LEARNING TO REFER: INFERENCE-TIME COLLABORATION IN DECENTRALIZED FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In decentralized federated learning, heterogeneous clients may develop complementary expertise, yet each query is typically evaluated only by the client where it arrives. To exploit this distributed expertise without global coordination, we formulate inference-time collaboration as budgeted local expert referral, a graph-constrained routing process in which a query can move only between neighboring clients and only visited clients contribute to its final prediction. `EdgeDRM` implements this framework by learning a density-based routing potential for each client: differences between neighboring potentials provide the comparisons used to select successive referrals, while local task models generate predictions that are aggregated along the resulting route. Both training and inference are decentralized, using only neighbor-local information without a central router or graph-wide client scoring. We establish estimation guarantees for the learned comparisons and characterize how local improvement and graph structure govern prior-weighted density gains from referral. Experiments on benchmark and real-world medical data show that `EdgeDRM` outperforms graph-matched decentralized baselines in most settings, with the clearest benefits under sparse connectivity.

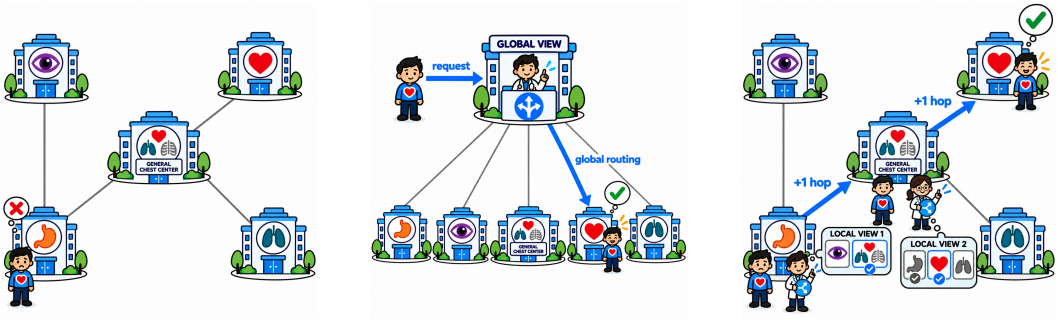
1 INTRODUCTION

Federated learning (FL) enables multiple clients to train a shared model without centralizing their raw data. Conventionally, a server distributes the model and aggregates locally computed updates (McMahan et al., 2017). In decentralized FL (Lalitha et al., 2019), clients form a communication graph whose nodes represent clients and whose edges specify which pairs can exchange model information directly, with no central server coordinating the network. The graph determines which clients can collaborate during training, but at inference a query is typically evaluated only by the client where it arrives. This inference limitation becomes consequential when client data distributions differ, a setting known as statistical heterogeneity. Consequently, a model may perform well on queries from its local distribution but poorly on those from other clients.

Such variation in local performance creates an opportunity for inference-time collaboration: a query that is poorly handled by its entry client may be better handled by another client. For instance, a patient may first visit a local hospital that lacks the expertise needed for the case, even though a better-suited specialist is available elsewhere in the hospital network. Reaching that specialist, however, may require a sequence of local referrals because each physician can refer the patient only to nearby or affiliated hospitals, which may in turn refer the patient onward. Each referral incurs delay and cost, while the referral path determines which specialists can ultimately be reached. We call the analogous inference process in decentralized FL *local expert referral*: a query enters at an arbitrary client, may traverse only a limited number of communication edges, and receives predictions only from the clients it visits. Here, a *referral* is one neighbor-to-neighbor transfer, *routing* is the process that selects successive referrals, and the *route* is the resulting sequence of visited clients. Unlike global expert selection, local expert referral is sequential and path-dependent: both the referral budget and the graph topology restrict which experts a query can reach. This raises our central question: *how can client specialization be exploited through graph-constrained local referral to improve on the entry client's prediction?* Figure 1 compares no routing, centralized routing, and local referral.

Existing work addresses only individual parts of this problem. Personalized and decentralized FL methods determine how clients collaborate during training, but typically leave each test query at

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.



(a) **No router.** (b) **Centralized router.** (c) **Proposed local router.**

Figure 1: Hospital analogy for inference-time routing. Gray lines indicate hospital pairs that can exchange information and know one another's capabilities, for example because they operate in the same region or health system. Without routing, a patient with a heart condition remains at an entry hospital whose expertise lies in gastroenterology (left). A centralized router sees all hospitals and dispatches the patient directly (center), but such global coordination is often infeasible because of administrative boundaries, privacy constraints, and geographic separation. In contrast, `EdgeDRM` (ours) makes stepwise local referrals (right): the gastroenterology hospital first routes the patient to a general chest center, whose thoracic expertise better matches the condition. The center then identifies the condition as cardiac and routes the patient to a dedicated cardiology hospital. Thus, each hop progressively improves the match between the patient and the hospital's expertise.

the client where it arrives (Sun et al., 2023; Zantedeschi et al., 2020; Dai et al., 2022; Lin et al., 2025; Li et al., 2025). Federated routing and mixture-of-experts methods exploit specialization through global client scores, learned routing gates, or expert subsets, but do not model budgeted neighbor-to-neighbor query movement (Wang et al., 2026b; Reisser et al., 2021; Farhat et al., 2025). Graph-based referral and search methods support local query movement, but learn from observed referral outcomes or graph-wide search supervision rather than from the static datasets held by neighboring FL clients (KhudaBukhsh et al., 2016; Baranchuk et al., 2019). Together, these lines leave open how to learn query–client compatibility from decentralized client data and compose neighbor-local comparisons into an effective prediction path. See Appendix A for details.

Addressing this gap requires overcoming two coupled difficulties specific to decentralized routing. First, query–client comparisons must be learned without the graph-wide normalization used by centralized routers such as `FedDRM` (Wang et al., 2026b). Second, neighbor-local comparisons must remain coherent when composed along a multi-hop referral path. We therefore propose `EdgeDRM`, a decentralized density-ratio routing framework. It avoids graph-wide normalization by learning each edge comparison through balanced endpoint source classification. To make these comparisons composable, `EdgeDRM` assigns each client one routing potential shared across its incident edges. Under a common representation, the resulting differences accumulate to the potential difference between the entry client and each reached client. We also characterize discrepancies caused by different local representations. At deployment, each client compares only its neighbors using its own query representation and stored neighbor coefficients, requiring neither a global score vector nor predictions from unvisited clients. This work makes four contributions:

- We formulate inference-time collaboration in decentralized FL as budgeted local expert referral, in which queries move only between neighbors and only visited clients contribute to prediction.
- We develop `EdgeDRM`, a neighbor-local framework that replaces graph-wide normalization with balanced endpoint source classification. One potential per client makes edge comparisons coherent across hops, and history-weighted aggregation uses only visited clients.
- We show that, under a shared representation, `EdgeDRM` learns increasingly accurate neighbor comparisons as training data and optimization steps increase. When locally better neighbors consistently guide the route, these comparisons yield cumulative gains in query–client compatibility and lead to the globally best-matched client given sufficient referral budget.
- Across benchmark, scaling, tiered-hospital, and preliminary real-world medical evaluations, `EdgeDRM` outperforms graph-matched decentralized baselines in most settings, with clear gains on sparse graphs but no consistent advantage under complete connectivity.

2 PROBLEM SETTING, DISTRIBUTIONAL MODELS, AND OBJECTIVE

Problem setting. Consider a K-class classification problem with m clients indexed by $V = [m] = \{1, 2, \ldots, m\}$. The clients communicate over a fixed graph $G = (V, E)$. An edge $\{i, j\} \in E$ allows clients i and j to exchange model information during training and refer queries directly to one another during inference. During inference, each client can communicate only with its neighbors in $\mathcal{N}(i)$ and cannot access the models or scores of non-neighboring clients. A query x first arrives at an entry client s. A *referral* transfers the query from its current client to one of its neighbors, and successive referrals form a path in G. The *movement budget* $L \in \mathbb{N}_0$ is an upper bound on the number of such neighbor-to-neighbor transfers after the query arrives. Thus, the query visits at most $L + 1$ clients, including the entry client s.

Each client i has a local dataset $D_i = \{(x_{ir}, y_{ir})\}_{r=1}^{n_i}$ consisting of samples drawn independently from a client-specific joint distribution $P_i^{X,Y}$. Let $N = \sum_{i=1}^m n_i$ and let $\rho_i = n_i/N$ denote client i's sample proportion. For each class $k \in [K]$, let $\pi_{ik} = P_i^Y(\{k\})$ denote the marginal probability of class k at client i, write $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$, and let P_{ik}^X denote the conditional feature distribution given $Y = k$. The feature marginal is then $P_i^X = \sum_{k=1}^K \pi_{ik} P_{ik}^X$. We allow the clients' distributions to differ, so that $P_i^{X,Y} \neq P_j^{X,Y}$ for some $i \neq j$.

Distributional models and conditional objective. The joint distribution factorizes as $P_i^{X,Y} = P_i^{Y|X} P_i^X$. Let g_θ denote a shared encoder and let $Z = g_\theta(X)$ be the resulting representation. We model $P_i^{Y|X}$ under the restriction that it depends on X only through Z. Then rather than modeling P_i^X directly, we model its induced representation distribution P_i^Z.

Conditional label distribution. For each client i, we model the conditional label distribution using the shared encoder g_θ and a private prediction head $\omega_i = (\alpha_i, \beta_i)$:

$$P_i^{Y|X}(\{k\} \mid x; \theta, \omega_i) = \frac{\exp\{\alpha_{ik} + \beta_{ik}^\top g_\theta(x)\}}{\sum_{c=1}^K \exp\{\alpha_{ic} + \beta_{ic}^\top g_\theta(x)\}}, \quad i \in V,\ k \in [K]. \tag{1}$$

For identifiability, we set $\sum_{k=1}^K \alpha_{ik} = 0$ and $\sum_{k=1}^K \beta_{ik} = 0$ for each client i. The shared encoder provides a common representation across clients, while the client-specific heads allow their conditional label distributions to differ.

Representation marginal distribution. To share information across clients, we jointly model the representation marginals $\{P_i^Z\}_{i=1}^m$ through a common-reference density-ratio model (DRM), following `FedDRM` (Wang et al., 2026b). Specifically, we assume that the measures $\{P_i^Z\}_{i=1}^m$ admit strictly positive Radon–Nikodym derivatives with respect to a common baseline distribution F and satisfy:

$$\varphi_i(z) := \{dP_i^Z/dF\}(z) = \exp\{b_i + \vartheta_i^\top z\}, \qquad i \in V, \tag{2}$$

and φ_i is the density of P_i^Z relative to F parameterized by b_i and ϑ_i. For identifiability, we take $F = P_1^Z$ and set $b_1 = 0$ and $\vartheta_1 = 0$. Because all clients share the same baseline distribution, it cancels from every pairwise comparison: $\log\{dP_j^Z/dP_i^Z\}(z) = \log\{\varphi_j(z)/\varphi_i(z)\} = (b_j - b_i) + (\vartheta_j - \vartheta_i)^\top z$. Each client is associated with one coefficient pair (b_i, ϑ_i). For a connected graph, differences across its edges determine these pairs relative to client 1.

From empirical likelihood to a conditional objective. The baseline distribution F is unknown, so an empirical-likelihood (EL) construction leaves it unrestricted and assigns masses to the distinct pooled representations $z_{ir} = g_\theta(x_{ir})$. These masses must normalize both F and every tilted distribution $\varphi_i\, dF$. To state the connection with source classification, let C denote the source client and define $P_\rho^{C,Z}$ by $P_\rho^C(\{i\}) = \rho_i$ and $P_\rho^{Z|C}(\cdot \mid i) = P_i^Z$. Under the normalized DRM, Bayes' rule gives:

$$s_i(z) := P_\rho^{C|Z}(\{i\} \mid z) = \frac{\rho_i \varphi_i(z)}{\sum_{c=1}^m \rho_c \varphi_c(z)} = \frac{\exp\{\log \rho_i + b_i + \vartheta_i^\top z\}}{\sum_{c=1}^m \exp\{\log \rho_c + b_c + \vartheta_c^\top z\}}.$$

Thus, $s_i(z)$ is the posterior probability that a representation z originates from client i. For fixed representations and finite slopes ϑ, maximizing EL over both the baseline masses and the free intercepts $b_2, \ldots, b_m$ is equivalent, up to a constant independent of these coefficients, to maximizing $\sum_{i,r} \log s_i(z_{ir})$ over the intercepts. Appendix B proves this identity, including existence of the

intercept optimum and the treatment of repeated representations. Profiling the masses alone does not give a pointwise equality at arbitrary intercepts. Motivated by this EL connection, we use the softmax expression as a conditional source model and combine it with the conditional-label model to define

$$\mathcal{L}(\theta, \omega, b, \vartheta) = \underbrace{\left[-\frac{1}{N}\sum_{i=1}^{m}\sum_{r=1}^{n_i} \log P_i^{Y|X}(\{y_{ir}\} \mid x_{ir}; \theta, \omega_i)\right]}_{\text{conditional-label loss}} + \underbrace{\left[-\frac{1}{N}\sum_{i=1}^{m}\sum_{r=1}^{n_i} \log s_i(z_{ir})\right]}_{\text{source-classification loss}}. \quad (3)$$

The conditional-label loss fits client-specific predictors, while the source-classification loss estimates density ratios by identifying each representation's source. For an edge $\{i, j\}$, the odds $s_j(z)/s_i(z)$ provide the comparison used to evaluate referral from client i to client j. These odds measure distributional compatibility, not prediction accuracy; predictions come from the conditional-label models at visited clients. Computing the global source-classification loss requires all client scores and a graph-wide normalization, which graph-local communication cannot provide. This limitation distinguishes decentralized FL from `FedDRM` and motivates a conditional source comparison on each communication edge.

3 Decentralized Training and Local Expert Referral

Section 2 defines the statistical model and its global objective. We now describe how clients train this model through neighbor communication and use the fitted quantities for referral and prediction.

3.1 Decentralized training with graph-local source likelihoods

Because the two terms in (3) impose different communication requirements, we train the model in two stages. Stage I trains the encoders and private prediction heads using the conditional-label loss and neighbor mixing. Stage II freezes the learned encoders, refines the private prediction heads, and estimates the density-ratio coefficients (b_i, ϑ_i) from graph-local source likelihoods.

For training, client i forms the local objective $\mathcal{L}_i$ in (17) from its local conditional-label loss (the first term in (3)) and one local source-classification loss for each incident edge, all evaluated on D_i. The graph objective in (18) then averages these client-local objectives with equal weights: $\mathcal{L}_G = m^{-1}\sum_i \mathcal{L}_i$. For the conditional-label component, this replaces the sample-proportion weight $\rho_i = n_i/N$ in (3) by $1/m$. This equal weighting is the canonical formulation in decentralized optimization and is widely used in decentralized FL, where doubly stochastic neighbor mixing gives each client equal weight (Lalitha et al., 2019; Koloskova et al., 2020; Sun et al., 2023). This weighting convention differs from the centralized setting.

