BEYOND MEMBERSHIP: BLIND SPOT IN GRAPH UNLEARNING AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph unlearning is commonly considered successful when a membership inference attack performs no better than chance in determining whether a deleted node was used for training. For graph data, however, this criterion is incomplete. A node's label can often be inferred from its neighbors, which remain in the graph after deletion. The label of a deleted node may therefore remain recoverable even when a membership attack detects no evidence of training participation. In an exactly solvable model, we show that these risks are governed by a common quantity: the leave-one-out residual, which measures the part of a node's label that its neighborhood cannot explain. They move in opposite directions. When the neighborhood explains the label well, the node leaves little membership signal in the model, but its label remains easy to reconstruct from the retained graph. When the neighborhood is less informative, reconstruction becomes harder while membership leakage increases. This exposure persists even under exact retraining—the reference standard that approximate unlearning methods seek to match—because unlearning modifies only the model and the information lives in the retained graph. Across multiple unlearning methods and real-world graphs, our experiments support the theoretical prediction: membership attacks are least effective precisely where label-reconstruction risk is highest. Membership inference alone is therefore an incomplete audit of graph unlearning and should be paired with reconstruction-based evaluation.

1 INTRODUCTION

Graph neural networks predict a node's label by message passing: each layer aggregates representations from the node's neighbors and combines them with its own features (Kipf & Welling, 2017; Gilmer et al., 2017). The predictive signal is therefore split between two sources — what the model learned from the node's own record, and what the surrounding graph says about it. On homophilous nodes the second source is strong: a node's label is often recoverable from its neighbors' labels alone, before any model is trained (McPherson et al., 2001; Zhu et al., 2020). This split is the object of our study, because a deletion request removes the first source and leaves the second in place.

When a node asks to be forgotten, graph unlearning (GU) removes its contribution from a trained model, with retraining from scratch on the reduced graph as the reference (Ginart et al., 2019; Bourtoule et al., 2021). Success is then certified in two steps. First, the unlearned model is compared with the retrained one on utility, typically by matching test accuracy (Ding et al., 2026). Second, a membership-inference attack (MIA) is run on the deleted nodes to measure the forgetting quality by measuring the failure of an attacker to tell apart unlearned examples from never-seen examples Zhao et al. (2024). The graph community follows the same recipe to assess whether GU methods effectively prevent information leakage Ding et al. (2026). Han et al. (2026) also use an MIA at chance to read as the mark of a successful deletion. Used for what it measures, this is sound evidence: the deleted node is indistinguishable from one that was never trained on.

Our concern is with what this evidence omits. Indistinguishability from a never-trained node is all a chance-level MIA establishes; it does not imply that the node's label is hard to recover, because a never-trained node can itself be highly predictable once its neighborhood is known. The gap is one of scope. The audit examines the released model, while the adversary might retain the graph — the target's former neighbors and what is known about them — and on a homophilous graph, that

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

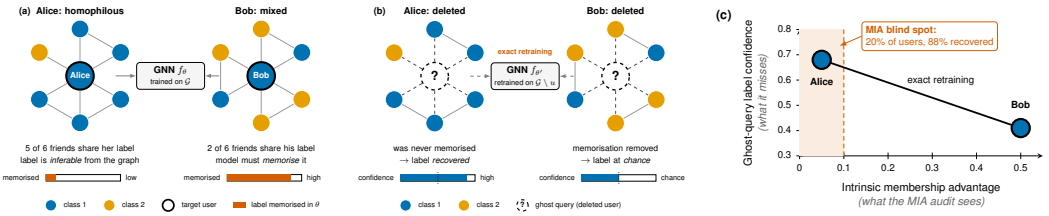


Figure 1: **Deletion removes what the model stored, not what the graph explains.** (a) Alice's neighbours share her label, so the model stores little about her; Bob's do not, so the model must memorise his. (b) After exact retraining, a ghost query on the archived friend list still recovers Alice's label but not Bob's. (c) The users whose membership audit clears most cleanly (shaded) are those whose label the neighbourhood gives away.

context alone often determines the target's label. Such scoping failures are familiar in privacy: the Netflix release met the anonymization contract it had set for itself, yet subscribers were re-identified through side information the contract never considered (Narayanan & Shmatikov, 2008). In graph unlearning, the side information is the retained graph itself. A subject whose deletion passes the audit may therefore be assured of erasure while remaining exactly as inferable as before.

Prior work has already seen the two lenses disagree. Structure is the dominant channel of GNN membership leakage (Olatunji et al., 2021), and Jamadandi et al. (2025) find that membership attacks succeed on low-homophily nodes and fall to chance on highly homophilic ones, concluding that low-homophily nodes "have a significantly higher risk of leaking." Through the membership lens, homophilous nodes are the safe ones. As noted above, they are also the nodes whose labels their neighborhoods most reliably determine.

Why should the two audits pull apart? Our answer is that a model can only remember what the data around a node fails to explain. Theorem 1 makes this exact for a linearized message-passing learner: both exposures are governed by one number, the *leave-one-out residual* which is the share of a node's label that its retained context cannot account for. Membership advantage rises with this residual; post-deletion reconstruction falls with it. The mechanism is simple. A learner that fits its labels stores the part its context cannot explain; presence inference reads that stored part, and reconstruction reads the explained part from the neighbors. For a homophilous node, the explained part dominates, so s_v is small; the model has almost nothing to forget, and the MIA is reassuring exactly because the neighborhood already gives the label away. This memorization is forced rather than a training defect, and the explained part outlives the model (Section 3).

Real graphs sit on this frontier. We audit $1{,}000$ deletion requests each on Pokec-n (Takáč & Zábovský, 2012; Dai & Wang, 2021) and Cornell5 (Traud et al., 2012; Lim et al., 2021), where the trained label — field of work, gender — is one a person would want forgotten. Ordered by how much of the label the retained graph leaves unexplained, the trained model's membership advantage climbs from at most 0.12 to 0.32–0.38, while one query to the *exactly retrained* model by anyone holding the user's friend list recovers the label with accuracy falling from 0.98–0.99 to 0.74–0.76: the users a membership audit clears most cleanly are those whose label a model that never saw them still gives away. Under the gold standard this is not approximation error; it is the tradeoff the theorem says must be there.

Our work does not challenge the guarantees of unlearning; it focuses on the limitation of those guarantees. Exact retraining removes everything the model stored about a node, and removes it perfectly — yet the node's label may remain readable from the retained graph, because that limit is set by the data and not the deletion mechanism. An audit confined to the model cannot detect this. Thus, unlearning the node is a promise about the model, not about the world on which the model was trained. This work tries to make that boundary precise and measurable.

Contributions.

1. **A residual-based exposure frontier.** For ridge regression on fixed message-passing representations, we derive a conditional frontier linking membership exposure before deletion to label reconstruction after supervision deletion through the leave-one-out residual.

2. **A structural floor under full node deletion.** In a contextual stochastic block model, we derive a mechanism-independent lower bound on label reconstruction from retained neighbor features and archived edges, without access to a trained model.
3. **Empirical evidence on real graphs.** On Pokec-n and Cornell5, lower original-model membership exposure is associated with greater post-deletion label recovery under exact retraining and four approximate unlearners.
4. **A two-axis audit of graph unlearning.** We pair membership inference with label reconstruction to assess detectable training participation and post-deletion label inferability separately.

2 Related works

Privacy leakage in GNNs and the structural channel. Privacy attacks on graph neural networks target membership, attributes, and graph structure (Dai et al., 2024; Guan et al., 2024). A recurring finding is that the graph itself is a major leakage channel: neighborhood structure can expose membership, edges, and node properties (Olatunji et al., 2021; He et al., 2023; 2021; Duddu et al., 2021; Zhang et al., 2022). For sensitive attributes, retained neighbors may be sufficient for inference, with exposure governed by local homophily (Yuan et al., 2024). Indeed, black-box access to a trained GNN provides little advantage over imputing the attribute directly from the released neighborhood (Olatunji et al., 2023). These results establish a participation-independent attribute channel: a node's attribute may remain recoverable from retained neighbors even if the node never contributed to training. We study whether this channel persists after unlearning.

