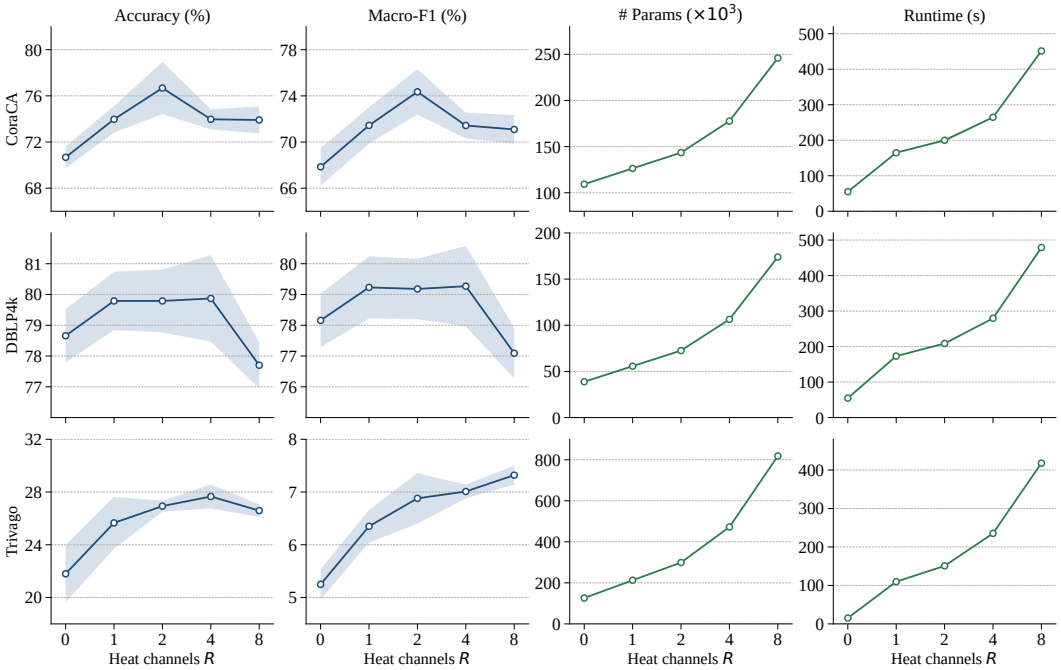


Figure 3: **Predictive utility and cost of the heat-basis width**. We fix $K = 6$ and $\lambda_{\text{pred}} = 1$ on all three datasets. The $R = 0$ control retains masked prediction and removes diffusion.

Table 16: Heat-basis width at $K = 6$ and $\lambda_{\text{pred}} = 1$. Parameter counts include the online model and predictive head. Runtime uses the synchronized fitting boundary described in Appendix D.

Dataset	R	Accuracy (%)	Macro-F1 (%)	Parameters	Runtime (s)
CoraCA	0	70.67 ± 0.94	67.85 ± 1.64	109,319	55.0 ± 1.7
	1	73.97 ± 1.16	71.44 ± 1.58	126,407	164.9 ± 0.3
	2	**76.68 ± 2.26**	**74.35 ± 1.96**	143,495	199.8 ± 4.7
	4	73.97 ± 0.87	71.43 ± 1.12	177,671	264.6 ± 1.8
	8	73.91 ± 1.17	71.09 ± 1.24	246,023	451.5 ± 2.9
DBLP4k	0	78.66 ± 0.87	78.16 ± 0.87	38,788	54.8 ± 0.4
	1	79.79 ± 0.95	79.23 ± 1.01	55,684	173.1 ± 2.1
	2	79.79 ± 1.02	79.18 ± 0.98	72,580	208.7 ± 2.4
	4	**79.87 ± 1.41**	**79.27 ± 1.31**	106,372	279.6 ± 0.3
	8	77.70 ± 0.75	77.09 ± 0.82	173,956	478.9 ± 2.8
Trivago	0	21.80 ± 2.16	5.25 ± 0.29	126,112	15.2 ± 0.7
	1	25.66 ± 1.97	6.35 ± 0.31	212,640	109.8 ± 0.1
	2	26.93 ± 0.42	6.88 ± 0.48	299,168	150.8 ± 0.3
	4	**27.66 ± 0.91**	7.01 ± 0.13	472,224	235.3 ± 0.6
	8	26.59 ± 0.47	**7.32 ± 0.18**	818,336	417.9 ± 0.1

E.4 Scalability on Large-scale Data

Table 17 reports the full-scale evaluation on the complete graph assets and official splits. It contains one audited execution rather than repeated runs. Repetition-based uncertainty is reserved for the finite studies. It records predictive quality, communication, wall-clock time, throughput, and peak

device allocation with complete root and citation coverage for every compared method. FedASTRA achieves the highest predictive performance in this full-data setting while maintaining parameter-only communication.