Stage I: conditional-label model training. The model in Section 2 uses a shared encoder parameter θ, but decentralized optimization requires each client to maintain a local copy θ_i. For compactness, let $\psi_i = (\theta_i, \omega_i)$ collect the local encoder and prediction-head parameters at client i. Stage I runs for R_{sh} communication rounds. At round r, client i starts from $\psi_i^{(r)}$ and takes B_{sh} SGD steps on its empirical conditional-label loss using D_i. Let $\widetilde{\psi}_i^{(r)}$ denote the resulting encoder and prediction-head parameters. The graph G determines which pairs of clients can exchange these intermediate parameters. Because no central server aggregates all client updates, each client instead combines its own intermediate parameters with those received from its neighbors. A mixing matrix W determines the weights in this local aggregation. Specifically, let W be symmetric and doubly stochastic, with $W_{ij} = 0$ whenever $j \notin \{i\} \cup \mathcal{N}(i)$. Client i then updates its parameters as

$$\psi_i^{(r+1)} = \sum_{j \in \{i\} \cup \mathcal{N}(i)} W_{ij}\widetilde{\psi}_j^{(r)}. \quad (4)$$

For each edge $\{i, j\} \in E$, we use the Metropolis weight $W_{ij} = 1/(1 + \max\{\deg(i), \deg(j)\})$ and set $W_{ii} = 1 - \sum_{j \neq i} W_{ij}$. Neighbor mixing is an established decentralized optimization mechanism (Koloskova et al., 2020), but finitely many rounds need not produce exact consensus, so the learned encoders may differ across clients.

Stage II: frozen-encoder density-ratio fitting. Stage I uses local SGD and neighbor mixing to optimize the client-averaged conditional-label objective described above. Stage II addresses the

remaining source-classification loss, which learns the density ratios by predicting the source client C from a representation Z. In centralized FL, a server can coordinate the scores of all m clients required to evaluate the source probabilities $s_i(z)$. In the decentralized setting, however, each client communicates only with its neighbors and therefore cannot evaluate the graph-wide normalization in $s_i(z)$. Because each referral traverses one edge, deciding whether to move the query requires only a comparison between the two endpoints. We therefore replace the global m-client source-classification problem with a binary source-classification problem on each edge.

For each edge $e = \{i, j\} \in E$ with $i < j$, we form a balanced binary source-classification problem by assigning equal weight to the two endpoint mean losses. Its posterior probability that z comes from client j rather than client i is $q_{ij}^{\mathrm{bal}}(z) := \varphi_j(z)/\{\varphi_i(z) + \varphi_j(z)\}$, with

$$\operatorname{logit} q_{ij}^{\mathrm{bal}}(z) = (b_j - b_i) + (\vartheta_j - \vartheta_i)^\top z. \tag{5}$$

This binary problem is the source-classification term in (3) restricted to the endpoints $\{i, j\}$, and it learns the coefficient differences needed to compare them. Because client i uses the same coefficient pair (b_i, ϑ_i) on all its incident edges, the graph links these pairwise comparisons through their shared endpoints. The two endpoint losses receive equal weight, irrespective of n_i and n_j. Thus, the sample-count prior is not included in the training logit. The exact graph-local objective is given in Appendix B.4. After Stage I, clients freeze their encoders and fit their private heads and density-ratio coefficients using local conditional-label and incident-edge source losses. Only sample counts, density-ratio coefficients, and their gradient contributions are exchanged across edges. Encoders, private heads, and raw data remain local. These fitted quantities drive referral and prediction.

3.2 Budgeted local expert referral & Prediction using the referral history

Budgeted local expert referral. For a query x, let $\Delta_{ij}(x) := \log(\rho_j/\rho_i) + \operatorname{logit} q_{ij}^{\mathrm{bal}}(g_\theta(x))$ denote the population pairwise log-odds obtained by adding the sample-count prior to the learned balanced comparison. This quantity provides the criterion for referring the query from client i to client j. At inference, client i substitutes the fitted density-ratio coefficients and its own fitted representation into this quantity, giving the plug-in estimator

$$\widehat{\Delta}_{ij}(x) := \log(\rho_j/\rho_i) + (\widehat{b}_j - \widehat{b}_i) + (\widehat{\vartheta}_j - \widehat{\vartheta}_i)^\top g_{\widehat{\theta}_i}(x). \tag{6}$$

Hence training estimates the density-ratio term without sample-count weighting, and inference adds $\log(\rho_j/\rho_i) = \log(n_j/n_i)$ exactly once. Applying these fitted endpoint comparisons sequentially produces a query-dependent route through the graph. For a query x entering at client s under movement budget L, the realized route is

$$\mathcal{R}_L(x, s) = (\widehat{s}_0 = s, \widehat{s}_1, \ldots, \widehat{s}_T), \quad T \le L, \quad \{\widehat{s}_{t-1}, \widehat{s}_t\} \in E. \tag{7}$$

The route construction is detailed in Algorithm 1. We describe its intuition below. During referral, x is passed between neighboring clients together with the visited-client set, accepted edge scores, and remaining movement budget. At client $\widehat{s}_t$, the router scores each unvisited neighbor using (6) and selects the largest score, with ties broken deterministically. The query moves to the selected neighbor only if this score exceeds η and otherwise terminates. Evaluating a candidate requires only the current client's representation of x and the stored sample counts and density-ratio coefficients of the two endpoints. The candidate need not inspect the query. The procedure uses neither the query label nor information about prediction correctness. It returns the visited route and accepted edge scores for the final prediction rule.

Algorithm 1: Budgeted local expert referral.

Input: Query x, entry client s, threshold $\eta \ge 0$, and movement budget L.
$\widehat{s}_0 \leftarrow s$, $V_{\mathrm{vis}} \leftarrow \{s\}$, $T \leftarrow 0$
while $T < L$ **do**
  $U \leftarrow \mathcal{N}(\widehat{s}_T) \setminus V_{\mathrm{vis}}$
  if $U = \varnothing$ **then break**
  $j^\star \leftarrow \arg\max_{j \in U} \widehat{\Delta}_{\widehat{s}_T, j}(x)$
  if $\widehat{\Delta}_{\widehat{s}_T, j^\star}(x) \le \eta$ **then break**
  Send the packet and accepted score to $j^\star$
  $T \leftarrow T + 1$, $\widehat{s}_T \leftarrow j^\star$
  $V_{\mathrm{vis}} \leftarrow V_{\mathrm{vis}} \cup \{j^\star\}$
end while
Output: $(\widehat{s}_0, \ldots, \widehat{s}_T)$ and the accepted scores.

Prediction using the referral history. Because referral scores reflect distributional compatibility rather than prediction accuracy directly, earlier visited clients may still provide useful complementary predictions. We therefore aggregate the predictions from all clients on $\mathcal{R}_L(x, s)$. Denote the fitted probability vector at client i as $f_i(x) := \left(P_i^{Y|X}(\{k\} \mid x; \widehat{\theta}_i, \widehat{\omega}_i)\right)_{k=1}^K$. For each

visited client $\widehat{s}_t$, let $u_0(x,s) = 0$ and $u_t(x,s) = \sum_{\ell=1}^{t} \widehat{\Delta}_{\widehat{s}_{\ell-1},\widehat{s}_\ell}(x)$ denote the cumulative accepted edge score from the entry client to $\widehat{s}_t$ for $t = 1, \ldots, T$. We set $(w_0(x,s), \ldots, w_T(x,s)) := \mathrm{softmax}(u_0(x,s), \ldots, u_T(x,s))$. The final prediction is:

$$f_{\mathrm{mix}}(x,s) = \sum_{t=0}^{T} w_t(x,s) f_{\widehat{s}_t}(x), \qquad \widehat{y} = \arg\max_{k \in [K]} f_{\mathrm{mix},k}(x,s).$$

4 Theoretical Results

We establish theoretical guarantees for coefficient estimation and referral gain in `EdgeDRM`. Our results characterize the accuracy and asymptotic normality of parameter estimation (Section 4.1) and referral gains (Section 4.2). Proofs and additional results are given in Appendix C.

4.1 How accurately can we estimate the parameters?

We first study the performance of parameter estimation in stage II. To separate coefficient estimation from representation learning, consider a common representation g_θ from Section 2, fixed independently of the coefficient-fitting samples. Let ζ collect the free density-ratio coefficients (ϑ_i, b_i), $i \neq 1$, and private-head parameters (α_i, β_i) under the normalizations in Section 2.

We analyze R updates of Stage-II optimization on $\mathcal{L}_G$ in (18), holding $(\vartheta_1, b_1) = 0$ fixed and using unbiased mini-batch gradients with step sizes $\eta_r = 2/\{\mu(r + r_0)\}$ for $r = 0, \ldots, R-1$. Here $0 < \mu \leq 1$ is a curvature constant and $r_0 > 0$ is a fixed offset. We refer to Appendix C.1 for further details on conditions on these constants. The following theorem bounds the statistical and optimization errors of the Stage-II output and establishes its asymptotic normality.

Theorem 1. *Let $\zeta^\star$ denote the true coefficient vector of the models in Section 2, and suppose $n_i/N \to \rho_i^* > 0$. Assume the coefficients are identifiable and assume the population Hessian at $\zeta^\star$ is positive definite and, with probability tending to one, $\mu I \preceq \nabla^2 \mathcal{L}_G(\zeta) \preceq \mu^{-1} I$, throughout a convex region containing a finite empirical minimizer and the optimization trajectory, for a constant $0 < \mu \leq 1$ independent of N. Under the remaining regularity conditions in Appendix C.1, the Stage-II output $\zeta^{(R)}$ from the specified SGD procedure satisfies, as $N, R \to \infty$,*

$$\|\zeta^{(R)} - \zeta^\star\|_2 = O_p\Big(N^{-1/2} + R^{-1/2}\Big). \tag{8}$$

If, in addition, $R/N \to \infty$, then for some Σ_G specified in Appendix C.1,

$$\sqrt{N}\{\zeta^{(R)} - \zeta^\star\} \xrightarrow{d} \mathcal{N}(0, \Sigma_G), \tag{9}$$

The first term in (8) is the statistical error from N training samples. The second is the optimization error after R updates. Thus, when $R/N \to \infty$, the optimization error is dominated by $N^{-1/2}$, and the output has the asymptotic distribution in (9).

We note that, under the stated models, changing client sample proportions does not change the population minimizer $\zeta^\star$. The proportions need only satisfy $n_i/N \to \rho_i^* > 0$, without a specified convergence rate. The coefficient-estimation bound in Theorem 1 implies a bound on the edge-comparison errors used for referral. In particular, for any fixed x with finite $g_\theta(x)$, we obtain

$$\max_{\{i,j\} \in E} |\widehat{\Delta}_{ij}(x) - \Delta_{ij}(x)| = O_p\Big(N^{-1/2} + R^{-1/2}\Big). \tag{10}$$

This comparison-error bound yields the referral guarantee in Section 4.2. We remark that the result applies to coefficient fitting with a common, independently fixed encoder.

4.2 How far and how much can referral improve?

Fix a query x and entry client s, and consider the route $\mathcal{R}_L(x,s)$ in (7). Using the population coefficients of Section 2, write $h_i(x) = \log \rho_i + b_i + \vartheta_i^\top g_\theta(x)$ and $h^\star(x) = \max_{i \in V} h_i(x)$. Thus, $\Delta_{ij}(x) = h_j(x) - h_i(x)$, and $\Gamma_L(x,s) = h_{\widehat{s}_T}(x) - h_s(x)$ is the referral gain. Under the DRM, this gain is the log ratio of the terminal and entry clients' prior-weighted densities. Our referral guarantees rely on the following intuitive structural properties.

Minimum local improvement. Assume there is $\gamma > 0$ such that every nonoptimal client has a neighbor whose potential is at least γ higher:

$$\max_{j \in \mathcal{N}(i)} \{h_j(x) - h_i(x)\} \geq \gamma \quad \text{whenever } h_i(x) < h^\star(x). \tag{11}$$

Thus, a client with no higher-potential neighbor must be globally optimal.

Ascent depth. For $\epsilon \geq 0$, let H_ϵ be the longest strictly ascending path length from s, with each step choosing a neighbor within 2ϵ of the highest neighbor potential. Then $H_\epsilon \leq m - 1$, and under the conditions below, it bounds the number of accepted referrals.

Theorem 2. *Under the conditions of Theorem 1, use $\zeta^{(R)}$ for referral at a fixed query x with finite $g_\theta(x)$. For any $\delta \in (0, 1)$, there exists a constant $C_\delta > 0$, independent of N and R, such that the following holds with $\epsilon = C_\delta(N^{-1/2} + R^{-1/2})$. Suppose* (11) *holds, $\eta \geq \epsilon$, and $\gamma > \eta + \epsilon$. For all sufficiently large N, R, with probability at least $1 - \delta$, Algorithm 1 with movement budget $L \in \mathbb{N}_0$ makes $T \leq \min\{L, H_\epsilon\}$ moves and satisfies*

$$\Gamma_L(x, s) \geq (\gamma - 2\epsilon) \min\{L, H_\epsilon\}. \tag{12}$$

On the same event, $L \geq H_\epsilon$ guarantees that the route reaches a maximizer.

The guarantee in (12) lower-bounds $\Gamma_L(x, s)$, the improvement in source compatibility, by a product of two terms. The first term, $\gamma - 2\epsilon$, lower-bounds the population gain per accepted move. Larger per-step gains allow substantial improvement with fewer referrals. The second term, $\min\{L, H_\epsilon\}$, is the effective step count. A smaller ascent depth H_ϵ requires less budget to guarantee reaching a global maximizer, while with $L < H_\epsilon$ the route may exhaust its budget first. Note that improvement in source compatibility does not by itself imply improved predictive performance. Appendix C.3 provides the proof and analyzes ascent depth and referral gain in several common graph types.

5 Experiments

5.1 Experimental setup and evaluation

Benchmarks and topologies. We first evaluate on CIFAR-10, CIFAR-20, and CIFAR-100 benchmarks under 2 non-IID partitions. In the shards-per-client (SPC) partition based on (McMahan et al., 2017), each client receives five equal-sized, label-homogeneous shards for CIFAR-10 and 25 for CIFAR-20/100. In the Dirichlet partition, each client's class proportions are drawn from a Dirichlet distribution with concentration 0.3. The main benchmark has eight clients. We consider four communication topologies: clients are connected as a Path, a Cycle, a connected random graph in which every client has four neighbors (RR4), or a Complete graph. These are standard decentralized communication topologies and are ordered by increasing connectivity. For further details on the benchmarks, graph topologies, and data construction, see Appendix D.1.

Training and inference. We train randomly initialized ResNet-18 models for 800 communication rounds using `DFedAvgM` (Sun et al., 2023), freeze each local encoder, and then fit the personalized classifier heads and routers for 200 rounds. The routing objective weights the two endpoints of every edge equally. At inference, `EdgeDRM` starts from a uniformly sampled non-source entry, excludes revisits, and averages predictions from visited experts using temperature-one score weights.

Comparisons and reporting. At test time, each query from source client s is assigned uniformly at random to an entry client other than s. `Local` trains one model per client using only local data and predicts with the model at that entry. The decentralized baselines `DFedAvgM` (Sun et al., 2023), `DisPFL` (Dai et al., 2022), and `FedSPD` (Lin et al., 2025) also predict with their learned model at that entry. By contrast, we include centralized `FedDRM` (Wang et al., 2026b) only as an illustrative reference because its client-selection head scores all clients globally and therefore cannot operate under the neighbor-only communication constraints of our decentralized setting. `EdgeDRM-NR` uses the same trained model but keeps each query at its entry client.

5.2 Main results

CIFAR benchmark results. The main results in Table 1 yield three main observations. First, `EdgeDRM` is the strongest graph-constrained method in most settings. It ranks first in all 18 dataset–partition configurations on the three sparse graphs and in 19 of the 24 graph–configuration pairs

Table 1: **Eight-client decentralized CIFAR benchmark** reporting system accuracy (%) across three datasets and two non-IID partitions. `Local` and centralized `FedDRM` are illustrative references. Bold cells with blue shading mark the best decentralized result within each topology and setting. All entries are means ± sample SD across 3 seeds.