Graph unlearning and attacks on unlearned models. Machine unlearning seeks to remove a training record's influence from a deployed model, with retraining on the retained data as the reference solution (Ginart et al., 2019; Bourtoule et al., 2021). This is harder on graphs because a node's features, label, and edges influence its neighbors through message passing. Graph-unlearning methods therefore use graph partitioning, approximate parameter updates, or deletion operators that explicitly remove the node's neighborhood influence (Chen et al., 2022; Wu et al., 2023b;a; Cheng et al., 2023). Existing attacks ask whether the unlearning process itself reveals the deleted record. They exploit differences between the original and unlearned models, or residual information left by approximate unlearning (Chen et al., 2021; Song & Palanisamy, 2026; Zhang et al., 2026). Thus, these attacks are update-dependent and target the deleted node's residual influence. They do not address our setting: a single model obtained through exact retraining. We ask what this model, together with the retained neighborhood, still reveals about a node that did not participate in its training.

Auditing unlearning: what "privacy" is measured. Unlearning is commonly evaluated by comparing the unlearned model with a retrained reference, using aggregate performance or membership inference. Recent work strengthens this evaluation through per-sample audits that separate unlearning efficacy from residual privacy risk and show that aggregate agreement is insufficient (Naderloui et al., 2025). Under this notion of privacy, exact retraining is sound by construction because the deleted record leaves no participation-dependent information in the model. On graphs, membership-based auditing is complicated by relational dependencies and by the weak residual signals left after unlearning (He et al., 2023; Zhang et al., 2026). Nevertheless, these audits share the same estimand: whether the deleted record's *participation* remains detectable. They do not measure whether its sensitive attribute remains recoverable from retained graph structure, despite the established distinction between membership and attribute inference in GNN privacy research (Dai et al., 2024; Guan et al., 2024).

Gap. Together, these results expose a blind spot. Existing unlearning audits measure whether information arising from a deleted node's participation remains in the released model (Naderloui et al., 2025). Yet the node's label — here a sensitive attribute the model was trained to predict — may remain recoverable from its retained neighborhood even when that participation is removed exactly (Yuan et al., 2024; Olatunji et al., 2023). This structural leakage is neither an artifact of the update nor a residual error from approximate unlearning, and it is not captured by membership-based audits. We address this gap by evaluating exact retraining under two distinct estimands: membership indistinguishability and label recoverability. We show that the former can hold while

the latter persists, and that the two are coupled through how much of the label the neighborhood explains. Our analysis does not challenge existing unlearning guarantees; it measures a privacy quantity they were not designed to capture.

3 THEORETICAL ANALYSIS

The introduction argued informally that a model can only store what a node's neighborhood fails to explain. In this section we analyze two privacy exposures of a training node in an exactly solvable model: detectability of its *presence* in training (membership inference) and recoverability of its *label* after deletion (reconstruction). Two results anchor the analysis. Under *supervision deletion* (only the training example removed), both exposures are exact monotone functions — in opposite directions — of a single scalar, the leave-one-out residual, along a node-specific, strictly decreasing frontier, and we characterize exactly when both attacks can succeed at once (Theorem 1). Under *full node deletion* (features and incident edges removed too) we prove a *structural floor* on reconstruction that no deletion mechanism can lower (Proposition 1). The ingredients are deliberately classical, the leave-one-out algebra is the Cook (1977) distance, the prediction shift is the memorization score defined in Feldman (2020). The membership–attribute coupling goes back to Yeom et al. (2018), but that coupling makes both exposures grow *together* with overfitting at training time, post-deletion they are exactly *anti*-monotone, with a closed-form frontier and noise threshold; the floor adds that inference from population structure is not a deletion failure. Formal assumptions, proofs, and supporting results are deferred to Appendix A.

3.1 PRELIMINARIES AND THREAT MODEL

Data, learner, deletion. A graph $G = (V, E)$ with node features X, a labeled subset $L \subset V$, binary labels $y_u \in \{\pm 1\}$, and a target $v \in L$ with secret label y_v. Each node u has a representation $\phi_u \in \mathbb{R}^d$ from a fixed message-passing map applied to (G, X); fixing the representation and training only a linear head keeps deletion exactly computable while retaining graph dependence — the linearized (NTK) regime of a GNN — and structure enters only through the joint law of $(\phi_u, y_u)_u$. The learner is ridge regression on L: $w = H^{-1}\Phi^\top y$ with $H := \Phi^\top\Phi + \lambda I$, $\lambda > 0$, Φ stacking $(\phi_u)_{u\in L}$, and predictions $f(u; w) = w^\top \phi_u$. Deletion yields the refit w_{-v} (ridge on $L \setminus \{v\}$, same λ) — exact for supervision deletion; full node deletion also perturbs retained representations (perturbation budget in Remark 10). Ridge is deterministic, so exact unlearning returns w_{-v} itself.

Observation channel. Outputs are released through a Gaussian channel after clipping to the label scale: a query at v returns $z = \mathrm{clip}(f(v; \cdot)) + \mathcal{N}(0, \sigma^2)$, $\mathrm{clip}(t) := \max\{-1, \min\{1, t\}\}$, $\sigma > 0$ — the games' sole source of randomness, so all statements below condition on the realized dataset with probability over the channel. Clipping never increases the membership advantage and closes an *overshoot* loophole (Remark 2).

Membership game. A fair bit b selects the queried model: the trained w ($b = +$) or the refit w_{-v} ($b = -$). The attacker sees one noisy query z at v, may know everything else — including y_v, the strong-adversary convention — and guesses b; its advantage $\mathrm{Adv}_{\mathrm{MIA}} := \mathrm{TV}\big(\mathcal{L}(z \mid b{=}+),\, \mathcal{L}(z \mid b{=}-)\big)$ is the total variation between the query's law under w and under w_{-v}.

Reconstruction game. After deletion is honored exactly, the adversary sees one noisy query z_- of the retained prediction (under node deletion, via a ghost query on archived context) and guesses $\hat{y} = \mathrm{sign}(z_-)$, with accuracy $a_{\mathrm{rec}} := \mathbb{P}[\hat{y} = y_v]$. This sign attacker is fixed and not optimal (see Remark 3), and "reconstruction succeeds" claims below are unconditional lower bounds, while "reconstruction fails" claims are attacker-specific. The two threat models are deliberately asymmetric — reconstructing a label the adversary already holds is vacuous — and the population statements keep the label-informed membership convention explicit (Appendix A.7).

Assumption 1 (No anti-prediction / alignment)**.** The retained fit does not point against the true label: the remaining data may explain nothing about y_v, but never systematically suggests its opposite. The Bayes predictor satisfies this in expectation; failure modes are classified in Remark 6 and Proposition 2. (Formal version, together with the near-interpolation assumption used only by the appendix corollaries: Appendix A.1.)

3.2 The Presence–Label Tradeoff under Supervision Deletion

Two data-dependent scalars govern everything: the *leverage* $h_v := \phi_v^\top H^{-1}\phi_v \in [0,1)$, measuring how strongly v's own label pulls the model's prediction at v, and the *leave-one-out residual* $r_v := y_v - w_{-v}^\top \phi_v$, the part of the label the retained data cannot explain. Under Assumption 1 the *released* unexplained fraction is $\tilde{s}_v := \max\{y_v r_v, 0\} \in [0,1]$, equal to $y_v r_v$ except when the retained fit overshoots the label, where clipping zeroes the released gap.

Every label splits as $y_v = w_{-v}^\top \phi_v + r_v$: an *explained* component the retained data already predicts, and an *unexplained* component only v supplies. The split is operational (Lemma 1): adding v moves the weights by exactly $\|w - w_{-v}\|_H = \sqrt{h_v}\,|r_v|$ and the prediction at v by $h_v r_v$. The membership attacker reads this shift — the stored, unexplained component; the reconstruction attacker reads the surviving prediction — the explained component. An interpolating learner ($h_v \to 1$, the regime of overparameterized models trained to fit their labels) fits y_v exactly, so it must store the entire unexplained component in its weights; since the components sum to the label, whatever one attacker gains at a given node, the other loses there.

Theorem 1 (Residual frontier)**.** *Under the setup and Assumption 1, the optimal membership advantage and the sign attacker's reconstruction accuracy are exactly* $\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\big(h_v\tilde{s}_v/2\sigma\big) - 1$ *and* $a_{\mathrm{rec}} = F_{\mathcal{N}}\big((1-\tilde{s}_v)/\sigma\big)$*, with* $F_{\mathcal{N}}$ *the standard Gaussian CDF: both are monotone in the single scalar* $\tilde{s}_v$*, in opposite directions, so the pair lies on the strictly decreasing curve*

$$\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\Big(\frac{h_v}{2\sigma}\big(1 - \sigma\, F_{\mathcal{N}}^{-1}(a_{\mathrm{rec}})\big)\Big) - 1.$$

Joint success is noise-limited: for $\delta \in (0, \frac{1}{2})$*, writing* $z_\delta := F_{\mathcal{N}}^{-1}(1-\delta)$*, some* $\tilde{s}_v \in [0,1]$ *attains both* $a_{\mathrm{rec}} \geq 1-\delta$ *and* $\mathrm{Adv}_{\mathrm{MIA}} \geq 1-\delta$ *if and only if* $\sigma \leq \big(z_\delta + 2z_{\delta/2}/h_v\big)^{-1}$.