Table 17: Results on full-data ogbn-papers100M under the official split. Wall-clock training time covers 20 rounds and is reported in seconds (s). Throughput counts training roots per second. Peak memory is reported in gibibytes (GiB), where 1 GiB = 2^{30} bytes. Communication counts bidirectional trainable-parameter traffic, excluding client-local BatchNorm buffers for all methods.

Method	Parameters	Parameter traffic per round (MiB)	Wall-clock (s)	Roots/s	Peak GPU (GiB)
Fed-MLP	983,380	60.02	47.2	511,594	0.58
Fed-SIGN	984,442	60.09	47.6	506,887	0.68
Fed-HGNN	982,672	59.98	54.6	442,051	1.52
Fed-HNHN	34,672	2.12	312.5	77,259	0.54
Fed-HyperGCN	984,442	60.09	406.6	59,386	8.89
Fed-GCN	983,380	60.02	61.8	390,637	2.92
FedASTRA	976,504	59.60	1069.7	22,570	19.53

Complete assets contain 111,059,956 nodes, 1,615,685,872 source edges, and 3,231,371,744 bidirectional incidences. The official root split contains 1,207,179 training, 125,265 validation, and 214,338 test nodes. All requested neighborhoods are evaluated without truncation.

F AI USE STATEMENT

In accordance with the ICLR 2027 policy on the use of large language models, we disclose the following. For the manuscript, we used LLM assistants together with automated compiler diagnostics to help debug LaTeX formatting, to suggest grammatical corrections at the word and sentence level, and to check cross-references and the internal consistency of terminology, notation, and reported numbers; these tools produced suggestions only, and every change was reviewed and adopted or rejected by the authors. We also used an AI assistant to help condense and tighten the presentation of the analysis and proofs in Appendix, which we had derived beforehand, and the condensed statements and derivations were subsequently checked independently by three of the authors. We have reviewed all AI-assisted work, and the authors take full responsibility for the final content of this paper, including its text, claims, proofs, figures, and experimental results.

G REPRODUCIBILITY STATEMENT

The main text defines the observation model, latent-correspondence setting, learner information boundary, FedASTRA architecture, and federated training protocol used throughout the analysis. Appendices B–C provide the formal assumptions, identification and computational results, and complete proofs for the compatibility bounds, multi-client matching complexity, partition censoring, correspondence invariance, information-value analysis, local-information recovery, and population-risk convergence. Appendix D specifies the complete experimental and implementation protocols, including dataset construction, client partitioning, graph lifting, model architectures, optimization, masking and EMA updates, numerical heat evaluation, checkpoint selection, repetition counts, information-access conditions, communication accounting, and the full-scale ogbn-papers100M execution. Appendix E reports the complete results for the native-hypergraph benchmarks, graph baselines with different information-access regimes, controlled-censoring studies, structural identification audits, mechanism and heat-width ablations, and large-scale system measurements. The supplementary source package contains the manuscript source, simulation and plotting programs, and numerical outputs used to produce the reported figures and tables, allowing each reported experimental result to be traced to its corresponding program and result record.

H ETHICS STATEMENT

This work studies federated hypergraph learning under latent cross-client correspondence, with the goal of improving our understanding of what structural information is identifiable, usable, and computationally accessible when cross-client relation identities are unavailable. Our experiments use established public benchmark datasets and controlled federated partitions, and do not involve the

collection of new human-subject data or direct interaction with individuals. FedASTRA communicates model parameters only and does not require direct cross-client node, edge, or source-hyperedge identifiers under the prescribed protocol. Together, this work don't have any ethical issue.