Group	Method	CIFAR-10		CIFAR-20		CIFAR-100	
		Dir-0.3	5-SPC	Dir-0.3	25-SPC	Dir-0.3	25-SPC
Reference	`Local`	16.53 ± 3.56	18.27 ± 1.28	13.30 ± 0.17	11.43 ± 0.60	4.10 ± 0.13	2.58 ± 0.08
	`FedDRM`	61.93 ± 1.61	58.96 ± 2.34	37.03 ± 0.99	40.09 ± 0.89	26.62 ± 0.62	30.86 ± 0.75
PATH	`DFedAvgM`	28.55 ± 7.62	27.41 ± 0.21	23.94 ± 0.03	21.47 ± 1.04	9.28 ± 0.61	7.03 ± 0.44
	`DisPFL`	25.91 ± 6.78	25.27 ± 2.01	24.46 ± 1.04	21.91 ± 1.02	9.54 ± 0.78	5.87 ± 0.36
	`FedSPD`	18.44 ± 4.78	19.31 ± 1.23	15.08 ± 0.65	14.12 ± 1.05	4.51 ± 0.34	4.25 ± 0.43
	`EdgeDRM-NR`	22.00 ± 4.85	21.53 ± 1.18	21.57 ± 0.35	18.51 ± 1.79	6.89 ± 0.62	4.21 ± 0.21
	`EdgeDRM`	**33.08 ± 3.53**	**32.88 ± 1.00**	**24.87 ± 0.33**	**23.97 ± 1.12**	**11.19 ± 0.58**	**11.15 ± 0.34**
CYCLE	`DFedAvgM`	30.34 ± 6.23	27.63 ± 2.22	25.00 ± 0.15	23.19 ± 1.11	10.39 ± 0.72	7.93 ± 0.45
	`DisPFL`	27.54 ± 6.33	26.55 ± 1.68	25.37 ± 1.17	23.40 ± 1.54	9.92 ± 0.59	6.47 ± 0.24
	`FedSPD`	19.41 ± 4.06	20.24 ± 0.33	15.54 ± 1.01	14.78 ± 0.79	4.93 ± 0.32	4.61 ± 0.32
	`EdgeDRM-NR`	22.92 ± 4.71	21.95 ± 0.75	21.99 ± 0.22	19.04 ± 1.62	7.49 ± 0.71	4.32 ± 0.39
	`EdgeDRM`	**35.20 ± 2.30**	**34.79 ± 0.63**	**25.99 ± 0.59**	**25.68 ± 0.94**	**12.40 ± 0.63**	**12.08 ± 0.42**
RR4	`DFedAvgM`	40.10 ± 1.86	37.58 ± 1.05	28.83 ± 0.73	28.14 ± 1.41	13.92 ± 0.90	12.14 ± 0.54
	`DisPFL`	28.39 ± 6.96	27.68 ± 1.66	26.24 ± 1.34	24.89 ± 1.23	10.67 ± 1.12	6.35 ± 0.54
	`FedSPD`	20.33 ± 3.36	20.47 ± 1.49	14.75 ± 0.36	13.73 ± 1.11	4.88 ± 0.02	4.47 ± 0.49
	`EdgeDRM-NR`	25.35 ± 4.54	23.44 ± 1.23	22.72 ± 0.60	20.09 ± 1.81	8.05 ± 0.40	5.53 ± 0.18
	`EdgeDRM`	**50.12 ± 2.21**	**46.19 ± 1.17**	**29.91 ± 0.79**	**31.51 ± 0.52**	**17.59 ± 0.59**	**20.01 ± 0.52**
COMPLETE	`DFedAvgM`	55.04 ± 1.47	**52.91 ± 1.43**	**32.62 ± 0.59**	**36.10 ± 0.38**	**21.02 ± 0.48**	**23.81 ± 1.08**
	`DisPFL`	33.19 ± 5.44	30.14 ± 0.34	27.29 ± 1.10	26.72 ± 1.26	12.81 ± 0.84	9.36 ± 0.01
	`FedSPD`	26.03 ± 5.08	25.41 ± 4.80	14.47 ± 0.82	14.75 ± 0.55	5.82 ± 0.49	5.82 ± 0.25
	`EdgeDRM-NR`	27.85 ± 4.03	25.80 ± 0.93	23.64 ± 0.46	21.12 ± 1.44	9.31 ± 0.37	6.88 ± 0.23
	`EdgeDRM`	**56.62 ± 2.87**	52.84 ± 1.86	31.31 ± 1.02	34.12 ± 0.31	20.02 ± 0.63	23.65 ± 0.67

overall. Against the second-best `DFedAvgM`, its mean gain across the sparse graphs is 4.21 percentage points. Second, the COMPLETE graph, equivalent to the central server case, `EdgeDRM` is effectively tied with `DFedAvgM`, averaging only 0.49 points lower across settings. So full graph connectivity narrows the advantage of `EdgeDRM`. Both methods nevertheless remain below centralized `FedDRM` in all six settings, yielding the gap between routing designed for centralized setting. Third, referral itself provides a large and universal gain. `EdgeDRM` outperforms `EdgeDRM-NR` in all configurations, by 3.30–28.77 points and 11.87 points on average.

Where the referral gain comes from. We conduct an ablation to isolate the effect of the routing rule. Every method uses the terminal expert's hard prediction, without averaging predictions across visited experts. Besides `EdgeDRM` and `EdgeDRM-NR`, we also compare `Count-only`, which triggers a referral when $\log(n_j/n_i) > 0$, and `Random`, which samples uniformly among unvisited neighbors. The improvement over `EdgeDRM-NR` and `Random` shows that selecting suitable neighbors is more effective than null/arbitrary referral. Comparison with `Count-only` further shows that the gain comes from the trained router rather than solely from differences in local sample sizes.

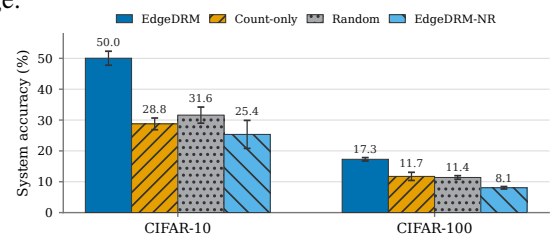


Figure 2: **Fixed-expert hard-terminal routing ablation** on CIFAR-10 and CIFAR-100 with Dirichlet-0.3 partitions and RR4.

Scaling in number of clients. We test whether the behavior observed with few clients persists as the client population grows substantially. Table 2 shows that `EdgeDRM` remains the strongest decentralized method in both settings. We note that, because the total CIFAR-10 data pool is fixed, increasing the number of clients reduces the average local sample sizes, which likely contributes to the lower accuracy relative to the results in Table 1.

Table 2: CIFAR-10/RR4 system accuracy with 32 clients (%; mean ± SD over three seeds).

	Reference	Decentralized					Reference
Setting	`Local`	`DFedAvgM`	`DisPFL`	`FedSPD`	`EdgeDRM-NR`	`EdgeDRM`	`FedDRM`
Dir-0.3	16.45 ± 1.15	27.79 ± 3.25	23.71 ± 2.65	17.82 ± 0.69	21.52 ± 1.50	**30.06 ± 1.16**	38.45 ± 0.52
5-SPC	18.11 ± 0.47	27.92 ± 1.22	24.87 ± 0.95	20.54 ± 0.75	22.76 ± 1.20	**31.70 ± 0.39**	35.96 ± 0.52

5.3 Referral traffic

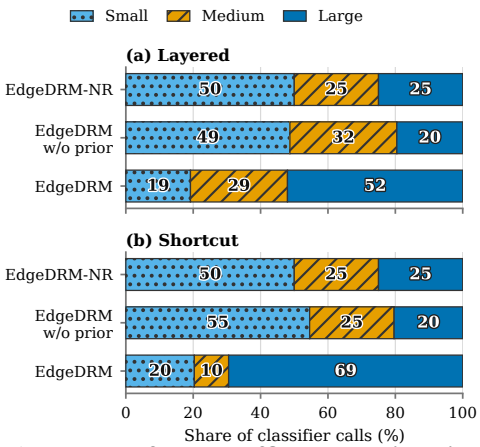


Figure 3: **Referral traffic by hospital tier on two topologies**. Bars show the mean share of visited-classifier calls over 3 seeds.

We conduct an ablation study to examine how a sample-count prior changes referral traffic across hospital tiers. We use 32 clients: 16 small, eight medium, and eight large hospitals holding 250, 1000, and 4000 training examples, respectively. We evaluate two graph structures. In the layered graph, each tier forms a cycle, every small hospital connects to one medium hospital, and medium and large hospitals are matched one-to-one. Thus, a small hospital can reach the large tier only through the medium tier, requiring at least two hops. The shortcut graph retains all edges and adds one direct small–large edge for each small hospital, making a large hospital reachable from every small hospital in one hop. `EdgeDRM w/o prior` applies the same history-weighted referral rule after removing the sample-count offset $\log(n_j/n_i)$ and serves as a sample-size-agnostic comparison. In Figure 3, `EdgeDRM w/o prior` largely preserves the `EdgeDRM-NR` tier proportions, whereas `EdgeDRM` substantially increases large-hospital traffic, with an even stronger concentration when shortcut edges provide direct access to large hospitals. This pattern captures a realistic healthcare trade-off: the sample-count prior is a plausible way to represent the stronger expertise of larger hospitals, but systematically trusting those hospitals concentrates demand and increases their service burden. The shortcut amplification further highlights the need for tiered referral systems: allowing patients to bypass intermediate hospitals and go directly to the largest regional centers can overwhelm scarce specialist capacity, whereas staged referrals distribute demand across levels of care.

5.4 Real medical data

Table 3: **ISIC2019 balanced accuracy** (%; three-seed mean). Parentheses show the `EdgeDRM` change from `DFedAvgM`.

Graph	`DFedAvgM`	`EdgeDRM`
Path	28.45	**28.62** (+0.18)
Cycle	27.20	**28.35** (+1.15)
RR3	27.45	**27.76** (+0.31)
Complete	**28.93**	28.79 (-0.14)

To assess `EdgeDRM` beyond synthetic benchmarks, we conduct a real-world skin-lesion classification study on ISIC2019 using the six source-center assignments distributed with FLamby (du Terrail et al., 2022). Unlike the CIFAR benchmarks, where client heterogeneity is imposed through synthetic label-skew partitions, these clients follow source and acquisition metadata from real medical image collections. Every test image is evaluated at all six entry clients, and we report pooled balanced accuracy, which weights recall equally across classes to account for the pronounced class imbalance in ISIC2019. Appendix D.3 provides the data provenance, split audit, graph, training, and evaluation details. `EdgeDRM` matches or improves upon the graph-matched `DFedAvgM` baseline on three of the four topologies. On Complete, it is 0.14 points lower, consistent with previous results, where full connectivity likewise narrows or eliminates the advantage of learned local referral.

6 Conclusion and Limitations

`EdgeDRM` formulates inference-time collaboration in decentralized federated learning as budgeted local expert referral. A query moves only between neighbors on the communication graph, and only visited experts contribute to its final prediction. Neighbor-local density-ratio comparisons make referral trainable without a central router or global expert scoring, while our analysis relates referral gains to comparison accuracy and graph structure.

Across six CIFAR settings, `EdgeDRM` improves over graph-matched `DFedAvgM` in all 18 sparse-graph comparisons, with a mean gain of 4.21 percentage points. The Complete graph gives `EdgeDRM` its highest absolute accuracy in every setting, but its mean advantage over `DFedAvgM` disappears (-0.49 points). Together, these results establish graph-constrained referral as a distinct inference setting and a useful mechanism under sparse connectivity, while showing that greater connectivity can improve absolute accuracy without increasing the relative benefit of learned referral. Closing the gap between the theory and learned neural representations remains future work.

AI Assistance Disclosure

In this work, we used generative AI tools to aid and polish writing, produce figures and tables, draft portions of the manuscript, and check mathematical proofs. The authors reviewed all AI-assisted work and take responsibility for the final content, including all text, claims, proofs, and reported results produced with the aid of generative AI.

References

Dmitry Baranchuk, Dmitry Persiyanov, Anton Sinitsin, and Artem Babenko. Learning to route in similarity graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 475–484. PMLR, 2019. URL http://proceedings.mlr.press/v97/baranchuk19a.html.

Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC), 2018. URL https://doi.org/10.1109/ISBI.2018.8363547.

Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4587–4604. PMLR, 2022. URL https://proceedings.mlr.press/v162/dai22b.html.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, Boris Muzellec, Constantin Philippenko, Santiago Silva, Maria Telenczuk, Shadi Albarqouni, Salman Avestimehr, Aurélien Bellet, Aymeric Dieuleveut, Martin Jaggi, Sai Praneeth Karimireddy, Marco Lorenzi, Giovanni Neglia, Marc Tommasi, and Mathieu Andreux. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/232eee8ef411a0a316efa298d7be3c2b-Abstract-Datasets_and_Benchmarks.html.

Yehya Farhat, Hamza ElMokhtar Shili, Fangshuo Liao, Chen Dun, Mirian Hipolito Garcia, Guoqing Zheng, Ahmed Awadallah, Robert Sim, Dimitrios Dimitriadis, and Anastasios Kyrillidis. Learning to specialize: Joint gating-expert training for adaptive moes in decentralized settings. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/03bb44eceb94537ee2e2f00a8dca60b1-Abstract-Conference.html.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. BCN20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Math. Program.*, 127(1, Ser. B):203–244, 2011. ISSN 0025-5610,1436-4646. doi: 10.1007/s10107-010-0419-x. URL https://doi.org/10.1007/s10107-010-0419-x.

Salma Kharrat, Marco Canini, and Samuel Horváth. DPFL: decentralized personalized federated learning. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025*, volume 258 of *Proceedings of Machine Learning Research*, pp. 5086–5094. PMLR, 2025. URL https://proceedings.mlr.press/v258/kharrat25a.html.

Ashiqur R. KhudaBukhsh, Peter J. Jansen, and Jaime G. Carbonell. Distributed learning in expert referral networks. In Gal A. Kaminka, Maria Fox, Paolo Bouquet, Eyke Hüllermeier, Virginia Dignum, Frank Dignum, and Frank van Harmelen (eds.), *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pp. 1620–1621. IOS Press, 2016. doi: 10.3233/978-1-61499-672-9-1620. URL https://doi.org/10.3233/978-1-61499-672-9-1620.

Jon M. Kleinberg. The small-world phenomenon: an algorithmic perspective. In F. Frances Yao and Eugene M. Luks (eds.), *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000, Portland, OR, USA*, pp. 163–170. ACM, 2000. doi: 10.1145/335305.335325. URL https://doi.org/10.1145/335305.335325.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 2020. URL http://proceedings.mlr.press/v119/koloskova20a.html.

Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173*, 2019. URL https://arxiv.org/abs/1901.11173.

Xiang Li, Chendi Wang, Buxin Su, Qi Long, and Weijie Su. Mitigating the privacy-utility trade-off in decentralized federated learning via f-differential privacy. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/f6913c4fd14f608077a4145f2744ec89-Abstract-Conference.html.

I-Cheng Lin, Osman Yagan, and Carlee Joe-Wong. Fedspd: A soft-clustering approach for personalized decentralized federated learning. In Silvia Chiappa and Sara Magliacane (eds.), *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2618–2641. PMLR, 2025. URL https://proceedings.mlr.press/v286/lin25a.html.

Liang Ma, Mudhakar Srivatsa, Derya Cansever, Xifeng Yan, Sue Kase, and Michelle Vanni. Performance bounds of decentralized search in expert networks for query answering. *ACM Trans. Knowl. Discov. Data*, 13(2):18:1–18:23, 2019. doi: 10.1145/3300230. URL https://doi.org/10.1145/3300230.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Matthias Reisser, Christos Louizos, Efstratios Gavves, and Max Welling. Federated mixture of experts. *arXiv preprint arXiv:2107.06724*, 2021.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press Cambridge, 2012.