The proof (Appendix A.4) rests on a *signal budget*: the mean gap read by the membership attacker and the margin read by the reconstruction attacker satisfy $h_v\tilde{s}_v + (1-\tilde{s}_v) \leq 1$ — disjoint shares of the label's unit scale — and σ only sets the exchange rate to success probability; the identity, not the Gaussian forms, is the release-invariant content (Remark 4). Two scope notes. The frontier is *per node*: h_v indexes the curve, so across nodes both exposures can rise together (a high-leverage explained node can dominate a low-leverage memorized one on both counts); the tradeoff is along $\tilde{s}_v$ at fixed leverage. And it is *per release*: query averaging lowers the effective σ past the threshold, so privacy claims must declare their noise scale and release interface (Remark 11). Table 1 summarizes the regimes; Appendix A.5 gives the exact conservation law at interpolation, and Appendix A.6 the status of each direction.

From residuals to structure. The frontier makes a node's exposure a function of its residual; the natural next question is *which* nodes have small ones. That depends on the data distribution: treat the data as random and compare the learner to the best possible predictor of the label from what survives deletion — the Bayes predictor. Two facts follow (Proposition 3, Appendix A.7). First, the mean-square residual is exactly the Bayes-unexplainable share of the label plus the learner's own excess over Bayes, so at least that share is supplied by the node alone and, for ridge, stored in the weights (Lemma 1): memorization is forced, not chosen. Second, in the contextual stochastic block model (CSBM) of the next subsection the share is computable: exponentially small in homophily$^2\times$ degree for learners close to Bayes, and total when structure is silent — so homophily and *exploitable* heterophily act only through the homophily magnitude, the sign mattering only through alignment, since a homophily prior fit to heterophilous structure anti-predicts (Proposition 2). Alignment itself is generic: atypical label draws merely realize the frontier and are budgeted by a population frontier needing no per-realization alignment (Theorem 2); only systematic, adversary-known sign errors break the tradeoff, and they are certifiably far from Bayes (Remark 6).

3.3 Reconstruction Floor under Full Node Deletion

Under full node deletion the exact algebra above no longer applies (Remark 10), but the practical question survives: *after deletion is honored exactly, how much of* y_v *is still recoverable from retained data alone?* We answer in a contextual stochastic block model.

Definition 1 (CSBM and structural channel)**.** Labels $y_u \in \{\pm 1\}$ are i.i.d. uniform (balanced in expectation, not by constraint; Remark 7). Conditionally on labels, edges are independent with intra-

Presence (MIA adv.)	**Label** (recon.)	**Regime**	**Learner behavior**	**Typical node**
low	high	$\tilde{s}_v \approx 0$, any h_v	standard training; retained context already explains the label	strongly homophilous neighborhood
high	low	$\tilde{s}_v \approx 1$, $h_v \approx 1$	interpolating fit of an unexplained label (forced memorization)	mixed/ambiguous neighborhood (retained prediction ≈ 0)
low	low	$\tilde{s}_v \approx 1$, $h_v \approx 0$	heavy regularization; the model declines to fit the label	ambiguous neighborhood†
high	high	$y_v r_v > 1$ (anti-prediction), h_v away from 0; or tiny σ	misaligned: smoothing architecture fits structured heterophily with the wrong sign; a sign-aware adversary flips the rule	consistently heterophilous neighborhood

Table 1: The four exposure regimes, at channel noise small on the unit label scale (e.g. $\sigma \approx 0.2$–0.3; "high"/"low" are relative to the range attainable at that σ). Rows 1–3 satisfy Assumption 1; rows 1–2 are the frontier of Theorem 1: under interpolating training exactly one leakage is high, decided by the residual alone. Rows 3–4 each require a nameable deviation: declining to fit (row 3, a utility cost) or anti-predicting exploitable structure (row 4, Proposition 2; clipping attenuates deeper excursions $y_v r_v > 2$). †Row 3 additionally requires an unpredictable node: on a predictable one, a context-only adversary reconstructs the label regardless of the model (Proposition 1).

/inter-community probabilities p and q, $p + q > 0$; write $\gamma := \frac{p-q}{p+q} \in [-1, 1]$ ($\gamma > 0$: homophily; $\gamma < 0$: heterophily). Features are $x_u = \mu\, y_u\, v_\star + \xi_u$ with $\|v_\star\| = 1$, $\mu \geq 0$, $\xi_u \sim \mathcal{N}(0, \nu^2 I_d)$ i.i.d., $\nu > 0$. The *structural channel* at v is the masked one-hop mean $\phi_v^{\mathrm{struct}} := \frac{1}{D_v} \sum_{u \in N(v)} x_u$, $D_v := |N(v)| \geq 1$: v's own features are masked, so everything it knows about y_v arrives through edges, at signal-to-noise of order $\mu|\gamma|\sqrt{D_v}/\nu$ (exact form in Remark 7). Structural results assume the adversary knows $v_\star$ and the sign of γ (Assumption 8).

The ghost query is structure-only: an adversary holding an archived edge list $N(v)$ can rebuild the masked aggregate ϕ_v^{struct} of Definition 1 from *retained* neighbor features — it needs neither x_v nor any deleted model state; in fact, no model at all.

Proposition 1 (Structure-only reconstruction floor)**.** *In the CSBM, under Assumption 8 and conditionally on $D_v \geq 1$, the sign test applied to the rebuilt ϕ_v^{struct} achieves $a_{\mathrm{rec}}^{\mathrm{struct}} \geq \max\{\frac{1}{2},\ 1 - e^{-D_v\gamma^2/8} - F_{\mathcal{N}}\big(-\mu|\gamma|\sqrt{D_v}/2\nu\big)\}$, irrespective of the unlearning mechanism, of the channel noise σ, and of whether any model is released at all.*

This is a floor, not the total exposure: the model-mediated query and the structural probe are separately available, computed on different conditioning bases, and a joint adversary combining complementary evidence does strictly better (Remark 9); the frontier governs only the model-mediated term. The regimes now read: at high $|\gamma|\sqrt{D_v}$, v sits at the label-leaking, presence-private end of the frontier, and the label was close to public both before and after deletion — it imprints on *which* nodes are v's neighbors, which no unlearning of the model touches — so deletion protects presence only. At low $|\gamma|\sqrt{D_v}$, structure is silent, and the residual is near one (given a weak or self-masked own-feature channel: supervision deletion retains x_v with signal μ/ν, Remark 8), so the label is memorized and its presence exposed at interpolation; exact deletion then protects both, at the price that the model had learned the label only by memorizing it. In short: node deletion erases what the *model* stored about y_v (the unexplained component) but not what the *graph* stores (the explained component, readable from neighbors at rate $\mu|\gamma|\sqrt{D_v}/\nu$ — forever).

4 Experiments: the frontier on real graphs

Theorem 1 predicts something that can be checked on a served model without opening it: presence exposure and post-deletion label exposure are two readings of one scalar per node, the unexplained

share $\tilde{s}_v$, and move in opposite directions. We test it on two social networks, with the trained label as the quantity a user would want forgotten: field of work on Pokec-n, gender on Cornell5.

4.1 SETUP

Graphs, cohort, worlds. Pokec-n (66,569 users) and Cornell5 (18,660) carry near-balanced labels predicted from profile features that exclude the label and every sensitive column; a Pokec-n user has a median of one labeled friend, a Cornell5 user 61 (Appendix B.1). From the training users we draw 1,000 *candidates* — the users who request deletion and the only users measured — and 1,000 *anchors*, held out of every model. A *world* trains a two-layer GCN (Kipf & Welling, 2017) on the retained graph plus a random half of the candidates; each candidate is a training member in 32 of the 64 worlds per dataset (32 shadow, 16 calibration, 16 evaluation; Appendix B.2).

Releases. Each world serves seven models: **No-op**, the trained model (the theorem's w); **GOLD**, retrained from scratch without the deleted candidates (exact deletion, the law of w_{-v}, the reference for every unlearner); **Never**, an independent second retrain whose gap to GOLD is the noise floor of every number below; and four unlearners applied to No-op — **GIF** (Wu et al., 2023a), **MEGU** (Li et al., 2024), **GNNDelete** (Cheng et al., 2023) and **CGU** (Chien et al., 2023) — each tuned per dataset on development worlds disjoint from the audit (Appendix B.2).