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4289–4301, 2023. doi: 10.1109/TPAMI.2022.3196503. URL https://doi.org/10.1109/TPAMI.2022.3196503.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Zijian Wang, Pengfei Li, Guangyu Yang, and Qiong Zhang. FedSPM: Routing-enabled federated learning under dual heterogeneity via semiparametric mixture. *arXiv preprint arXiv:2607.04085*, 2026a. URL https://arxiv.org/abs/2607.04085v1.

Zijian Wang, Xiaofei Zhang, Xin Zhang, Yukun Liu, and Qiong Zhang. Beyond aggregation: Guiding clients in heterogeneous federated learning. In *International Conference on Learning Representations*, volume 2026, pp. 149981–150004, 2026b.

Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39801–39817. PMLR, 2023. URL https://proceedings.mlr.press/v202/ye23b.html.

Valentina Zantedeschi, Aurélien Bellet, and Marc Tommasi. Fully decentralized joint learning of personalized models and collaboration graphs. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 864–874. PMLR, 2020. URL http://proceedings.mlr.press/v108/zantedeschi20a.html.

A Related Work

Federated routing and mixtures of experts. FedDRM is the closest starting point for our work. It learns a globally indexed source-classification head and asks a server to compare all clients for every query (Wang et al., 2026b). FedSPM instead uses a semiparametric mixture to represent inter- and intra-client heterogeneity (Wang et al., 2026a). Federated mixture-of-experts methods also construct specialized predictors: FedMix adapts a client-specific subset of experts, whereas DDOME jointly trains gates and experts for on-the-fly selection (Reisser et al., 2021; Farhat et al., 2025). These methods establish the value of specialization, but they do not address our information constraint: at inference, no server or client can evaluate a global gate. EdgeDRM replaces global selection with a sequence of graph-supported comparisons and forms its prediction only from the experts reached by that sequence.

Personalized and decentralized federated learning. Personalized FL decides which information clients should share during training. pFedGraph learns a weighted collaboration graph at a server (Ye et al., 2023). Fully decentralized methods learn personalized models and sparse collaboration graphs, exchange sparse model masks, optimize budgeted collaborator selection, or form soft clusters through peer-to-peer updates (Zantedeschi et al., 2020; Dai et al., 2022; Kharrat et al., 2025; Lin et al., 2025). DFedAvgM combines local momentum with neighbor model averaging (Sun et al., 2023), while decentralized SGD theory characterizes gossip and local-update convergence (Koloskova et al., 2020). These approaches determine how models collaborate during training and ordinarily return a model at each client. Our problem begins after training: for each new query, the system must decide whether and where that query should move. We therefore use DFedAvgM as the primary graph-matched training baseline and isolate referral with EdgeDRM−NR, evaluated on the same EdgeDRM predictors.

Expert referral and local graph search. Expert-referral networks are closer to our deployment protocol because they model accept-or-forward decisions on sparse graphs. DIEL learns topic-dependent neighbor expertise from observed referral outcomes through a distributed active-learning

rule (KhudaBukhsh et al., 2016). Decentralized-search theory studies when local connections yield short query paths (Kleinberg, 2000; Ma et al., 2019), and learned similarity-graph routing can use graph-wide search supervision (Baranchuk et al., 2019). `EdgeDRM` requires neither logged referral outcomes nor target paths. Its local scores are learned from the static datasets held at neighboring clients, and its movement budget makes the access constraint explicit.

Density ratios and pairwise consistency. Classifier-based density-ratio estimation recovers pointwise density ratios from prior-corrected posterior odds (Sugiyama et al., 2012). Statistical Hodge theory shows how pairwise comparison flows decompose into gradient, cyclic, and harmonic components (Jiang et al., 2011). We do not propose a new density-ratio estimator in isolation. Instead, one potential per client induces compatible edge differences, while balanced endpoint classification estimates pairwise density-ratio terms and the sample-count prior is added only at inference. We then analyze how estimation error, representation disagreement, and graph structure affect the resulting referral policy.

B Derivation of the EdgeDRM Distributional Objective

This appendix establishes the precise connection between empirical likelihood (EL) and the global client-source conditional likelihood in Section 2. The equality requires profiling both the baseline masses and the density-ratio intercepts at fixed representations and slopes. We then derive the graph-local conditional objective used by `EdgeDRM`.

B.1 Empirical likelihood under the reference density-ratio model

For client i, suppose

$$dP_i^Z(z) = \varphi_i(z)dF(z), \qquad \varphi_i(z) = \exp\{b_i + \vartheta_i^\top z\}, \qquad F = P_1^Z, \quad b_1 = 0, \quad \vartheta_1 = 0.$$

The reference density satisfies $\varphi_1 \equiv 1$, and the baseline distribution F is unrestricted. Fix a common representation $z_{ur} = g_\theta(x_{ur})$ and assume $n_i > 0$ for every client. Let $\xi_1, \dots, \xi_{n_z}$ be the distinct pooled representations, let $k_t = \#\{(u,r) : z_{ur} = \xi_t\}$, and write $t(u,r)$ for the index satisfying $z_{ur} = \xi_{t(u,r)}$. Thus, $k_t \geq 1$ and $\sum_t k_t = N$. EL places a single mass $p_t = F(\{\xi_t\})$ at each distinct representation, subject to

$$p_t \geq 0, \qquad \sum_{t=1}^{n_z} p_t = 1, \qquad \sum_{t=1}^{n_z} p_t \varphi_i(\xi_t) = 1, \quad i = 2, \dots, m.$$

Aggregating repeated representations avoids counting the same atom of F more than once. Let

$$\ell_Y(\theta, \omega) = \sum_{u=1}^{m} \sum_{r=1}^{n_u} \log P_u^{Y|X}(\{y_{ur}\} \mid x_{ur}; \theta, \omega_u).$$

The label term combined with the representation-marginal empirical log-likelihood is

$$\ell_{\mathrm{EL}}(p, \theta, \omega, b, \vartheta) = \ell_Y(\theta, \omega) + \sum_{t=1}^{n_z} k_t \log p_t + \sum_{u=1}^{m} \sum_{r=1}^{n_u} \log \varphi_u(z_{ur}).$$

A finite value requires $p_t > 0$ for every observed atom. Profiling only p at fixed (b, ϑ) means maximizing over the displayed normalization constraints; it does not in general produce the client-source likelihood pointwise in b.

B.2 Exact equivalence after profiling masses and intercepts

Let $P_\rho^{C,Z}$ denote the pooled source–representation distribution with

$$P_\rho^C(\{i\}) = \rho_i, \qquad P_\rho^{Z|C}(\cdot \mid i) = P_i^Z, \qquad \rho_i = n_i/N.$$

For a normalized DRM, Bayes' rule gives

$$s_i(z) := P_\rho^{C|Z}(\{i\} \mid z) = \frac{\rho_i \varphi_i(z)}{D(z)}, \qquad D(z) := \sum_{c=1}^{m} \rho_c \varphi_c(z).$$

For arbitrary finite candidate coefficients, the same formula defines a valid conditional source model, whether or not those coefficients admit a normalized baseline F. Define its log-likelihood by

$$\ell_{\mathrm{src}}(b,\vartheta)=\sum_{u=1}^{m}\sum_{r=1}^{n_u}\log s_u(z_{ur}).$$

Fix the representations and finite slopes ϑ, with $\vartheta_1=0$. We prove

$$\sup_{\substack{p,b:\,p\text{ satisfies the EL constraints}\\ b_1=0}}\ell_{\mathrm{EL}}=\ell_Y+\max_{b:\,b_1=0}\ell_{\mathrm{src}}(b,\vartheta)+C_{\mathrm{emp}},\tag{13}$$

where $C_{\mathrm{emp}}:=\sum_{t=1}^{n_z}k_t\log(k_t/N)-\sum_{i=1}^{m}n_i\log\rho_i$. The supremum on the left is attained, and the source log-likelihood has a unique finite maximizer in the free intercepts $b_2,\dots,b_m$.

Upper bound for every feasible baseline. For any feasible (p,b), the normalization constraints give $\sum_t p_tD(\xi_t)=\sum_i\rho_i=1$. Since $\sum_t k_t/N=1$, Jensen's inequality for the logarithm yields

$$\begin{aligned}\ell_{\mathrm{EL}}-\ell_Y-\ell_{\mathrm{src}}-C_{\mathrm{emp}}&=\sum_{t=1}^{n_z}k_t\log\frac{Np_tD(\xi_t)}{k_t}\\&\le N\log\left\{\sum_{t=1}^{n_z}\frac{k_t}{N}\frac{Np_tD(\xi_t)}{k_t}\right\}=0.\end{aligned}$$

Equality holds precisely when $p_t=k_t/\{ND(\xi_t)\}$ for every t. These proposed masses need not be feasible at a fixed arbitrary intercept vector.

Existence and uniqueness of the intercept maximizer. Write $h_{it}=\log\rho_i+\vartheta_i^\top\xi_t$ and $M=\max_{i,t}|h_{it}|<\infty$. For $b_1=0$, let $B=\max_i b_i$, $A=\min_i b_i$, and $n_{\min}=\min_i n_i>0$. Bounding each numerator score above by b_u+M and each log normalizer below by $B-M$ gives

$$\ell_{\mathrm{src}}(b,\vartheta)\le 2NM+\sum_i n_ib_i-NB\le 2NM-n_{\min}(B-A)\le 2NM-n_{\min}\|(b_2,\dots,b_m)\|_\infty.$$

Thus ℓ_{src} tends to $-\infty$ as the free intercepts diverge and, being continuous, attains a finite maximum. For a nonzero vector of free intercept perturbations, extend it by $u_1=0$. Its Hessian quadratic form is

$$u^\top\nabla_b^2\ell_{\mathrm{src}}\,u=-\sum_{t=1}^{n_z}k_t\,\mathrm{Var}_{C\sim s(\xi_t)}(u_C)<0.$$

Every source probability is strictly positive, and $u_1=0$ makes any nonzero free perturbation nonconstant across clients, proving the strict inequality and uniqueness.

Attainment of the EL upper bound. Let $b^\star(\vartheta)$ denote this intercept maximizer, and evaluate φ_i and D at $b^\star$. For each $i\ge 2$, its score equation is

$$0=\partial_{b_i}\ell_{\mathrm{src}}=n_i-\rho_i\sum_{t=1}^{n_z}k_t\frac{\varphi_i(\xi_t)}{D(\xi_t)}.$$

Hence the positive masses $p_t^\star=k_t/\{ND(\xi_t)\}$ satisfy $\sum_t p_t^\star\varphi_i(\xi_t)=1$ for every $i\ge 2$. Their baseline normalization follows from $\varphi_1\equiv 1$ and $\rho_1>0$:

$$1=\sum_t p_t^\star D(\xi_t)=\rho_1\sum_t p_t^\star+\sum_{i=2}^{m}\rho_i,\qquad\text{so}\qquad\sum_t p_t^\star=1.$$

Thus $p^\star$ is feasible and attains equality in the upper bound, proving (13).

Scope of the equivalence. At fixed representations, optimizing the slopes as well preserves equality of the suprema in (13); finite global maximizers correspond through the displayed masses whenever such maximizers exist. Existence of a finite slope maximizer is a separate condition and can fail under source separation. The identity does not assert that $\sup_p \ell_{\mathrm{EL}} = \ell_Y + \ell_{\mathrm{src}} + C_{\mathrm{emp}}$ for every fixed (b, ϑ). For example, take two clients with one observation each, $z_{11} = 0$, $z_{21} = 1$, and $(\varphi_2(0), \varphi_2(1)) = (1/2, 3/2)$, while $\varphi_1 \equiv 1$. The EL constraints force both baseline masses to be $1/2$, giving marginal likelihood $3/8$, whereas the source likelihood is $(2/3)(3/5) = 2/5$ and $C_{\mathrm{emp}} = 0$. This example is itself reference normalized, so fixing the reference does not remove the need to optimize the intercepts. If all pooled representations are distinct, $C_{\mathrm{emp}} = -N \log N - \sum_i n_i \log \rho_i$. With repeated representations, C_{emp} is constant for a fixed representation but can change if learning θ changes their multiplicities. We therefore use (3) as a conditional-label plus source-classification objective; the EL identity above is a statement at fixed representations, not an unrestricted identity throughout encoder training.

B.3 From the global source loss to graph-local comparisons

Evaluating $s_i(z)$ requires the scores and normalizing sum for all clients. `EdgeDRM` instead restricts the pooled source model to the endpoints of each communication edge. For an edge $e = \{i, j\}$, training assigns probability $1/2$ to each endpoint before drawing a representation from that endpoint. Bayes' rule for this balanced endpoint mixture yields

$$q_{ij}^{\mathrm{bal},\star}(z) := \varphi_j(z)/\{\varphi_i(z) + \varphi_j(z)\},$$

and therefore

$$\operatorname{logit} q_{ij}^{\mathrm{bal},\star}(z) = \log\{\varphi_j(z)/\varphi_i(z)\} = (b_j - b_i) + (\vartheta_j - \vartheta_i)^\top z.$$

The negative empirical conditional log-likelihood of this edge-restricted source model is

$$\begin{aligned}\mathcal{L}_{e,\mathrm{src}} &= -\frac{1}{2n_i}\sum_{r=1}^{n_i} \log\{1 - q_{ij}^{\mathrm{bal}}(z_{ir})\} - \frac{1}{2n_j}\sum_{r=1}^{n_j} \log q_{ij}^{\mathrm{bal}}(z_{jr}) \\ &= \tfrac{1}{2}\mathcal{L}_e^{(i)} + \tfrac{1}{2}\mathcal{L}_e^{(j)}.\end{aligned}$$

Equal weighting of the endpoint means makes the population logit target the pairwise density ratio without a sample-proportion offset. This balanced edgewise conditional likelihood requires no graph-wide normalization. Summing these edge terms produces a conditional composite-likelihood objective, rather than the global likelihood in the EL identity above.

B.4 Balanced graph-local objective used in training

The main text keeps the label and graph-local source components separate to distinguish prediction from source classification. The implementation combines them using the fixed relative weight specified below. Client u maintains a local representation g_{θ_u} and evaluates $z_{ur} := g_{\theta_u}(x_{ur})$ on its own examples. The local parameters need not coincide after a finite number of neighbor-mixing rounds; setting all $\theta_u = \theta$ recovers the common-representation model used in the derivation above.

The softmax model in (1) gives the client-normalized task loss

$$\mathcal{L}_{\mathrm{task},i} = 1/n_i \sum_{r=1}^{n_i} \ell_{\mathrm{CE}}(y_{ir}, z_{ir}; \omega_i). \tag{14}$$

For an edge $e = \{i, j\}$ with orientation $i < j$ and endpoint $u \in \{i, j\}$, define

$$\mathcal{L}_e^{(u)} = 1/n_u \sum_{r=1}^{n_u} \ell_{\mathrm{BCE}}\big(\mathbf{1}\{u = j\}, q_{ij}^{\mathrm{bal}}(z_{ur})\big). \tag{15}$$

The equal-weighted sum $\{\mathcal{L}_e^{(i)} + \mathcal{L}_e^{(j)}\}/2$ equals the edge-source loss derived above.

Let $\lambda \in (0, 1)$ denote the relative weight assigned to task prediction and set

$$\kappa := \frac{m(1-\lambda)}{2|E|}. \tag{16}$$

With one routing potential $a_i = (\vartheta_i, b_i)$ per client, the client-local contribution is

$$\mathcal{L}_i = \lambda \mathcal{L}_{\text{task},i} + \kappa \sum_{e \in \delta(i)} \mathcal{L}_e^{(i)}, \tag{17}$$

and the corresponding graph-level objective is

$$\mathcal{L}_G = 1/m \sum_i \mathcal{L}_i = \lambda/m \sum_i \mathcal{L}_{\text{task},i} + (1-\lambda)/(2|E|) \sum_{e=\{i,j\} \in E} \{\mathcal{L}_e^{(i)} + \mathcal{L}_e^{(j)}\}. \tag{18}$$

Each endpoint loss is a mean over its local data; the two endpoints and all edges receive equal weight, while degrees are not separately normalized. The reported experiments use $\lambda = 0.9$, while $\lambda = 1$ gives the task-only ablation. For fixed features, the prediction-head and routing-potential subproblems separate.