Membership audit. A per-user test in the style of LiRA (Carlini et al., 2022) on the logit-scaled confidence in the user's true label, with per-user Gaussians fitted on the shadow worlds, the threshold calibrated on the calibration worlds and the evaluation worlds scored once (Appendix B.3). Per user we also use the *membership advantage* $\mathrm{Adv}_v = 2\,\mathrm{AUROC}_v - 1$ of the same score over the 48 shadow and calibration worlds (0: the release's output at v does not depend on whether v was trained on; 1: perfect separation). On No-op, Adv_v compares w with w_{-v} and is the theorem's $\mathrm{Adv}_{\mathrm{MIA}}$; on a release it is the presence that release left behind. We call $|\mathrm{Adv}_v| \le 0.1$ the *pass band*, the per-user counterpart of an AUROC within 0.05 of chance.

The ghost query. The reconstruction adversary of Section 3 sees one output of the released model at the deleted node. Under node deletion it must rebuild that node: the user is re-inserted with a *blank* profile on the friendships the adversary archived and the release is queried once; the guess is the label with the larger posterior, and with $\hat{p}_v$ the posterior on the true label the *margin* $m_v^{\mathrm{ghost}} = 2\hat{p}_v - 1 \in [-1, 1]$ is that posterior on the label scale of Section 3. Nothing else is used — no other users' labels, no shadow models, no model internals — so this is the adversary anyone holding the friend list can run, and exactly the sign attacker of Theorem 1. We measure $a_{\mathrm{rec}}(v)$ as the fraction of the 8 evaluation worlds in which v was deleted where the guess is right and $\bar{m}_v^{\mathrm{ghost}}$ as the mean margin over them; the two axes of every figure below therefore share no model. Two baselines that use no model calibrate the ghost: the *class prior* (52.7% on Pokec-n, 50.3% on Cornell5) and a *model-free attacker*, a logistic decoder over the labels of the user's retained friends and friends-of-friends, fit on the anchors (Appendix B.4).

4.2 RESULTS

Which releases pass. Table 2 gives the audit's verdict: GOLD and Never sit at chance, No-op is traceable (0.62; 0.68), and of the unlearners GIF and CGU pass on Pokec-n and CGU alone on Cornell5 — the rest fall short despite dedicated tuning (Appendix B.2). We carry every release through the section, since the theorem is about nodes, not unlearners: a failing release shows its residue moving along the frontier, and a passing one is exactly where the label axis must be read, because membership is silent there.

Table 2: Membership AUROC per release (16 evaluation worlds; brackets: 95% CI over 8 world pairs). Red: passes (within 0.05 of GOLD). TPR and utility: Table 5.

Release	Pokec-n	Cornell5
No-op	0.619 [.611, .626]	0.680 [.672, .690]
GIF	0.519 [.506, .530]	0.666 [.657, .676]
MEGU	0.560 [.552, .579]	0.599 [.589, .616]
GNNDelete	0.729 [.714, .747]	0.697 [.691, .705]
CGU	0.493 [.483, .502]	0.506 [.494, .517]
GOLD	0.498 [.489, .507]	0.506 [.500, .512]
Never	0.500 [.494, .507]	0.497 [.489, .506]

Exact deletion follows the frontier. Figure 2 plots every deleted user under GOLD at the trained model's membership advantage against the ghost's confidence

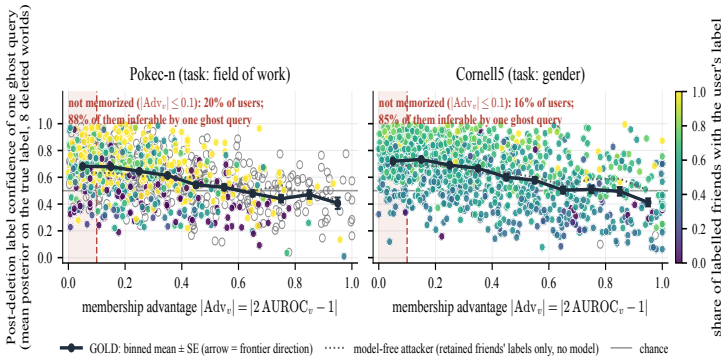


Figure 2: **Exact deletion follows the predicted frontier.** One point per deleted user under GOLD (exact retraining). x: how strongly the *trained* model's output at the user depended on whether the user was trained on (its membership advantage $|\mathrm{Adv}_v|$; 0 = not at all). y: the ghost query's confidence in the true label after exact deletion, mean over the user's 8 deleted worlds. Color: share of labeled friends with the user's label (hollow: none). Navy: binned mean $\pm$ SE; dotted: the model-free attacker (retained friends' labels only); red band: $|\mathrm{Adv}_v| \leq 0.1$.

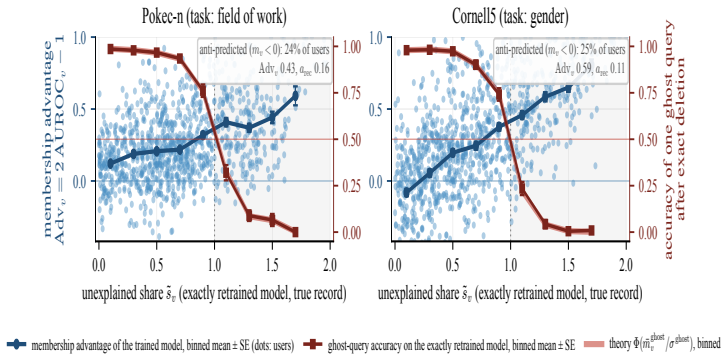


Figure 3: **One scalar, two directions.** Each deleted user is placed at its unexplained share $\tilde{s}_v$, read from the exactly retrained model (GOLD) shown the user's true record: 0 when the retained data fully explain the label, 1 when they explain none of it, > 1 when they predict the wrong label. Blue, left: membership advantage Adv_v of the *trained* model (No-op), signed. Red, right: accuracy of one ghost query on the *exactly retrained* model (GOLD); shaded: the theorem's $F_{\mathcal{N}}(\bar{m}_v^{\mathrm{ghost}}/\sigma)$. The two curves — the *presence blade* and the *label blade* — open like scissors. Grey: anti-predicted users.

in its true label. The curve falls on both graphs: the mean posterior on the true label drops from 0.68 in the first bin to 0.41 in the last on Pokec-n and from 0.72 to 0.41 on Cornell5. At the left are the users whose presence the trained model barely encoded: 20% of Pokec-n and 16% of Cornell5 candidates have $|\mathrm{Adv}_v| \leq 0.1$ under No-op, and one ghost query on the *exactly retrained* model recovers the label of 88% and 85% of them in at least six of eight deleted worlds — against 53% and 50% for the class prior, and 80% and 82% for the model-free attacker on the same users. At the right, the users the trained model had to memorize are at or below chance after exact deletion, and the model-free attacker (dotted) follows the same path: the drop is not a property of the model. The reason is the theorem's: when the retained neighborhood explains a user's label, the model never needed to store it, presence inference has nothing to detect, and the neighborhood explains the label still after the user is gone; when it does not, the model must store the label (Proposition 3), presence inference detects the stored part, and exact deletion removes it. On Pokec-n, where the label is homophilous, agreement with labeled friends grades the frontier from top-left to bottom-right; on Cornell5 gender agreement is near chance for everyone yet the ghost still recovers 72% of labels — the explained component travels through friends' features, a channel the label histogram of Definition 1 does not see. All seven releases show GOLD's shape (Appendix B.7): the ghost reads the explained component whichever model serves it.