During Stage II, endpoint i differentiates each incident loss with respect to both endpoint potentials and returns the neighbor-owned component. The unregularized routing-potential gradient received by owner i is

$$G_i = \nabla_{a_i} \mathcal{L}_i + \kappa \sum_{j \in \mathcal{N}(i)} \nabla_{a_i} \mathcal{L}_{\{i,j\}}^{(j)}.$$

This expression equals $\nabla_{a_i} \sum_u \mathcal{L}_u$ in unrestricted potential coordinates when all endpoint contributions are evaluated at the same local-step snapshot. For the reference-constrained theoretical update, a_1 remains fixed at zero and only the components G_i for $i \geq 2$ are used. The private head receives only $\lambda \nabla_{\omega_i} \mathcal{L}_{\text{task},i}$, while the frozen encoder receives no gradient.

Training estimates the density-ratio term from the balanced endpoint likelihood. At inference, the sample-count prior is added to the learned comparison:

$$\log\{s_j(z)/s_i(z)\} = \log(\rho_j/\rho_i) + \log\{\varphi_j(z)/\varphi_i(z)\} = \log(n_j/n_i) + (b_j - b_i) + (\vartheta_j - \vartheta_i)^\top z.$$

Thus, the fitted graph-local logit excludes the sample-count prior during training, and inference adds $\log(n_j/n_i)$ exactly once.

When the frozen local representations differ, the fitted reverse-edge scores need not be antisymmetric; Section 4 isolates this representation-disagreement error from coefficient-estimation error.

C Proofs and Additional Theory

Appendix C.1 proves the coefficient-estimation guarantee, Appendix C.2 derives the comparison identities, and Appendix C.3 proves the referral guarantee. For the routing arguments, fix the positive training counts and query (x, s). We use the route $\mathcal{R}_L(x, s) = (\widehat{s}_0 = s, \ldots, \widehat{s}_T)$ from (7), with terminal client $\widehat{s}_T$, and the scores u_t and weights w_t from Section 3.2. Empty sums are zero and $\max \varnothing = -\infty$.

C.1 Proof of Theorem 1

Statistical conditions. Fix a connected graph with $m \geq 2$, the number of classes, the representation and coefficient dimensions, and the task weight $\lambda \in (0, 1)$. The common representation is fixed independently of the fitting samples. Within each client the observations are i.i.d. from a fixed distribution $P_i^{X,Y}$, and different clients are independent. The deterministic allocations satisfy $\rho_{i,N} := n_i/N \to \rho_i^* > 0$; for a fixed N, these are exactly the proportions ρ_i used by the inference-time prior in Section 3. The conditional-label model and the common-reference DRM are correctly specified by the same finite coefficient vector $\zeta^\star$ for all N. For identifiability, we use $F = P_1^Z$ and fix $b_1 = 0$ and $\vartheta_1 = 0$, so that $\varphi_1 \equiv 1$. In identifiable coordinates the limiting population objective defined below has a unique interior minimizer $\zeta^\star$. On a compact convex parameter region containing this point in its interior, the observation losses and their first two derivatives have integrable envelopes; the score at $\zeta^\star$ has finite second moment, and the population Hessian is positive definite. Losses are measurable and twice continuously differentiable on a neighborhood of the region, and empirical minimizers are measurable. These compact-region regularity conditions apply directly to the balanced training losses below. The optimization conditions below specify existence of a finite empirical minimizer and containment of the iterates.

Observation contributions. We decompose the graph objective $\mathcal{L}_G$ in (18) into observation-level contributions. With $z = g_\theta(x)$ and $\sigma(t) = (1 + e^{-t})^{-1}$, write

$$q_{ij}^{\mathrm{bal}}(z;\zeta) = \sigma\big(b_j - b_i + (\vartheta_j - \vartheta_i)^\top z\big)\,,$$
$$\ell_e^{(u)}(W;\zeta) = \ell_{\mathrm{BCE}}\big(\mathbf{1}\{u=j\}, q_{ij}^{\mathrm{bal}}(g_\theta(X);\zeta)\big)\,, \qquad u \in \{i,j\}.$$

The density-tilt coefficients retain their definition in (2); no sample-count offset appears in this training logit. For $W_{ir} = (X_{ir}, Y_{ir})$, define

$$\phi_i(W;\zeta) = \frac{\lambda}{m}\ell_{\mathrm{task},i}(W;\zeta) + \frac{1-\lambda}{2|E|}\sum_{e\ni i}\ell_e^{(i)}(W;\zeta),$$

$$\mathcal{L}_G(\zeta) = \sum_i \frac{1}{n_i}\sum_{r=1}^{n_i}\phi_i(W_{ir};\zeta), \qquad R_G(\zeta) = \sum_i \mathbb{E}_i\phi_i(W_{i1};\zeta),$$

where $\mathbb{E}_i$ denotes expectation under $P_i^{X,Y}$. All losses using the same observation remain inside the same ϕ_i. Thus $\mathbb{E}\mathcal{L}_G = R_G$ for every allocation: the sample proportions affect the estimator's variance and the inference-time prior, but not the balanced population training objective.

Identifiable coordinates and the gradient update. The routing coordinates are the free blocks $a_i = (\vartheta_i, b_i)$ for $i = 2, \dots, m$, with $a_1 = 0$ held fixed. For each private head, retain $\sum_k \alpha_{ik} = 0$ and $\sum_k \beta_{ik} = 0$ and use orthonormal coordinates on this class-contrast subspace. All parameter derivatives and norms in Theorem 1 are taken in these coordinates. All unqualified vector norms below denote the Euclidean norm. Choosing client 1 as the density reference preserves every pairwise ratio: from any common baseline F_0 with coefficients (b_i^0, ϑ_i^0), take $F = P_1^Z$ and $(b_i, \vartheta_i) = (b_i^0 - b_1^0, \vartheta_i^0 - \vartheta_1^0)$. Then $dP_i^Z/dF = \exp\{b_i + \vartheta_i^\top z\}$ remains a normalized density. The private-head gradient sums to zero across classes, so common-step unpenalized updates preserve its class-contrast subspace. The analyzed routing update acts only on the free blocks $a_2, \dots, a_m$ and keeps $a_1 = 0$.

We analyze synchronous, unpenalized plain SGD in these free coordinates for Stage II. Write $\mathcal{L}_i$ and $\mathcal{L}_e^{(u)}$ for the method's local and endpoint-average losses. For each free routing-potential owner $i \neq 1$, the gradient is exactly

$$G_i^a = \nabla_{a_i}\mathcal{L}_i + \kappa\sum_{j\in\mathcal{N}(i)}\nabla_{a_i}\mathcal{L}_{\{i,j\}}^{(j)}, \qquad \kappa = \frac{m(1-\lambda)}{2|E|}.$$

The head owner receives $G_i^\omega = \lambda\nabla_{\omega_i}\mathcal{L}_{\mathrm{task},i}$, expressed in the head coordinates above. Client 1 still contributes its task gradient and the source gradients of its observations to neighboring free routing blocks. With the same snapshot for every endpoint contribution, concatenating the free routing blocks and all head-coordinate blocks gives $\nabla\sum_i \mathcal{L}_i = m\nabla\mathcal{L}_G$, including the gradients returned by neighbors. Divide the assembled gradient by m; equivalently, a step of size η_r on the normalized gradient uses step size η_r/m on the raw owner gradients. The analyzed update is

$$\zeta^{(r+1)} = \zeta^{(r)} - \eta_r\widehat{G}_r, \qquad \mathbb{E}[\widehat{G}_r \mid \mathcal{F}_r] = \nabla\mathcal{L}_G(\zeta^{(r)}), \tag{19}$$

where the data and initialization are included in the optimization history $\mathcal{F}_r$. At each update, independently across clients, sample B_i^{mb} local indices uniformly with replacement and average $\nabla\phi_i$ over them, then sum these averages over i. This gives exactly the conditional mean in (19). If $\widehat{P}_i$ is the empirical distribution at client i, the conditional covariance of this normalized mini-batch gradient is

$$\sum_i \frac{1}{B_i^{\mathrm{mb}}}\operatorname{Var}_{\widehat{P}_i}\{\nabla\phi_i(W;\zeta^{(r)})\}.$$

Its trace is the conditional noise second moment. This covariance includes the correlations between incident-edge losses evaluated on the same observation. A variance bound for raw owner gradients must be divided by m^2 to give the corresponding normalized bound. Other samplers are covered only when the stated conditional unbiasedness and joint noise bound hold.

Optimization conditions. The optimization analysis uses the following finite-minimizer, curvature, and bounded-iterate conditions. On a data-and-initialization event whose probability tends to one, let

$\widehat{\zeta}_N$ be a finite empirical minimizer with zero coordinate gradient, and let a convex region $\mathcal{K}_N$ contain this minimizer and the initial iterate, with

$$\mu I \preceq \nabla^2 \mathcal{L}_G(\zeta) \preceq \mu^{-1} I, \qquad 0 < \mu \le 1, \qquad \zeta \in \mathcal{K}_N.$$

Any uniform positive lower and finite upper curvature bounds admit this form after choosing μ sufficiently small. Assume the conditional noise second moment is at most σ^2 while the iterate is in this region. The constants are uniform in N, and $\|\zeta^{(0)} - \widehat{\zeta}_N\| = O_p(1)$. As in the bounded-iterate premise, all iterates up to the analyzed budget must remain in the region, either almost surely conditional on these data or with probability tending to one. For high-probability containment, the proof uses the first exit time from $\mathcal{K}_N$. On a bounded-logit region, positive-definite augmented-feature Gram matrices for every task head and balanced edge mixture, together with graph connectivity, are sufficient for positive curvature in the free reference-client and private-head coordinates. Existence of a finite minimizer and containment of the iterates are separate assumptions; complete separation and unidentified design directions are excluded. The curvature condition is local to $\mathcal{K}_N$.

Proof of Theorem 1. **Step 1: A common population minimizer.** For each edge let

$$\overline{\varphi}_e(z) = \{\varphi_i(z) + \varphi_j(z)\}/2.$$

Under the correctly specified DRM the true endpoint-source posterior is

$$q_{ij}^{\mathrm{bal},\star}(z) = \frac{\varphi_j(z)}{\varphi_i(z) + \varphi_j(z)} = q_{ij}^{\mathrm{bal}}(z; \zeta^\star).$$

Let $\xi_e(z)$ be the gradient in identifiable coordinates of $(b_j - b_i) + (\vartheta_j - \vartheta_i)^\top z$. It does not depend on the parameter value, since these logits are affine. The source-label gradient for an observation from endpoint u is $(q_{ij}^{\mathrm{bal}} - \mathbf{1}\{u = j\})\xi_e$. Consequently, at $\zeta^\star$,

$$\begin{aligned} &\tfrac{1}{2}\mathbb{E}_i[q_{ij}^{\mathrm{bal},\star}(Z)\xi_e(Z)] + \tfrac{1}{2}\mathbb{E}_j[(q_{ij}^{\mathrm{bal},\star}(Z) - 1)\xi_e(Z)] \\ &\quad = \int \{\overline{\varphi}_e(z) q_{ij}^{\mathrm{bal},\star}(z) - \tfrac{1}{2}\varphi_j(z)\}\xi_e(z)\, dF(z) = 0. \end{aligned}$$

The two equally weighted endpoint means therefore cancel exactly. For the task term, write $p_i(z;\zeta)$ for the model class-probability vector and $p_i^\star(z) = p_i(z;\zeta^\star)$ for its true value. Let $A_i(z)$ be the Jacobian of the K task logits with respect to ζ, and let $\mathbf{e}_Y$ denote the class indicator vector. Correct specification gives

$$\mathbb{E}_i[\nabla \ell_{\mathrm{task},i}(W;\zeta^\star) \mid Z] = A_i(Z)^\top \{p_i^\star(Z) - \mathbb{E}_i[\mathbf{e}_Y \mid Z]\} = 0.$$

Thus, with $s_i(W) = \nabla \phi_i(W;\zeta^\star)$ and $m_i = \mathbb{E}_i s_i(W)$, the population-score identity is

$$\nabla R_G(\zeta^\star) = \sum_i m_i = 0, \qquad \text{while generally } m_i \neq 0.$$

Convexity already makes $\zeta^\star$ a minimizer. More explicitly, $R_G(\zeta) - R_G(\zeta^\star)$ is the sum of

$$\frac{\lambda}{m} \sum_i \mathbb{E}_i \,\mathrm{KL}\{p_i^\star(Z) \,\|\, p_i(Z;\zeta)\}$$

and

$$\frac{1-\lambda}{|E|} \sum_{e=\{i,j\}} \int \overline{\varphi}_e(z)\, \mathrm{KL}\Big\{\mathrm{Ber}(q_{ij}^{\mathrm{bal},\star}(z)) \,\|\, \mathrm{Ber}(q_{ij}^{\mathrm{bal}}(z;\zeta))\Big\}\, dF(z).$$

All weights are positive. Equality requires the correct task probabilities and density-score differences on every edge. The stated identification condition therefore makes $\zeta^\star$ the unique minimizer of the balanced population objective R_G.

Step 2: Score covariance. The endpoint score second moments follow from the information identity. At the true balanced posterior,

$$\begin{aligned} &\tfrac{1}{2}\sum_{u \in e} \mathbb{E}_u\Big[(q_{ij}^{\mathrm{bal},\star}(Z) - \mathbf{1}\{u = j\})^2 \|\xi_e(Z)\|^2\Big] \\ &\quad = \int \overline{\varphi}_e(z) q_{ij}^{\mathrm{bal},\star}(z)\{1 - q_{ij}^{\mathrm{bal},\star}(z)\}\|\xi_e(z)\|^2\, dF(z) < \infty. \end{aligned}$$

The right-hand side is the trace of the edge's population Hessian before its positive graph-level coefficient. All Hessian contributions are positive semidefinite, so finiteness follows from the stated Hessian envelope. Each endpoint term is finite. The task-score second moment follows from the stated moment condition (and from its correctly specified softmax information identity). Uniform integrable gradient envelopes also follow by expanding each gradient from $\zeta^\star$ over the compact region using its Hessian envelope.

Here are the exact score and Hessian for the balanced objective. All probabilities in the following displays are evaluated at $\zeta^\star$ and the displayed representation:

$$\begin{aligned} s_i(W) &= \frac{\lambda}{m} A_i(Z)^\top \{p_i^\star(Z) - \mathbf{e}_Y\} \\ &\quad + \frac{1-\lambda}{2|E|} \sum_{e=\{v,w\} \ni i,\ v<w} (q_{vw}^{\mathrm{bal},\star}(Z) - \mathbf{1}\{i=w\}) \xi_e(Z), \\ H_G := \nabla^2 R_G(\zeta^\star) &= \frac{\lambda}{m} \sum_i \mathbb{E}_i\big[A_i(Z)^\top \{\mathrm{diag}(p_i^\star(Z)) - p_i^\star(Z) p_i^\star(Z)^\top\} A_i(Z)\big] \\ &\quad + \frac{1-\lambda}{|E|} \sum_{e=\{i,j\}} \int \overline{\varphi}_e(z) q_{ij}^{\mathrm{bal},\star}(z)\{1 - q_{ij}^{\mathrm{bal},\star}(z)\} \xi_e(z) \xi_e(z)^\top \, dF(z). \end{aligned}$$

Set

$$\begin{aligned} V_i &= \mathbb{E}_i[(s_i - m_i)(s_i - m_i)^\top], \\ J_{G,N} &= \mathrm{Var}\{\sqrt{N} \nabla \mathcal{L}_G(\zeta^\star)\} = \sum_i \frac{N}{n_i} V_i, \\ J_G &= \lim_N J_{G,N} = \sum_i (\rho_i^*)^{-1} V_i, \\ \Sigma_G &= H_G^{-1} J_G H_G^{-\top}. \end{aligned}$$

Here V_i is centered within client stratum i, whose score mean m_i may be nonzero. If $\psi_e^{(i)} = (q_{vw}^{\mathrm{bal},\star} - \mathbf{1}\{i=w\})\xi_e$ for $e=\{v,w\}$, $v<w$, then the routing part of V_i contains the full double sum

$$\left(\frac{1-\lambda}{2|E|}\right)^2 \sum_{e,f \ni i} \mathrm{Cov}_i(\psi_e^{(i)}, \psi_f^{(i)}).$$

The terms with $e \neq f$ account for the reuse of each observation on incident edges. The full vector definition of V_i also retains task contributions and their covariance with routing contributions. Under correct task specification, the task–routing cross-covariances vanish because the task score has conditional mean zero given Z. In general J_G is not the same as H_G.