One scalar, two directions. Figure 2 orders users by presence exposure; the theorem says both exposures are set by the unexplained share. We measure it as $\tilde{s}_v = \max\{y_v r_v, 0\}$, with r_v the residual of the exactly retrained model on the user's true record averaged over the 48 shadow and calibration worlds (Appendix B.4), and plot both exposures against it in Figure 3: the membership advantage of the trained model, which we call the *presence blade*, and the ghost's accuracy on the exactly retrained model, the *label blade*; the two curves open like scissors. On the aligned users ($\tilde{s}_v \leq 1$; 76% and 75% of the cohort) the membership advantage rises from 0.12 to 0.32 on Pokec-n and from -0.08 to 0.38 on Cornell5, while the ghost's accuracy after exact deletion falls from 0.99 to 0.76 and from 0.98 to 0.74: the scissors of Theorem 1. The theoretical curve $F_{\mathcal{N}}(\bar{m}_v^{\text{ghost}}/\sigma)$, with σ the seed-to-seed spread of the ghost margin, lies on the measured points; the scatter of the blue blade is the leverage h_v, and the stored shift $h_v\tilde{s}_v$ reproduces each user's advantage through the Gaussian channel to within 0.04 (Appendix B.8). The grey region is a third regime, not an exception: the quarter of users whose retained fit points *against* their label ($y_v r_v > 1$; Table 1, row 4). There the sign attacker is wrong (0.16; 0.11) while presence exposure is highest (0.43; 0.59), so a sign-aware adversary recovers the label 84% and 89% of the time and both exposures are high at once — exactly where Assumption 1 fails. Because $\tilde{s}_v$ is read from GOLD's output, we also order users by a quantity that involves no model: the posterior the model-free attacker assigns to the true label from retained friends' labels alone. Both directions persist (Appendix B.6): the trained model's advantage falls from 0.47 to 0.20 on Pokec-n and from 0.48 to 0.26 on Cornell5, while the ghost's accuracy after exact deletion rises from 0.24 to 0.98 and from 0.22 to 0.97.

What unlearning moves, and what it cannot. Figure 4 (Appendix B.5) repeats both blades for every release. The presence blade responds to the unlearner: GOLD, Never and CGU are flat at zero for every $\tilde{s}_v$; GIF is flat on Pokec-n (own advantage 0.05 vs. No-op's 0.26) and equals No-op on Cornell5 (0.25 vs. 0.26; its update is inert); MEGU retains roughly half (0.15; 0.17); GNNDelete's is *negative* on explained users (-0.49; -0.33 at $\tilde{s}_v < 0.4$) and positive on unexplained ones ($+0.15$; $+0.26$ at $\tilde{s}_v \geq 0.8$): its operator over- and under-corrects. The label blade does not respond: the seven curves lie on GOLD's on both graphs, and the only release-dependent part is a residue on the most unexplained users ($\tilde{s}_v \geq 0.8$), where No-op keeps $+2.9$ and $+5.3$ pp over GOLD, inert GIF on Cornell5 $+5.9$ pp, and every other unlearner sits at or below GOLD. Together the blades give the central observation about auditing: where the label is most inferable — $\tilde{s}_v \leq 0.2$ — the presence blade is closed for every release including No-op (Adv_v 0.12; -0.08) while one ghost query recovers the label 99% and 98% of the time. A membership audit is therefore silent exactly where inferability is greatest, because there was nothing for deletion to remove. What it certifies, unlearning delivers; the explained component is a property of the retained graph, which no parameter update changes.

5 Conclusion

Membership audits and label reconstruction ask different questions about a deleted node, and on graphs their answers disagree most where it matters. The nodes the model barely memorized (20% of Pokec-n, 16% of Cornell5) pass the membership audit under every release, yet after exact retraining one query on an archived friend list recovers their label 88% and 85% of the time. The pattern holds for exact retraining and four approximate unlearners alike, so it comes from the graph, not from imperfect deletion; for a linearized GNN, the leave-one-out residual explains it. The theory is exact for a linear head under supervision deletion, whereas the experiments use full node deletion on a nonlinear GCN and cover two graphs and one architecture. Membership inference remains the right test for participation. Alongside it, an audit should add one ghost query per deleted node, report recovery per node rather than on average — exposure concentrates where the membership audit is quiet — and compare against never-trained nodes, separating what deletion should remove from what the graph reveals about anyone. A passing membership audit shows only that the model has forgotten. **Unlearning can remove what the model remembers, but not what the retained graph already reveals.**

6 AI USE STATEMENT

In this work, we used large language models (LLMs) to edit the text of the paper for readability, to assist with LaTeX formatting, and to identify and summarize relevant literature, and support the review process by helping identify potential edge cases, inconsistencies, and opportunities to strengthen the rigor and clarity of the theoretical presentation. We did not use LLMs to develop the method or its theoretical framework, to formulate or prove the mathematical claims, to design the experiments, or to interpret the results; generating synthetic data, translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used LLMs as a coding assistant for the implementation and for the scripts that render the figures and tables. We reviewed all AI-assisted output: every edited passage was checked against the intended meaning, every reference suggested by an LLM was verified against the original source, and LLM-assisted code was read, tested and verified for correctness by the authors. We take full responsibility for the final content of this work, including any text or code prepared with the aid of generative AI.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 896–911, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484756. URL https://doi.org/10.1145/3460120.3484756.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 499–513, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3559352. URL https://doi.org/10.1145/3548606.3559352.

Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. GNNDelete: A general strategy for unlearning in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=X9yCkmT5Qrl.

Eli Chien, Chao Pan, and Olgica Milenkovic. Efficient model updates for approximate unlearning of graph-structured data. In *International Conference on Learning Representations (ICLR)*, 2023.

R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1): 15–18, 1977. doi: 10.1080/00401706.1977.10489493. URL https://doi.org/10.1080/00401706.1977.10489493.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 2021.

Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6):1011–1061, Dec 2024. ISSN 2731-5398. doi: 10.1007/s11633-024-1510-8. URL https://doi.org/10.1007/s11633-024-1510-8.

Pengfei Ding, Yan Wang, and Guanfeng Liu. Re-understanding graph unlearning through memorization. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 2940–2951, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792383. URL https://doi.org/10.1145/3774904.3792383.

Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous '20, pp. 76–85, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450388405. doi: 10.1145/3448891.3448939. URL https://doi.org/10.1145/3448891.3448939.

Bowen Fan, Yuming Ai, Xunkai Li, Zhilin Guo, Rong-Hua Li, and Guoren Wang. OpenGU: A comprehensive benchmark for graph unlearning. *arXiv preprint arXiv:2501.02728*, 2025.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pp. 954–959, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384290. URL https://doi.org/10.1145/3357713.3384290.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf.

Faqian Guan, Tianqing Zhu, Wanlei Zhou, and Kim-Kwang Raymond Choo. Graph neural networks: a survey on the links between privacy and security. *Artificial Intelligence Review*, 57(2):40, Feb 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10656-4. URL https://doi.org/10.1007/s10462-023-10656-4.

Haoke Han, Yan Huang, Zhenzhen Xie, and Junjie Pang. Federated illusion: Multi-level geometric privacy audit for federated graph unlearning. *Information*, 17(5), 2026. ISSN 2078-2489. doi: 10.3390/info17050424. URL https://www.mdpi.com/2078-2489/17/5/424.

Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing links from graph neural networks. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2669–2686. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/he-xinlei.

Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-level membership inference attacks against graph neural networks, 2023. URL https://openreview.net/forum?id=XKHNu4OF6wn.

Adarsh Jamadandi, Jing Xu, Adam Dziedzic, and Franziska Boenisch. Memorization in graph neural networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 29513–29550. Curran Associates, Inc., 2025. doi: 10.52202/085713-0988. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2a54def490213ee10631b991c5acc6b5-Paper-Conference.pdf.

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

Xunkai Li, Yulin Zhao, Zhengyu Wu, Wentao Zhang, Rong-Hua Li, and Guoren Wang. Towards effective and general graph unlearning via mutual evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. ISSN 1545-2115. doi: https://doi.org/10.1146/annurev.soc.27.1.415. URL https://www.annualreviews.org/content/journals/10.1146/annurev.soc.27.1.415.

Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, Weiran Liu, and Yuan Hong. Rectifying privacy and efficacy measurements in machine unlearning: A new inference attack perspective. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5545–5564, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/naderloui.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, 2008. doi: 10.1109/SP.2008.33.

Iyiola E. Olatunji, Wolfgang Nejdl, and Megha Khosla. Membership inference attack on graph neural networks. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pp. 11–20, 2021. doi: 10.1109/TPSISA52974.2021.00002.

Iyiola E. Olatunji, Anmar Hizber, Oliver Sihlovec, and Megha Khosla. Does black-box attribute inference attacks on graph neural networks constitute privacy risk?, 2023. URL https://arxiv.org/abs/2306.00578.

Ying Song and Balaji Palanisamy. Graphtoxin: Reconstructing full unlearned graphs from graph unlearning, 2026. URL https://arxiv.org/abs/2511.10936.

Lubos Takáč and Michal Zábovský. Data analysis in public social networks. In *International Scientific Conference and International Workshop Present Day Trends of Innovations*, 2012.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

Jiancan Wu, Yi Yang, Yuchun Qian, Yongduo Sui, Xiang Wang, and Xiangnan He. Gif: A general graph unlearning strategy via influence function. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 651–661, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583521. URL https://doi.org/10.1145/3543507.3583521.