Step 3: Statistical consistency and asymptotic normality. The uniform-law conditions give, on the compact parameter region,

$$\sup_\zeta |\mathcal{L}_G(\zeta) - R_G(\zeta)| \to_p 0,$$

and the analogous uniform convergence for the Hessians. Uniqueness and interior positive curvature yield $\widehat{\zeta}_N \to_p \zeta^\star$. On any sufficiently small sphere about $\zeta^\star$, the population objective exceeds its value at $\zeta^\star$ uniformly, and so does the empirical objective with probability tending to one. Convexity then excludes a better empirical minimizer outside that sphere, by considering the segment to $\zeta^\star$.

The exact cancellation of total score means gives

$$Z_N := \sqrt{N} \nabla \mathcal{L}_G(\zeta^\star) = \sum_i \frac{\sqrt{N}}{n_i} \sum_{r=1}^{n_i} \{s_i(W_{ir}) - m_i\}.$$

The centered summands have independent client strata. For each client the ordinary finite-variance multivariate i.i.d. central limit theorem applies to $s_i - m_i$. Since $\sqrt{N/n_i} \to (\rho_i^*)^{-1/2}$ and the number of clients is fixed, $Z_N \xrightarrow{d} \mathcal{N}(0, J_G)$. Thus the allocation proportions enter the asymptotic covariance, although they do not enter the balanced population objective.

On the event that the empirical minimizer is interior, its exact first-order condition gives

$$0 = \nabla \mathcal{L}_G(\zeta^\star) + \widehat{H}_N(\widehat{\zeta}_N - \zeta^\star), \qquad \widehat{H}_N = \int_0^1 \nabla^2 \mathcal{L}_G\{\zeta^\star + t(\widehat{\zeta}_N - \zeta^\star)\}\, dt \to_p H_G.$$

Consequently,

$$\sqrt{N}(\widehat{\zeta}_N - \zeta^\star) = -H_G^{-1}\sqrt{N}\nabla \mathcal{L}_G(\zeta^\star) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Sigma_G). \tag{20}$$

Step 4: Optimization error. Take $\eta_r = 2/\{\mu(r + r_0)\}$ with $r_0 \geq 2/\mu^2$, so that $r_0 \geq 2$ and $\eta_r \leq \mu$. An integral Hessian expansion along the segment between ζ and $\widehat{\zeta}_N$ gives, for $0 < \eta^{\mathrm{opt}} \leq \mu$,

$$\|\zeta - \eta^{\mathrm{opt}} \nabla \mathcal{L}_G(\zeta) - \widehat{\zeta}_N\| \leq (1 - \mu\eta^{\mathrm{opt}})\|\zeta - \widehat{\zeta}_N\|.$$

Conditional unbiasedness makes the cross term with the gradient noise zero. Let τ_N be the first iterate index outside $\mathcal{K}_N$ and put

$$v_r = \mathbb{E}_{\mathrm{opt}}\left[\|\zeta^{(r)} - \widehat{\zeta}_N\|^2 \mathbf{1}\{\tau_N > r\}\right].$$

This is the usual mean-square error when containment holds almost surely. The event $\{\tau_N > r\}$ is measurable with respect to $\mathcal{F}_r$. Since $\{\tau_N > r + 1\} \subseteq \{\tau_N > r\}$,

$$v_{r+1} \leq \mathbb{E}_{\mathrm{opt}}\left[\mathbf{1}\{\tau_N > r\}\, \mathbb{E}\left[\|\zeta^{(r+1)} - \widehat{\zeta}_N\|^2 \mid \mathcal{F}_r\right]\right].$$

Applying conditional unbiasedness on the $\mathcal{F}_r$-measurable event inside this expectation gives

$$v_{r+1} \leq \left(1 - \frac{2}{r + r_0}\right)^2 v_r + \frac{4\sigma^2}{\mu^2(r + r_0)^2} \leq \left(1 - \frac{2}{r + r_0}\right) v_r + \frac{4\sigma^2}{\mu^2(r + r_0)^2}.$$

The last inequality uses $r + r_0 \geq 2$. Induction gives

$$v_R \leq \frac{C_{\mathrm{opt}}}{R + r_0}, \qquad C_{\mathrm{opt}} = \max\left\{r_0 v_0, \frac{4\sigma^2}{\mu^2}\right\}. \tag{21}$$

Indeed, writing $d = r + r_0$ and $A = 4\sigma^2/\mu^2$, if $v_r \leq C/d$ with $C \geq A$, then $v_{r+1} \leq C/d - (2C - A)/d^2 \leq C/d - C/d^2 \leq C/(d+1)$. The initial condition starts the induction. Since $C_{\mathrm{opt}} = O_p(1)$, conditional Markov's inequality and the stated containment condition $\Pr(\tau_N \leq R) \to 0$ imply

$$\zeta^{(R)} - \widehat{\zeta}_N = O_p(R^{-1/2}).$$

With high-probability containment, (21) bounds the stopped error v_R; the exit probability is controlled separately by the containment assumption.

Step 5: The SGD output. Combining this expansion with (21) yields

$$\|\zeta^{(R)} - \zeta^\star\| \leq \|\widehat{\zeta}_N - \zeta^\star\| + \|\zeta^{(R)} - \widehat{\zeta}_N\| = O_p(N^{-1/2} + R^{-1/2}),$$

which proves (8). If $R/N \to \infty$, the stopped mean-square bound and the containment condition imply $\sqrt{N}(\zeta^{(R)} - \widehat{\zeta}_N) = o_p(1)$, so (20) proves (9). This proves Theorem 1.

□

Full-gradient updates. For full gradients and $0 < \eta^{\mathrm{opt}} \leq \mu$, the same curvature and iterate-containment conditions give

$$\|\zeta^{(R)} - \widehat{\zeta}_N\| \leq (1 - \mu\eta^{\mathrm{opt}})^R \|\zeta^{(0)} - \widehat{\zeta}_N\|. \tag{22}$$

Adding the statistical error gives $O_p\{N^{-1/2} + (1 - \mu\eta^{\mathrm{opt}})^R\}$. Thus $\sqrt{N}(1 - \mu\eta^{\mathrm{opt}})^R \to 0$ suffices for the same normal limit. For fixed $c > 0$, it is sufficient to take $R \geq (1/2 + c)(\mu\eta^{\mathrm{opt}})^{-1} \log N$; when $\mu\eta^{\mathrm{opt}} = 1$ the contraction is zero after one update.

C.2 Balanced source fitting and prior-weighted comparison identities

For $e = \{i, j\}$, $i < j$, the source-label distribution in the edge risk assigns probability $1/2$ to each endpoint. Bayes' rule under this balanced mixture gives

$$q_{ij}^{\mathrm{bal},\star}(z) = \frac{\varphi_j(z)}{\varphi_i(z) + \varphi_j(z)}, \qquad \operatorname{logit} q_{ij}^{\mathrm{bal},\star}(z) = \log \frac{\varphi_j(z)}{\varphi_i(z)}.$$

The equally weighted binary cross-entropy is minimized by this posterior, which contains no sample-count term and agrees with (5). Correct specification, connectivity, and the stated identifiable-design condition identify the routing coefficients relative to the fixed client 1. Correctly specified task cross-entropies identify the private-head contrasts. At inference, adding $\log(\rho_j/\rho_i)$ to the fitted balanced logit converts the learned density ratio into the desired prior-weighted source odds. Thus the sample-count prior is absent from training and appears exactly once in deployment.

With a common fitted representation $z = g_{\widehat{\theta}}(x)$,

$$u_t(x, s) = \log \frac{\rho_{\widehat{s}_t}}{\rho_s} + (\widehat{b}_{\widehat{s}_t} - \widehat{b}_s) + (\widehat{\vartheta}_{\widehat{s}_t} - \widehat{\vartheta}_s)^\top z.$$

The identity holds for any fitted coefficients with a common representation. If the DRM holds for this representation and its density-score differences are recovered exactly, then

$$e^{u_t(x,s)} = \frac{\rho_{\widehat{s}_t}\varphi_{\widehat{s}_t}(z)}{\rho_s\varphi_s(z)}, \qquad w_t(x, s) = \frac{\rho_{\widehat{s}_t}\varphi_{\widehat{s}_t}(z)}{\sum_{\ell=0}^{T} \rho_{\widehat{s}_\ell}\varphi_{\widehat{s}_\ell}(z)}.$$

The history weights normalize the prior-weighted densities over the visited clients. With different local encoders, the count terms still telescope but the remaining score sum can be path dependent.

For a reference representation $g_\theta(x)$ and its coefficients, adding and subtracting the fitted comparison at this representation gives

$$\begin{aligned} \widehat{\Delta}_{ij}(x) - \Delta_{ij}(x) &= e_{ij}^{\mathrm{coef}}(x) + (\widehat{\vartheta}_j - \widehat{\vartheta}_i)^\top \{g_{\widehat{\theta}_i}(x) - g_\theta(x)\}, \\ e_{ij}^{\mathrm{coef}}(x) &= (\widehat{b}_j - b_j) - (\widehat{b}_i - b_i) \\ &\quad + \{(\widehat{\vartheta}_j - \vartheta_j) - (\widehat{\vartheta}_i - \vartheta_i)\}^\top g_\theta(x). \end{aligned} \tag{23}$$

Both scores use the same finite-N count prior, so their known offsets cancel exactly, irrespective of how quickly the proportions converge. If the coefficient term is bounded by ϵ_a, the fitted slope contrast by M, and the representation difference by δ, the score error is at most $\epsilon_a + M\delta$. For common representations and the same count prior, each comparison is affine in the identifiable coefficients. Using client 1 as the reference for both fitted and true coefficients, its error is bounded by $2\|(g_\theta(x)^\top, 1)^\top\| \max_i \|\widehat{a}_i - a_i\|$. Theorem 1 therefore proves (10) at a fixed query; bounded representations give a uniform version. For different local encoders, (23) remains an algebraic decomposition. However, these encoders also change the representation distributions used for coefficient fitting. The coefficient term may therefore contain effects of this change in addition to statistical estimation error, and Theorem 1 does not automatically bound it. Extending the rate to this setting requires controlling both the coefficient-fitting error and the representation-disagreement term.

The reverse-edge discrepancy is

$$\widehat{\Delta}_{ij}(x) + \widehat{\Delta}_{ji}(x) = (\widehat{\vartheta}_j - \widehat{\vartheta}_i)^\top \{g_{\widehat{\theta}_i}(x) - g_{\widehat{\theta}_j}(x)\}. \tag{24}$$

Here the two count offsets and the intercept differences cancel even with different encoders. With positive reference population proportions ρ_i^{pop} instead, the additional error is $\log(\rho_j/\rho_j^{\mathrm{pop}}) - \log(\rho_i/\rho_i^{\mathrm{pop}})$; hence

$$|\widehat{\Delta}_{ij} - \Delta_{ij}^{\mathrm{pop}}| \le \epsilon_a + M\delta + 2 \max_i |\log(\rho_i/\rho_i^{\mathrm{pop}})|. \tag{25}$$

The last term measures the difference between the training-source prior and the reference population prior.

C.3 Proof of Theorem 2 and structural examples

From estimation error to comparison tolerance. Under Theorem 1, (10) bounds the maximum comparison error over all graph edges at the fixed query x. By the definition of O_p, for every fixed $\delta \in (0,1)$ there is a finite constant $C_\delta > 0$ such that, for all sufficiently large N, R, the event

$$\mathcal{E}_\delta := \left\{ \max_{\{i,j\}\in E} |\widehat{\Delta}_{ij}(x) - \Delta_{ij}(x)| \le C_\delta \left(N^{-1/2} + R^{-1/2} \right) = \epsilon \right\} \tag{26}$$

has probability at least $1-\delta$ over the training data and SGD randomness. The constant may depend on the fixed query, graph, and model, but not on N, R, or the realized route. With the common representation, reverse comparisons are negatives of each other, so this event controls both edge directions. It therefore controls every candidate edge on the adaptively selected route without an additional probability bound for individual steps. All remaining arguments are deterministic on $\mathcal{E}_\delta$; the reference potentials and H_ϵ are fixed given the training counts and query. For an independently random frozen representation, the argument is conditional on that representation, as in Theorem 1.

Definition of the depth. Suppress x in potentials and let H_ϵ be the largest integer $q \ge 0$ for which there exists a path $(i_0 = s, i_1, \dots, i_q)$ satisfying, at every step $r < q$,

$$i_{r+1} \in \mathcal{N}(i_r), \qquad h_{i_{r+1}} > h_{i_r}, \qquad h_{i_{r+1}} \ge \max_{j\in\mathcal{N}(i_r)} h_j - 2\epsilon.$$

The one-node path is permitted. Strict increase excludes revisits, so $0 \le H_\epsilon \le m-1$. The maximum is attained because the graph is finite. Throughout the proof, H_ϵ denotes this exact maximum for the given graph, potentials, entry, and tolerance.

Proof of Theorem 2. Work on $\mathcal{E}_\delta$ and put $d_\epsilon = \gamma - 2\epsilon > 0$.

Step 1: Accepted moves and stopping. Every accepted edge satisfies $\Delta_{ij} \ge \widehat{\Delta}_{ij} - \epsilon > \eta - \epsilon \ge 0$. Hence every previously visited client has smaller potential than the current client, and every positive-gain neighbor is unvisited. At a nonoptimal client i, let $j^\star$ maximize the true neighboring gain. By (11), it is eligible and $\widehat{\Delta}_{ij^\star} \ge \gamma - \epsilon > \eta$. Thus the route cannot stop at i while budget remains. For the selected neighbor j,

$$\Delta_{ij} \ge \widehat{\Delta}_{ij} - \epsilon \ge \widehat{\Delta}_{ij^\star} - \epsilon \ge \Delta_{ij^\star} - 2\epsilon \ge d_\epsilon.$$

This also gives $h_j \ge \max_{k\in\mathcal{N}(i)} h_k - 2\epsilon$, so the realized route is admissible in the definition of H_ϵ. Consequently $T \le \min\{L, H_\epsilon\}$ and

$$\Gamma_L(x,s) \ge T(\gamma - 2\epsilon). \tag{27}$$

At a global maximizer, every candidate fitted score is at most $\epsilon \le \eta$, so no further move is accepted. An early stop therefore occurs at a maximizer.