Kun Wu, Jie Shen, Yue Ning, Ting Wang, and Wendy Hui Wang. Certified edge unlearning for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '23, pp. 2606–2617, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599271. URL https://doi.org/10.1145/3580305.3599271.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Hanyang Yuan, Jiarong Xu, Cong Wang, Ziqi Yang, Chunping Wang, Keting Yin, and Yang Yang. Unveiling privacy vulnerabilities: Investigating the role of structure in graph data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 4059–4070, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672013. URL https://doi.org/10.1145/3637528.3672013.

Jiahao Zhang, Yilong Wang, Zhiwei Zhang, Xiaorui Liu, and Suhang Wang. Unlearning inversion attacks for graph neural networks. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26, pp. 934–945, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722929. doi: 10.1145/3773966.3777929. URL https://doi.org/10.1145/3773966.3777929.

Zhikun Zhang, Min Chen, Michael Backes, Yun Shen, and Yang Zhang. Inference attacks against graph neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4543–4560, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/zhang-zhikun.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 12293–12333. Curran Associates, Inc., 2024. doi: 10.52202/079017-0394. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/16e18fa3b3add076c30f2a2598f03031-Paper-Conference.pdf.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7793–7804. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/58ae23d878a47004366189884c2f8440-Paper.pdf.

A THEORY: FORMAL SETUP, PROOFS, AND EXTENDED DISCUSSION

This appendix collects the formal assumptions and key quantities (A.1), the leave-one-out algebra (A.2), the exact exposure computations with the clipping and sign-optimality discussion (A.3), the proof of Theorem 1 with the signal-budget remark (A.4) and its corollaries (A.5), the status of each direction of the tradeoff and its only failure mode (A.6), the population statements including a population frontier requiring no per-realization alignment (A.7), the CSBM proofs (A.8), the scope of the analysis (A.9), and extensions we do not prove (A.10).

A.1 KEY QUANTITIES AND FORMAL ASSUMPTIONS

Definition 2 (Leverage, residual, contribution, margin)**.**

$h_v := \phi_v^\top H^{-1}\phi_v \in [0,1)$	(leverage)
$r_v := y_v - w_{-v}^\top \phi_v$	(leave-one-out residual)
$\Delta w_v := w - w_{-v}$	(the label's contribution to the weights)
$m_v := y_v\, f(v; w_{-v}) = 1 - y_v r_v$	(retained margin)
$s_v := y_v r_v = 1 - m_v$	(signed unexplained fraction)

The released (clipped) counterparts are $\tilde{m}_v := \mathrm{clip}(m_v)$ and $\tilde{s}_v := 1 - \tilde{m}_v$; note $y_v \,\mathrm{clip}\big(f(v; w_{-v})\big) = \mathrm{clip}(m_v)$ because clip is odd, and $\tilde{m}_v = m_v$ whenever $|m_v| \le 1$. Under Assumption 5 ($m_v \ge 0$) this reduces to $\tilde{s}_v = \max\{s_v, 0\}$, the form quoted in the main text.

Notation. To keep symbols disjoint: Φ denotes exclusively the stacked representation matrix; the standard Gaussian CDF is written $F_{\mathcal{N}}$ (quantile $F_{\mathcal{N}}^{-1}$) and is never written Φ. The letter h_v is reserved for leverage; homophily is γ, never h. Lower-case d is the representation dimension; node degrees are written D_v.

Assumption 2 (Model class)**.** Representations ϕ_u are fixed (not retrained); training is ridge regression as in Section 3.1. This is the linearized/NTK regime of a message-passing GNN.

Assumption 3 (Deletion scope)**.** Deletion removes the example (ϕ_v, y_v) from the fit and leaves the retained representations $(\phi_u)_{u \ne v}$ unchanged. Exact for supervision deletion; an $\varepsilon_{\mathrm{repr}}$ perturbation term covers node deletion (Remark 10).

Assumption 4 (Observation channel)**.** Adversaries observe through the Gaussian channel of Section 3.1 with known σ (resp. σ_w for the white-box variant, which releases $w + \mathcal{N}(0, \sigma_w^2 I_d)$, unclipped); the release is a *single* query at v (resp. one noisy weight vector). Multi-query releases are discussed in Remark 11.

Assumption 5 (Formal version of Assumption 1)**.** $m_v \ge 0$: the retained fit does not point away from the true label. This is a *realization-level, directional* condition, assumed for the dataset draw under study; it is distinct from the *population, metric* alignment ε_v of Section A.7, and Remark 6 quantifies how the two relate. With the clipped release this gives $\tilde{m}_v \in [0,1]$, hence $\tilde{s}_v \in [0,1]$. (Note $m_v \ge 0$ alone does *not* bound the raw residual: an overshooting fit has $m_v > 1$ and $s_v < 0$; it is the clipping of the release that removes this case from the observation, Remark 2.)

Remark 1 (Realism of Assumption 1). The Bayes predictor is aligned in expectation, exactly: $\mathbb{E}[Y \cdot \mathbb{E}[Y \mid W]] = \mathbb{E}[\mathbb{E}[Y \mid W]^2] \ge 0$, so the assumption asserts directional consistency with the Bayes predictor. It is a device for the exact conditional identities, not load-bearing for the phenomenon: the population frontier (Theorem 2) needs no per-realization alignment. Remark 6 quantifies how often the assumption holds and separates benign label-luck violations from exploitable systematic anti-prediction (Proposition 2).

Assumption 6 (Near-interpolation; used only for the lower bound)**.** $h_v \geq 1 - \eta$ for some small $\eta \geq 0$ (e.g. $\lambda \to 0$ with ϕ_v not in a degenerate span). Then the trained model fits v's label: $y_v f(v; w) = 1 - (1 - h_v)s_v \geq 1 - \eta$ by Lemma 1(iv) and Assumption 5.

Assumption 7 (Adversary knowledge)**.** The membership attacker is unrestricted (the bound is a total variation, hence attacker-independent, and is attained by a threshold test). The reconstruction adversary knows the sign convention of Assumption 1 and is the fixed *sign-based* attacker $\hat{y} = \mathrm{sign}(z_-)$; a_{rec} is its accuracy, hence a lower bound on the best reconstruction attacker's. It is Bayes-optimal only in the symmetric single-query scenario of Remark 3. Under node deletion it additionally holds archived pre-deletion context sufficient to pose the query (Remark 10).

Assumption 8 (Orientation knowledge; structural results only)**.** In Proposition 3(ii)–(iii) and Proposition 1 the structural adversary knows the feature direction $v_\star$ and the sign of γ (e.g. estimated from a held-out labelled subset, with estimation error charged to the bounds separately; a single labelled node does not by itself determine $v_\star$ or the correlation sign). This supplements Assumption 7, which is stated before any structural model is defined and therefore cannot carry it; no non-structural result uses it.

A.2 The leave-one-out algebra

Lemma 1 (Leave-one-out algebra)**.** *Under Assumptions 2–3, with $\hat{y}_v := w^\top \phi_v$:*

(i) $r_v = \dfrac{y_v - \hat{y}_v}{1 - h_v}$*;*

(ii) $\Delta w_v = H^{-1}\phi_v\, r_v$*, hence* $\|\Delta w_v\|_H = \sqrt{h_v}\,|r_v|$ *and* $\|\Delta w_v\|_2 = \|H^{-1}\phi_v\|_2\,|r_v|$*;*

(iii) $f(v; w) - f(v; w_{-v}) = h_v r_v$*;*

(iv) $y_v f(v; w) = 1 - (1 - h_v)\, s_v$*.*

Proof. Write $\phi := \phi_v$, $h := h_v$. The reduced fit is $w_{-v} = (H - \phi\phi^\top)^{-1}(\Phi^\top y - \phi y_v)$. By Sherman–Morrison, $(H - \phi\phi^\top)^{-1} = H^{-1} + \frac{H^{-1}\phi\phi^\top H^{-1}}{1-h}$. Expanding and collecting terms,

$$w_{-v} = w - H^{-1}\phi\, y_v + \frac{H^{-1}\phi\,(\hat{y}_v - h y_v)}{1 - h} = w - H^{-1}\phi\,\frac{y_v - \hat{y}_v}{1 - h}.$$

Taking inner products with ϕ gives $w_{-v}^\top\phi = \hat{y}_v - h\frac{y_v - \hat{y}_v}{1-h}$, so $r_v = y_v - w_{-v}^\top\phi = \frac{y_v - \hat{y}_v}{1-h}$, proving (i) and, substituting back, (ii). For (ii)'s norms: $\|\Delta w_v\|_H^2 = r_v^2\,\phi^\top H^{-1} H H^{-1}\phi = h\, r_v^2$. (iii) is $\phi^\top \Delta w_v = h r_v$. For (iv): $y_v f(v; w) = y_v(w_{-v}^\top\phi + h r_v) = y_v(y_v - r_v + h r_v) = 1 - (1-h)s_v$. □

The reading of (ii) is the model's central sentence: *a training label moves the weights exactly in proportion to its unexplained part,* $\|\Delta w_v\|_H = \sqrt{h_v}\,|r_v|$ — regardless of *why* the rest of the data explains it or not.

A.3 Exact exposures

Lemma 2 (Membership advantage)**.** *Under Assumptions 2–4, with the released mean gap* $\tilde{\Delta}_v := \big|\mathrm{clip}(f(v; w)) - \mathrm{clip}(f(v; w_{-v}))\big|$*,*

$$\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\Big(\frac{\tilde{\Delta}_v}{2\sigma}\Big) - 1 \quad \textit{(black-box)}, \qquad \mathrm{Adv}^{\mathrm{wb}}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\Big(\frac{\|\Delta w_v\|_2}{2\sigma_w}\Big) - 1 \quad \textit{(white-box)},$$

both attained by a likelihood-ratio (midpoint-threshold) test. Moreover $\tilde{\Delta}_v \leq h_v|r_v|$ *always, with equality whenever neither prediction is clipped, and under Assumption 5 the exact identity* $\tilde{\Delta}_v = h_v\tilde{s}_v$ *holds. In particular* $\mathrm{Adv}_{\mathrm{MIA}} \leq \frac{h_v|r_v|}{\sigma\sqrt{2\pi}}$*.*

Proof. The two observation laws are $\mathcal{N}(\mathrm{clip}(f(v; w)), \sigma^2)$ and $\mathcal{N}(\mathrm{clip}(f(v; w_{-v})), \sigma^2)$. For equal-variance Gaussians, $\mathrm{TV}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = 2F_{\mathcal{N}}\big(\frac{|\mu_1 - \mu_2|}{2\sigma}\big) - 1$, attained by thresholding

at the midpoint. The white-box case is identical in the one-dimensional projection along Δw_v (sufficiency of the likelihood ratio). For the mean gap, Lemma 1(iii) is the convex-combination identity $f(v;w) = (1-h_v)f(v;w_{-v}) + h_v y_v$, and clip is nonexpansive, so $\tilde{\Delta}_v \le |f(v;w) - f(v;w_{-v})| = h_v|r_v|$, with equality when neither prediction is clipped. Under Assumption 5 there are two cases. If $m_v \in [0,1]$, both predictions lie on the segment between $f(v;w_{-v})$ and y_v, inside $[-1,1]$, so no clipping occurs and $\tilde{\Delta}_v = h_v s_v = h_v \tilde{s}_v$. If $m_v > 1$, then $y_v f(v;w) = 1 + (1-h_v)(m_v - 1) \ge 1$ as well, both predictions clip to y_v, and $\tilde{\Delta}_v = 0 = h_v \tilde{s}_v$. The linear bound is $2F_{\mathcal{N}}(x) - 1 \le 2x\varphi(0) = x\sqrt{2/\pi}$ with $x = h_v|r_v|/2\sigma$. □

Remark 2 (Why the release is clipped). Without clipping, Assumption 5 alone leaves an *overshoot* loophole: a retained fit with $m_v > 1$ (e.g. $f(v;w_{-v}) = 2y_v$) reconstructs with accuracy $F_{\mathcal{N}}(m_v/\sigma)$, near one at small noise, yet has $|r_v| = m_v - 1 > 0$, so adding the true label *pulls the prediction back* and the move is detectable — presence and label exposed simultaneously, the mirror image of anti-prediction. Clipping removes this from the prediction release mechanically: overshooting predictions in both worlds clip to the same value ($\tilde{\Delta}_v = 0$). It is cheap: applied identically in every world before the noise, it contracts the released mean gap (clip is nonexpansive), and the equal-variance Gaussian total variation is monotone in that gap, so the membership advantage never increases; the attacker's rule and the sign of every released mean are unchanged, and a_{rec} is unchanged whenever $|m_v| \le 1$. For an overshooting fit ($m_v > 1$) the sign attacker's accuracy does decrease, from $F_{\mathcal{N}}(m_v/\sigma)$ to $F_{\mathcal{N}}(1/\sigma)$ — a small loss only in the small-noise regime $\sigma \ll 1$ — so clipping trades reconstruction accuracy on overshoots for closing the presence leak there entirely ($\tilde{\Delta}_v = 0$). The *white-box* release is not clipped, so overshoot remains a white-box presence leak; equivalently, Theorem 1 is a statement about the prediction release, while the white-box channel requires the two-sided condition $m_v \in [0,1]$. Overshoot, like anti-prediction, is misalignment in the precise sense of Section A.7: the Bayes predictor $\mathbb{E}[Y_v \mid W]$ lies in $[-1,1]$, so $|f(v;w_{-v})| > 1$ forces $\varepsilon_v \ge \mathbb{E}[(|f(v;w_{-v})| - 1)_+^2] > 0$.

Lemma 3 (Reconstruction accuracy of the sign-based attacker)**.** *Under Assumptions 2–7,*

$$a_{\mathrm{rec}} = F_{\mathcal{N}}\Big(\frac{\tilde{m}_v}{\sigma}\Big) = F_{\mathcal{N}}\Big(\frac{1-\tilde{s}_v}{\sigma}\Big), \qquad 1 - a_{\mathrm{rec}} \le e^{-\tilde{m}_v^2/2\sigma^2} \ \ (\tilde{m}_v \ge 0).$$

Proof. $z_- = \mathrm{clip}(f(v;w_{-v})) + \mathcal{N}(0,\sigma^2)$ and $y_v \,\mathrm{clip}(f(v;w_{-v})) = \tilde{m}_v$, so $\mathbb{P}[\mathrm{sign}(z_-) = y_v] = \mathbb{P}[y_v z_- > 0] = F_{\mathcal{N}}(\tilde{m}_v/\sigma)$. The tail bound is the standard Gaussian $F_{\mathcal{N}}(-t) \le e^{-t^2/2}$. □

Remark 3 (When the sign-based attacker is optimal). Conditional on a realized dataset, "optimal reconstruction" is not a meaningful notion — the best measurable rule is the constant y_v — so optimality must be judged in the population game, where the adversary knows the law of the data but not the draw. There the sign rule is Bayes-optimal among attackers observing *only* z_- if: (a) the prior on Y_v is uniform; (b) the joint law is *flip-symmetric*: conditional on $Y_v = y$ the released score is $y\tilde{M}$ with the margin $\tilde{M} := \tilde{m}_v$ distributed independently of y — which holds for label-flip-equivariant learners (ridge: flipping all labels maps $w \mapsto -w$) on flip-symmetric data (e.g. the balanced CSBM of Section 3.3); (c) Assumption 1 holds on (almost) every draw, so $\tilde{M} \ge 0$; and (d) the channel noise density is even and nonincreasing in $|z|$ (Gaussian suffices; symmetry alone does not[1]). Proof in one line: (b) and (d) give the mirror identity $p(z \mid +) = p(-z \mid -)$, and for $z > 0$, $\tilde{M} \ge 0$ implies $\varphi\big(\frac{z-\tilde{M}}{\sigma}\big) \ge \varphi\big(\frac{z+\tilde{M}}{\sigma}\big)$ pointwise, so $p(z \mid +) \ge p(z \mid -)$ and the Bayes rule under (a) is $\mathrm{sign}(z)$. These hypotheses are *sufficient, not necessary*,[2] and a failed hypothesis removes the guarantee without establishing suboptimality. When the failures are material, the sign rule can be strictly beaten: a non-uniform prior shifts the optimal threshold away from zero; *systematically* negative margins on an adversary-known event favour the flipped rule there (Proposition 2); and non-redundant side information beyond z_- (e.g. the context W of Section A.7) admits a posterior rule, up to the Bayes ceiling $a^*(W, z_-)$. Optimality here is population-averaged, not pointwise dominance on each realized dataset. Consequently, "reconstruction fails" claims (the second regime of Corollary 2) extend to all single-query attackers under (a)–(d); outside them they are claims about the sign-based attacker only, while "reconstruction succeeds" claims, being lower bounds, hold unconditionally.

[1] A symmetric bimodal noise can reverse the likelihood ordering; unimodality is essential.

[2] Negative margins on a minority of draws need not break optimality: margins ± 0.8 with probabilities $0.9/0.1$ still leave $p(z \mid +) - p(z \mid -)$ signed like z.