Step 2: The budget bound. If $T = L$, (27) gives $\Gamma_L \ge L d_\epsilon$. If $T < L$, Step 1 gives $\Gamma_L = h^\star - h_s$. Thus

$$\Gamma_L \ge \min\{h^\star - h_s, L(\gamma - 2\epsilon)\}. \tag{28}$$

Step 3: The total potential gap and ascent depth. Let $(i_0 = s, \dots, i_{H_\epsilon})$ attain the maximum in the definition of H_ϵ. Its terminal vertex is a maximizer: otherwise, a highest-potential neighbor has positive gain by (11), is unvisited by strict ascent, and extends the admissible path. For each $r < H_\epsilon$, the near-best-neighbor condition gives

$$h_{i_{r+1}} - h_{i_r} \ge \max_{j\in\mathcal{N}(i_r)} (h_j - h_{i_r}) - 2\epsilon \ge d_\epsilon.$$

Telescoping along this path yields the separate total-gap bound

$$h^\star - h_s = \sum_{r=0}^{H_\epsilon - 1} (h_{i_{r+1}} - h_{i_r}) \ge H_\epsilon d_\epsilon.$$

Together with Step 2, this proves (12):

$$\Gamma_L \ge \min\{h^\star - h_s, L d_\epsilon\} \ge d_\epsilon \min\{L, H_\epsilon\},$$

Step 4: A sufficient budget for reaching a maximizer. Suppose $L \geq H_\epsilon$ and the terminal client is suboptimal. Step 1 then implies budget exhaustion, so $T = L$. Together with $T \leq H_\epsilon$, this gives $T = L = H_\epsilon$. A highest-potential neighbor would extend the realized admissible path, contradicting the definition of H_ϵ. Thus the terminal client is a maximizer. The same arguments include $L = 0$ and an optimal entry, and hold under any tie breaking among score maximizers. □

Examples and the proportional-ascent alternative. The following sufficient conditions concern the prior-adjusted potentials at a fixed query. Their conclusions hold on the comparison-error event used above.

Assumption 1 (Uniform ascent)**.** *Every nonoptimal client has a neighbor with gain at least $\gamma > 0$, as in* (11)*.*

Assumption 2 (Referral hierarchy)**.** *A spanning tree is oriented toward a maximizing root. Each nonroot client i has a parent $p_x(i)$ such that*

$$\Delta_{i,p_x(i)} \geq \gamma, \qquad \Delta_{i,p_x(i)} - \max_{j \in \mathcal{N}(i) \setminus \{p_x(i)\}} \Delta_{ij} \geq \kappa_{\mathrm{hier}} > 0.$$

When $\kappa_{\mathrm{hier}} > 2\epsilon$, only the parent can belong to an admissible ascent path. Thus H_ϵ is exactly the number of parent edges from s to the root. Under the theorem's threshold condition,

$$T = \min\{L, H_\epsilon\}, \qquad \Gamma_L \geq \gamma \min\{L, H_\epsilon\}. \tag{29}$$

The preferred-parent condition therefore identifies ascent depth with the distance to the root in this hierarchy.

Assumption 3 (Mixture of expert types)**.** *Clients of the same type $\tau(i)$ have equal potential. There is a best type $c^\star(x)$ with $h_j - h_i \geq \gamma > 0$ whenever $\tau(j) = c^\star(x)$ and $\tau(i) \neq c^\star(x)$. Every nonbest-type client has a best-type neighbor.*

If $\gamma > \eta + \epsilon$ and $\eta \geq \epsilon$, then $\gamma > 2\epsilon$. Only best-type neighbors are admissible, so $H_\epsilon = 1$ for a nonoptimal entry and $H_\epsilon = 0$ for an optimal entry. Therefore

$$\Gamma_L = (h^\star - h_s)\mathbf{1}\{L \geq 1\}. \tag{30}$$

Equal representation distributions and equal source priors within each type suffice for the equal-potential condition.

Assumption 4 (Graph–PL ascent)**.** *There is $\alpha \in (0, 1]$ such that each nonoptimal client satisfies $\max_{j \in \mathcal{N}(i)} \Delta_{ij} \geq \alpha(h^\star - h_i)$.*

This is an alternative to uniform ascent: the available improvement scales with the remaining potential gap. Under this assumption, the candidate error bound and $\eta \geq \epsilon$ alone imply

$$\Gamma_L \geq \{1 - (1 - \alpha)^L\} \left[h^\star - h_s - \frac{\eta + \epsilon}{\alpha} \right]_+ . \tag{31}$$

To prove it, hold a stopped route at its terminal node until index L only for the analysis, and put $e_t = h^\star - h_{\widehat{s}_t}$. On a moving step, the preceding noisy-greedy inequality gives $e_{t+1} \leq (1 - \alpha)e_t + 2\epsilon$. On a stopped step, the best neighboring gain is at most $\eta + \epsilon$, so $\alpha e_t \leq \eta + \epsilon$ and $e_{t+1} = e_t \leq (1 - \alpha)e_t + \eta + \epsilon$. Using $2\epsilon \leq \eta + \epsilon$ and solving the common recursion proves (31). In particular an early stop satisfies

$$T < L \quad \Longrightarrow \quad h^\star - h_{\widehat{s}_T} \leq \frac{\eta + \epsilon}{\alpha}. \tag{32}$$

When $T = L$, the sharper 2ϵ numerator follows from the moving-step recursion. Use $(1 - \alpha)^0 = 1$, including $\alpha = 1$. The holding steps are an analytical convention; T, the prediction history, and communication counts are those of the original route.

C.4 Scope of the guarantees

Estimation setting. Theorem 1 concerns correctly specified coefficient models with an independently fixed common representation and the unpenalized SGD updates in (19). The empirical-minimizer, local-curvature, noise, and iterate-containment assumptions are required. If the frozen representation

is independently random, the rate and normal limit are conditional on it, including the covariance Σ_G. The normal limit for the SGD output requires $R/N \to \infty$; $R \asymp N$ gives the stated rate alone. The analysis covers neither Stage-I representation learning nor the experimental momentum, weight-decay, and finite cosine schedule. A fixed nonzero penalty changes the estimation target.

Ascent structure. On the path 1–2–3 with potentials $(1, 0, 2)$, entry 1, exact scores, and threshold zero, the route makes no move despite radius one and diameter two. A path to a better client exists but requires a downhill step. Even on the same graph with potentials $(0, a, 1)$ and $0 < a < 1/2$, the minimum local improvement is a and can be arbitrarily small. Thus the graph geometry alone does not determine the ascent margin.

Comparison threshold. If $0 \le \eta < \epsilon$, take $a = (\epsilon - \eta)/4$ and $b = (\eta + \epsilon)/2$. On a two-node graph let $h_i = 0$, $h_j = -a$, and $\widehat{\Delta}_{ij} = b$. The comparison error is $(3\epsilon + \eta)/4 < \epsilon$, but the accepted move decreases potential. Thus the condition $\eta \ge \epsilon$ is needed for the stated worst-case monotonicity guarantee; a zero threshold is covered when the comparisons are exact.

Prediction target. Theorem 2 bounds Γ_L, the change in prior-weighted source-density potential. Prediction risk also depends on the conditional label distributions and the prediction rule. The source-odds identities normalize history weights over visited clients, so the density-potential bound alone gives no prediction-risk ordering between the entry head, terminal head, and history mixture.

D Experiment Details

The subsections below collect the data construction, communication graphs, optimization schedules, baseline provenance, sensitivity analyses, inference diagnostics, and real-data protocols used throughout Section 5.

D.1 CIFAR Data, Graphs, Baselines, and Traffic Construction

The benchmark uses the client construction from the released `FedDRM` code (Wang et al., 2026b). CIFAR-10 has ten labels, CIFAR-20 maps the CIFAR-100 fine labels to 20 superclasses, and CIFAR-100 retains all 100 fine labels. The official training and test collections are pooled before partitioning among eight clients using one of two label-skew constructions. In the SPC construction, label-sorted data are divided into equal-sized, label-homogeneous shards, and each client receives five uniformly sampled shards for CIFAR-10 or 25 for CIFAR-20/100, yielding equal client dataset sizes. In the Dirichlet construction, client-wise class proportions are randomly drawn from a Dirichlet distribution with concentration 0.3, yielding heterogeneous label marginals and unequal client dataset sizes. Each client's assigned images are shuffled and divided into 70% training and 30% evaluation data. The latter is split into 20% validation and 80% test data. Training, validation, and test indices are disjoint.

The eight deterministic client transformations combine $\gamma \in \{0.6, 1.4\}$, hue shift in $\{-0.1, 0.1\}$, and saturation in $\{0.5, 1.5\}$. The CIFAR models use a GroupNorm ResNet-18 followed by a 256-dimensional task projection with LayerNorm and a local classifier. Within each paired run, all clients share a random initialization; no pretrained `FedDRM` parameters are loaded. Both CIFAR training phases retain the released random-resized crop to 32×32 pixels with scale 0.9–1.0, followed by the fixed client color transformation and normalization. Evaluation uses one seeded crop per image, shared across paired models and readouts. Freezing fixes encoder parameters and buffers; it does not fix the augmented input at every training access. The fixed evaluation view and non-source entry assignments are shared between each 800+200 model and its `DFedAvgM` control. The source ID is used to construct the evaluation and for observer-only diagnostics; it is not given to the routing or history-weighting rule.

The Path has edges $\{\{i, i+1\} : 0 \le i < 7\}$, and the Cycle adds $\{7, 0\}$. The fixed RR4 graph has neighbor sets

$$\mathcal{N}(0) = \{1,2,3,4\}, \quad \mathcal{N}(1) = \{0,2,5,6\}, \quad \mathcal{N}(2) = \{0,1,3,7\}, \quad \mathcal{N}(3) = \{0,2,4,6\},$$
$$\mathcal{N}(4) = \{0,3,5,7\}, \quad \mathcal{N}(5) = \{1,4,6,7\}, \quad \mathcal{N}(6) = \{1,3,5,7\}, \quad \mathcal{N}(7) = \{2,4,5,6\}.$$

The Complete graph contains all 28 undirected edges. The sparse-graph movement budgets are the diameters 7, 4, and 2. The primary Complete control uses $L = 8$; $L = 1$ is retained for the

training-duration diagnostics. Previously visited experts are excluded from subsequent moves. No posttraining representation consensus is applied.

Thirty-two-client scaling. The CIFAR-10 scaling benchmark uses gamma levels $\{0.6, 1.0, 1.4\}$, hue shifts $\{-0.15, 0, 0.15\}$, saturation levels $\{0.4, 1.0, 1.6\}$, and two-bit posterization for clients 27–31. The pooled-data construction and within-client split remain the same. The connected degree-four graph uses seed 82025, has 64 edges and diameter four, and retains the primary movement budget $L = 8$.

Hospital-network construction. The 16/8/8 small/medium/large hospitals receive disjoint, exactly class-balanced subsets of the official CIFAR-10 training split with 250/1,000/4,000 examples each. The layered graph has a cycle within each tier, one small–medium edge per small hospital, and a one-to-one medium–large matching (56 edges). The shortcut graph adds one predetermined small–large edge per small hospital (72 edges). Each graph uses its own shared-training prefix and frozen-encoder continuation.

The common diagnostic panel contains 20 official test images per class. Each of the 200 distinct images is reused in 64 arrival records, yielding 12,800 records with exactly uniform entry counts across hospitals. Seeds 2025–2027 vary training and entry assignments. The prior-on and prior-off evaluations share the model, query panel, and arrival records; both exclude revisits and allow at most two referrals. The panel is an existing diagnostic set rather than a newly blinded test set. Workload quantities count the diagnostic calls made along these finite query paths.

Cumulative diagnostic workload. Let $\mathcal{V}_q$ be the visited hospitals for arrival q, and let $\mathcal{H}_g$ contain the hospitals in tier g. Each hospital supplies one prediction when visited, including the entry and terminal hospitals. Because the route excludes revisits, its cumulative call counts are

$$C_i = \sum_{q=1}^{Q} \mathbf{1}\{i \in \mathcal{V}_q\}, \qquad C_g = \sum_{i \in \mathcal{H}_g} C_i, \qquad C_{\text{all}} = \sum_i C_i = \sum_{q=1}^{Q} (T_q + 1),$$

where T_q is the number of moves and $Q = 12{,}800$. The reported tier share, peak individual share, and calls per arrival are, respectively,

$$W_g = \frac{C_g}{C_{\text{all}}}, \qquad P_{\max} = \frac{\max_i C_i}{C_{\text{all}}}, \qquad \overline{C} = \frac{C_{\text{all}}}{Q}.$$

Table 4 reports both the large-tier share W_{large} and $100C_{\text{large}}/Q$, retaining the absolute large-hospital workload scale as the total number of calls changes. It also reports $\overline{C}$ and $P_{\max}$ to distinguish additional collaboration from concentration at any single hospital. Each seed's counts and ratios are computed first, followed by their mean and sample SD across the three seeds. Per-hospital counts reconstructed from the saved paths match the recorded classifier calls. These quantities weight each classifier invocation equally.

Table 4: Accuracy and diagnostic workload in the hospital-network experiment. Large share and peak share are percentages of all visited-classifier calls; large calls/100 is the absolute number of large-hospital calls per 100 arrivals. Accuracy uses history-weighted predictions. Entries are mean ± sample SD over seeds 2025–2027; `EdgeDRM-NR` workload quantities are fixed by uniform entry.

Condition	Accuracy (%)	Calls/query	Large share (%)	Large calls/100	Peak share (%)
Layered: `EdgeDRM-NR`	45.29 ± 1.80	1	25	25	3.125
Layered: History, no prior	45.74 ± 1.60	2.01 ± 0.03	19.55 ± 0.44	39.22 ± 1.56	5.04 ± 0.44
Layered: History + prior	47.36 ± 3.06	2.61 ± 0.01	52.02 ± 0.12	135.52 ± 0.63	7.14 ± 0.20
Shortcut: `EdgeDRM-NR`	45.49 ± 2.03	1	25	25	3.125
Shortcut: History, no prior	46.12 ± 1.73	2.01 ± 0.02	20.36 ± 0.33	40.96 ± 0.39	4.61 ± 0.59
Shortcut: History + prior	47.93 ± 2.53	2.46 ± 0.01	69.49 ± 0.10	170.81 ± 0.81	9.77 ± 0.18

D.2 CIFAR Training, Evaluation, and Additional Results

The balanced-endpoint schedule below applies to the benchmark and hospital protocol. The readout diagnostic uses CIFAR-100/Dirichlet-0.3, the sensitivity sweep uses CIFAR-10/Dirichlet-0.3, and the

routing-cost table covers all six benchmark settings. All three use eight clients and seeds 2025–2027. Each primary model uses its saved 200-round head and router checkpoint with the original 800-round encoder state. The same continuation length is used for every graph and seed. Evaluation preserves the training random-number and loader states.

D.2.1 Training schedule and implementation details

The routing loss assigns equal weight to the two endpoint means of every edge. At inference, the deployed incident score adds the single sample-count offset $\log(n_j/n_i)$. The pretraining duration and subsequent encoder freezing are experimental optimization choices, rather than defining components of `EdgeDRM`.

Stage I: 800 shared-training rounds. All clients start from the same model initialization. Each round takes local SGD steps on the task loss and then mixes the full model $\psi_i = (\theta_i, \omega_i)$ by (4). Both representation and classifier are exchanged, using only physical neighbors. Thus *shared* refers to full-model collaboration toward a common task objective, not a server-owned model or identical local parameters. This stage uses the same graph and full-model mixing as the matched `DFedAvgM` control (Sun et al., 2023).

Stage II: 200 rounds with frozen local encoders. Each expert inherits its own Stage-I model and sampler state, fixes $\bar{\theta}_i = \theta_i^{(800)}$, and initializes its routing potential $a_i = (\vartheta_i, b_i)$ to zero. The encoder includes the backbone, feature projection, and feature normalization. Its parameters and persistent buffers remain unchanged, and it runs in evaluation mode. There are no encoder gradient updates or representation-mixing steps:

$$\theta_i^{(800+r)} = \bar{\theta}_i, \qquad r = 0, \ldots, 200.$$

Each local step minimizes (17) only over the private classifier ω_i and the endpoint potentials. Classifier heads are not exchanged, and potentials are not averaged. Freezing keeps each local feature map stationary during this stage; it does not align different experts' feature maps. The fitted representation used at inference is $\widehat{\theta}_i = \bar{\theta}_i$ under this experimental protocol.

The head and routing-potential updates instantiate Stage II of Section 3.1, with no parameter mixing. Routing gradients do not flow into the encoder or task head. Training counts are exchanged before coefficient fitting. After round 200, final potentials are published to neighbors. All parameters remain fixed during inference, and no additional consensus phase is applied.

Algorithm 2 Experimental 800+200 training schedule

Require: Graph G, neighbor weights W, local steps B, λ; $R_{\mathrm{sh}} = 800$, $R_{\mathrm{per}} = 200$
Initialize all local task models identically
for $r = 1, \ldots, R_{\mathrm{sh}}$ **do**
  Each client takes B task-only SGD steps
  Mix full models with physical neighbors using W
end for
Inherit local models and sampler states; set $\bar{\theta}_i \leftarrow \theta_i^{(800)}$
Freeze encoder parameters and buffers in evaluation mode; initialize owner potentials to zero and exchange training counts along edges
for $r = 1, \ldots, R_{\mathrm{per}}$ **do**
  for $b = 1, \ldots, B$ **do**
    Owners send potential snapshots to incident neighbors
    Each endpoint computes local features with its frozen encoder
    Differentiate task loss for its private head and balanced endpoint losses for endpoint potentials
    Return neighbor-owned potential gradients along their edges
    Update only private heads and owner potentials with accumulated gradients
  end for
end for
Publish final potentials to neighbors; freeze heads and routers

Fixed budget and learning-rate schedule. For the CIFAR protocol, each round has five local mini-batches of size 128. SGD uses momentum 0.9, reset each round, and weight decay 5×10^{-4}. The phase-local learning rate is

$$\alpha_r = 0.005\{1 + \cos(\pi(r-1)/H_{\mathrm{lr}})\}, \qquad H_{\mathrm{lr}} = 800 \text{ in Stage I}, \quad H_{\mathrm{lr}} = 600 \text{ in Stage II}. \tag{33}$$

Stage II restarts at 0.01 but stops at round 200, with $\alpha_{200} \simeq 0.00752264$; it does not run to the end of the 600-round decay horizon. The encoder remains frozen throughout this phase; the schedule applies only to heads and routing potentials. The model at total round 1000 is the fixed endpoint, without test-based early stopping. The finite neural optimizer with weight decay is distinct from the unpenalized fixed-representation estimator in Section 4.

This schedule keeps local features stationary during coefficient fitting, but does not make different clients' representations identical or independent of the reused training data. It therefore does not by itself satisfy the statistical assumptions of Theorem 1.

D.2.2 LOGICAL INFERENCE COST AT THE 800+200 ENDPOINT

Table 5 reports the incident edge-score evaluations required by the deployed routing rule across the six main benchmark settings. All entries aggregate the same three training seeds used in the main tables.

Table 5: Mean incident edge-score evaluations per query across the six benchmark settings. Entries are final-checkpoint means ± sample SD across seeds 2025–2027.

	CIFAR-10		CIFAR-20		CIFAR-100	
Graph	Dir-0.3	5-SPC	Dir-0.3	25-SPC	Dir-0.3	25-SPC
PATH	2.34 ± 0.04	2.34 ± 0.00	2.35 ± 0.01	2.37 ± 0.02	2.35 ± 0.02	2.39 ± 0.01
CYCLE	2.79 ± 0.05	2.79 ± 0.03	2.79 ± 0.01	2.81 ± 0.04	2.81 ± 0.01	2.82 ± 0.01
RR4	6.59 ± 0.05	6.55 ± 0.02	6.54 ± 0.02	6.55 ± 0.02	6.56 ± 0.01	6.59 ± 0.02
COMPLETE	12.75 ± 0.03	12.72 ± 0.02	12.63 ± 0.01	12.65 ± 0.02	12.66 ± 0.01	12.68 ± 0.01

There are three independent training seeds; graphs reuse each seed's partition. Validation/test query counts are 3,600/14,403 for seed 2025, 3,603/14,400 for seed 2026, and 3,601/14,402 for seed 2027. All controls match each readout's query indices, labels, source IDs, entry nodes, and shared-checkpoint hash. Accuracy is measured over the fixed non-source arrivals, with one view per query.

D.2.3 FIXED-EXPERT REFERRAL-SCORE ABLATION

The comparison in Figure 2 reuses the final CIFAR-10 and CIFAR-100 Dirichlet-0.3 RR4 checkpoints for all three seeds. No model parameter is updated. `EdgeDRM` routing uses the deployed incident score, including the single $\log(n_j/n_i)$ offset. `EdgeDRM-NR` reports the same entry expert without referral. Count-only replaces the learned part of every incident score by zero, leaving exactly $\log(n_j/n_i)$, and retains the learned rule's strict zero threshold. Random routing samples uniformly from the unvisited incident neighbors and stops only at the two-move budget or when none remain; its seed is fixed before evaluation. All three routing rules exclude revisits. The primary prediction is the terminal expert's classifier, which changes only through the selected route.

D.2.4 SENSITIVITY AND READOUT CONTROLS

Table 6: Task–routing weight on CIFAR-10/Dirichlet-0.3, eight clients, RR4, and $L = 2$. All branches reuse the same seed-specific 800-round prefix and fit heads and routers for 200 rounds. Accuracy is in percent, referral gain in percentage points, and calls per query. Entries are final-checkpoint means ± sample SD across seeds 2025–2027.

λ	`EdgeDRM`	`EdgeDRM-NR`	Referral gain	Calls
0.10	52.11 ± 2.19	35.01 ± 3.55	$+17.11 \pm 3.98$	2.16 ± 0.03
0.25	51.76 ± 2.04	30.29 ± 4.06	$+21.46 \pm 4.99$	2.16 ± 0.03
0.40	51.57 ± 2.03	28.36 ± 4.23	$+23.21 \pm 5.36$	2.15 ± 0.03
0.50	51.51 ± 2.04	27.48 ± 4.32	$+24.03 \pm 5.60$	2.15 ± 0.03
0.60	51.26 ± 2.04	26.75 ± 4.40	$+24.51 \pm 5.71$	2.14 ± 0.03
0.75	50.95 ± 2.09	25.96 ± 4.49	$+24.99 \pm 5.99$	2.13 ± 0.04
0.90	50.12 ± 2.21	25.35 ± 4.54	$+24.77 \pm 6.35$	2.11 ± 0.05
0.95	49.30 ± 2.33	25.18 ± 4.56	$+24.12 \pm 6.51$	2.09 ± 0.06
1.00	30.65 ± 2.81	24.97 ± 4.54	$+5.68 \pm 2.00$	1.93 ± 0.28

Table 7: Readout and inference-cost controls using identical frozen models and paths. Accuracy is in percent (three-seed mean $\pm$ sample SD). Hard terminal uses the final visited expert; uniform and weighted history use only visited predictions. Edge scores and classifier calls are mean logical operations per query, not measured latency.

Graph	Hard terminal	Uniform history	Weighted history	Calls	Incident scores
PATH	11.09 ± 0.60	10.43 ± 0.79	$\mathbf{11.19 \pm 0.58}$	1.72	2.35
CYCLE	12.29 ± 0.63	11.62 ± 0.79	$\mathbf{12.40 \pm 0.63}$	1.81	2.81
RR4	17.28 ± 0.55	15.62 ± 0.47	$\mathbf{17.59 \pm 0.59}$	2.10	6.56
COMPLETE	$\mathbf{20.05 \pm 0.55}$	17.71 ± 0.67	20.02 ± 0.63	1.94	12.66

The sweep in Table 6 is a sensitivity analysis rather than test-set model selection. Referral remains effective across the nonzero routing-loss settings, whereas removing the routing term at $\lambda = 1$ weakens performance.

Table 7 compares prediction readouts on identical routes. Weighted history performs best on the sparse graphs and is comparable to hard-terminal prediction on COMPLETE.

D.3 ISIC2019: DATA, PROTOCOLS, AND ADDITIONAL RESULTS

The study in Section 5.4 uses a fixed 400-round shared-training prefix followed by 50 rounds of balanced-endpoint frozen-encoder adaptation, with three training seeds.

D.3.1 SHARED DATA, PREPROCESSING, AND GRAPHS

Data provenance and center identities. We use the 23,247 images and six source-center assignments in the ISIC2019 split distributed with FLamby (du Terrail et al., 2022). Images come from the public ISIC2019 training collection, including BCN20000, HAM10000, and MSK contributions (Hernández-Pérez et al., 2024; Tschandl et al., 2018; Codella et al., 2018). In particular, the evaluation set here is FLamby's held-out subset of that collection, not the separate ISIC challenge test collection. Center identities follow source and acquisition metadata and should not be interpreted as six independent hospitals. The `FedSPM` package (Wang et al., 2026a) serves only as a reference for the data input pipeline; we do not use its within-server subpopulation model or its pooled resplitting procedure. Each observed center is represented by one expert.

Leakage control and fixed split. The released FLamby partition contains 18,597 training and 4,650 test images. Our metadata audit finds 2,302 lesion identifiers shared across these folds. We preserve every test image and label and remove the 6,102 training images whose lesion identifier also occurs in the test fold. We additionally link lesion identifiers through identical original or preprocessed image files and one visually confirmed near-duplicate pair. These links are computed globally, including across centers; groups with conflicting labels are excluded from the training candidates. The image-level audit removes four additional training candidates. A deterministic grouped split (seed 42) targets a 20% validation share within center–class cells, while reserving at least one training group for each represented cell. The resulting counts are given in Table 8. Training, validation, and test are disjoint in the audited lesion/duplicate groups. The grouping metadata do not establish patient-level independence, and the near-duplicate review is not an exhaustive guarantee against all transformed duplicates. This modified split is not directly comparable to the unmodified FLamby leaderboard.

Table 8: Fixed ISIC2019 split after global lesion and identified duplicate grouping. Excluded images come from the original training fold. The n_i used for the deployed sample-count offset are the final training counts.

Node	FLamby source	Train	Validation	Test	Excluded
0	BCN	3,802	954	2,483	5,174
1	vidir_molemax	2,528	632	791	3
2	vidir_modern	1,686	424	672	581
3	rosendahl	1,221	308	452	278
4	MSK4	520	131	164	4
5	vienna_dias	229	56	88	66
	Total	9,986	2,505	4,650	6,106

Images and model. We resize each RGB image with bilinear interpolation so that its shorter side is 224 pixels, apply the FLamby shades-of-gray color-constancy transform with power six, and save a JPEG with quality 75. Training uses a random 200×200 crop, independent horizontal and vertical flips, rotation within ± 50 degrees, and brightness/contrast jitter of 0.15/0.1. Validation and test use a deterministic center crop. All roles use ImageNet channel normalization. These torchvision transforms are shared by all methods; we do not claim exact reproduction of FLamby's augmentation code. The model is the existing `FedDRM` GroupNorm ResNet-18 with its stride-one stem, a $512 \to 256$ projection with LayerNorm/ReLU, and a private eight-class linear head. Within each seed, all six experts and all methods start from the same random initialization; no pretrained weights are used. `EdgeDRM` adds a 256-dimensional slope and scalar intercept per node.

Communication graphs. The four undirected graphs—PATH, CYCLE, RR3, and COMPLETE—are fixed before training. PATH has edges $\{i, i+1\}$ for $i = 0, \ldots, 4$, and CYCLE adds the edge $\{0, 5\}$. The intermediate regular graph is

$$E(\text{RR3}) = \{01, 04, 05, 12, 13, 24, 25, 34, 35\},$$

where ij denotes the unordered pair $\{i, j\}$. COMPLETE contains all pairs. Their edge counts are 5, 6, 9, and 15 and their diameters are 5, 3, 2, and 1. We use symmetric Metropolis mixing, with $W_{ij} = 1/(1 + \max\{d_i, d_j\})$ on an edge and the diagonal completing each row to one. Graph diameter alone does not ensure that the positive-gain, no-revisit referral rule reaches a globally best expert.

D.3.2 BALANCED-ENDPOINT FROZEN-ENCODER 400+50 PROTOCOL

Shared-training prefix. Each graph uses six identically initialized local models, three training seeds, and 400 rounds of task-only SGD followed by full-model Metropolis mixing. There are ten local optimizer steps per round, effective batch size 32, and microbatch size 16. The initial learning rate is 0.03 with a 400-round cosine horizon; SGD momentum is 0.9, reset each round, and weight decay is 5×10^{-4}. The RR3 prefix reuses the unchanged final checkpoint of the earlier `DFedAvgM` run, before its evaluation-only consensus. The other graphs require their own task-only prefixes; old joint-training `EdgeDRM` checkpoints are not substituted for them. The comparison uses fixed round 400, not the historical validation-selected checkpoint.

Frozen adaptation. The backbone, feature projection, feature normalization, and their buffers are fixed to the prefix checkpoint and evaluated in inference mode. Private classifier heads inherit their prefix values, and all node routing potentials start at zero. For 50 rounds, five local mini-batches per round update only the heads and balanced-endpoint routing potentials, using equal endpoint weights and $\lambda = 0.9$. Microbatch gradients are weighted by their fractions of the full batch. Endpoint owners accumulate both local and neighbor contributions from a common snapshot before updating. No encoder or head mixing is performed, and no extra consensus phase is added. The head/router learning rate starts at 0.01 and uses (33) with horizon 600, stopping at round 50; momentum, weight decay, and batch sizes are as above. Training retains random image augmentation despite the frozen representation parameters.

Selection, inference, and budget interpretation. Validation readouts are saved at adaptation rounds 0, 10, 25, and 50. Within each topology, test comparisons use the fixed prefix and round-50 endpoint without selecting a model from the test scores. The main table reports PATH, CYCLE, RR3, and COMPLETE; topology comparisons are descriptive rather than a preregistered test. The full prefix uses 4,000 local optimizer steps per client; adaptation adds 250 head/router steps. Thus comparison with `DFedAvgM400` estimates the change from adding adaptation and referral to that prefix, and does not establish superiority at equal total training effort. The adapted entry, hard terminal, uniform history, and score-weighted history controls use identical final model parameters and matched queries. The graph, split, loss weights, and inference protocol are fixed before each run. Each graph has three fixed training seeds.

Metric and access checks. For a prediction rule $\widehat{y}(x, s)$ and the test images $\mathcal{T}_c$ of class c, the uniform-entry metric is

$$\mathrm{BA}(\widehat{y}) = \frac{1}{8}\sum_{c=0}^{7}\frac{1}{6|\mathcal{T}_c|}\sum_{(x,y)\in\mathcal{T}_c}\sum_{s=0}^{5}\mathbf{1}\{\widehat{y}(x,s)=y\}.$$

For the primary rule, $\widehat{y}$ is the argmax of the visited-history probability mixture. Each of 4,650 unique test images contributes six entry records, for 27,900 records per model. These records are paired comparisons of repeated images, not independent test subjects. The router sees neither the test label nor the source-center identity. The deployed router adds the single sample-count offset $\log(n_j/n_i)$ to the balanced-endpoint logit, and the cumulative history weights use the resulting scores without another correction. All completed runs have zero terminal-neighbor prediction requests and exactly $T+1$ classifier visits for a T-move route. Source-checkpoint and encoder hashes are checked, and the saved paths, accepted scores, history weights, and probability vectors reproduce the reported BA values. No cross-seed SD is reported in the main table. Its zero threshold and local frozen representations do not establish the additional statistical or routing assumptions in Section 4.