A.4 Proof of Theorem 1

Proof of Theorem 1. Under Assumption 5, $\tilde{s}_v \in [0,1]$ and Lemma 2 gives the exact gap $\tilde{\Delta}_v = h_v \tilde{s}_v$, hence $\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}(h_v \tilde{s}_v/2\sigma) - 1$, nondecreasing in $\tilde{s}_v$ (strictly increasing when $h_v > 0$); Lemma 3 gives $a_{\mathrm{rec}} = F_{\mathcal{N}}((1-\tilde{s}_v)/\sigma)$, strictly decreasing. Solving for $\tilde{s}_v = 1 - \sigma F_{\mathcal{N}}^{-1}(a_{\mathrm{rec}})$ and substituting yields the frontier curve. We also record a conditional bound omitted from the main statement: if $a_{\mathrm{rec}} \geq F_{\mathcal{N}}(c/\sigma)$ for some $c \in (0,1]$, then $\tilde{m}_v \geq c$, i.e. $\tilde{s}_v \leq 1-c$, and $2F_{\mathcal{N}}(x) - 1 \leq x\sqrt{2/\pi}$ with $x = h_v \tilde{s}_v/2\sigma$ gives $\mathrm{Adv}_{\mathrm{MIA}} \leq h_v(1-c)/(\sigma\sqrt{2\pi})$. For joint attainability, with $z_\delta := F_{\mathcal{N}}^{-1}(1-\delta)$: $a_{\mathrm{rec}} \geq 1-\delta \iff \tilde{s}_v \leq 1 - \sigma z_\delta$ and $\mathrm{Adv}_{\mathrm{MIA}} \geq 1-\delta \iff \tilde{s}_v \geq 2\sigma z_{\delta/2}/h_v$; the interval $[2\sigma z_{\delta/2}/h_v,\ 1-\sigma z_\delta]$ (automatically inside $[0,1]$ when nonempty) is nonempty iff $\sigma(z_\delta + 2z_{\delta/2}/h_v) \leq 1$. In particular, whenever $\sigma > (z_\delta + 2z_{\delta/2}/h_v)^{-1}$ — and uniformly in leverage whenever $\sigma > (z_\delta + 2z_{\delta/2})^{-1}$ — no aligned residual attains both targets at level $1-\delta$. □

Remark 4 (Signal budget and the role of σ). Per node, both attacks reach level $1-\delta$ exactly when $\sigma \leq \sigma_v^*(\delta) := \min\{\frac{1-\tilde{s}_v}{z_\delta},\ \frac{h_v \tilde{s}_v}{2z_{\delta/2}}\}$ — the per-node counterpart of the existential threshold in Theorem 1, strictly positive whenever $\tilde{s}_v \in (0,1)$ and $h_v > 0$. What is noise-invariant is the *signal budget*: the released mean gap read by the membership attacker and the released margin read by the reconstruction attacker satisfy

$$\underbrace{h_v \tilde{s}_v}_{\text{MIA signal}} + \underbrace{(1-\tilde{s}_v)}_{\text{recon signal}} = 1 - (1-h_v)\,\tilde{s}_v \ \leq\ 1 :$$

the two attackers read disjoint shares of the label's unit scale — its unexplained and explained components — and σ only sets the exchange rate from signal to success probability. Two consequences. (i) Privacy claims derived from the frontier must declare the noise scale: the exclusion regime is σ comparable to the label scale, or equivalently a target membership level δ with σ calibrated Gaussian-mechanism style. (ii) The single-query restriction of Assumption 4 is load-bearing: k queries with independent channel noise average to effective noise $\sigma/\sqrt{k}$, so an unrestricted-query attacker manufactures $\sigma \to 0$ and crosses $\sigma_v^*(\delta)$ for every node with $h_v > 0$ and $\tilde{s}_v \in (0,1)$ (a zero gap or a zero margin is not created by averaging). Drawing the randomness *once at release* (one noisy weight vector, all queries functions of that draw and world-shared side information) neutralizes the averaging by data processing; it is a mechanism change whose own frontier we do not prove — see Section A.10.

A.5 Corollaries of Theorem 1

Corollary 1 (Exact conservation at interpolation)**.** *Add Assumption 6 with $\eta = 0$ ($h_v \to 1$). Then*

$$\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\Big(\frac{\tilde{s}_v}{2\sigma}\Big) - 1, \qquad a_{\mathrm{rec}} = F_{\mathcal{N}}\Big(\frac{1-\tilde{s}_v}{\sigma}\Big),$$

and as $\tilde{s}_v$ runs over $[0,1]$ the pair traverses the full frontier from $\big(0,\ F_{\mathcal{N}}(1/\sigma)\big)$ (undetectable presence, near-certain label recovery) to $\big(2F_{\mathcal{N}}(1/2\sigma) - 1,\ \frac{1}{2}\big)$ (near-certain presence detection, chance-level label recovery). Moreover the trained model fits v exactly ($y_v f(v;w) = 1$) at every point of the frontier, so the sweep is not explained by differing fit quality.

Corollary 2 (The two regimes)**.** *Under Assumptions 2–6:*

- ***Explained label*** *($\tilde{s}_v \leq \tau$):* $a_{\mathrm{rec}} \geq F_{\mathcal{N}}\big(\frac{1-\tau}{\sigma}\big)$ *while* $\mathrm{Adv}_{\mathrm{MIA}} \leq \frac{\tau}{\sigma\sqrt{2\pi}}$ *— membership auditing is blind exactly where the label leaks.*
- ***Unexplained label*** *($\tilde{s}_v \geq 1-\tau$):* $a_{\mathrm{rec}} \leq F_{\mathcal{N}}\big(\frac{\tau}{\sigma}\big) \to \frac{1}{2}$ *while* $\mathrm{Adv}_{\mathrm{MIA}} \geq 2F_{\mathcal{N}}\big(\frac{(1-\eta)(1-\tau)}{2\sigma}\big) - 1$ *— the label is stored in the weights (Lemma 1(ii) with $|r_v| = \tilde{s}_v \geq 1-\tau$, valid since $\tilde{s}_v > 0$ forces $m_v < 1$: $\|\Delta w_v\|_H \geq \sqrt{1-\eta}\,(1-\tau)$) and its presence is detectable, but the retained data reveals nothing.*

A.6 Is it really a tradeoff? Status of each direction

1. *Reconstructible ⇒ presence-private.* Proved: conditionally on the draw (Theorem 1, under Assumptions 2–5) and in population form without any per-realization alignment (Theorem 2(i)–(ii)).
2. *Unreconstructible ⇒ presence-exposed.* Proved under interpolation (Corollary 2; Theorem 2(iii)); genuinely conditional — regularization defeats it, by design.
3. *Both-low is possible.* Small leverage (λ large) silences the membership signal without aiding reconstruction; the feasible set is the region below the frontier, and presence-privacy is bought by not learning the node — a utility cost, not a free lunch.
4. *Both-high.* Excluded for aligned residuals when $\sigma > (z_\delta + 2z_{\delta/2}/h_v)^{-1}$ (Theorem 1); achievable below that threshold at any $\tilde{s}_v \in (0,1)$ (Remark 4), and under systematic anti-prediction, which escapes the aligned signal budget at every noise level, though near-one exposures still require h_v bounded away from zero and small σ (Proposition 2).

Proposition 2 (The only failure mode)**.** *Drop Assumption 5 and suppose $m_v \in [-1, 0)$ (the retained fit anti-predicts y_v). On this range neither prediction is clipped: $|f(v; w_{-v})| = |m_v| \le 1$, and $y_v f(v; w) = m_v + h_v(1 - m_v) \in [m_v, 1] \subset [-1, 1]$. A sign-aware adversary using $\hat{y} = -\operatorname{sign}(z_-)$ achieves $a_{\mathrm{rec}} = F_{\mathcal{N}}(|m_v|/\sigma)$, while $|r_v| = 1 + |m_v| > 1$ gives $\mathrm{Adv}_{\mathrm{MIA}} = 2F_{\mathcal{N}}\big(\frac{h_v(1+|m_v|)}{2\sigma}\big) - 1$. Both exposures can then be simultaneously near one when h_v is bounded away from zero and σ is small; at vanishing leverage the membership signal $h_v(1 + |m_v|)$ vanishes with it. (Deeper excursions, $m_v < -1$, are attenuated by clipping and, at small leverage, hidden from the black-box channel entirely.)*

Proof. Immediate from Lemmas 2 and 3 with the flipped decision rule. □

Remark 5. Proposition 2 is not an artefact: it says that a learner which fits exploitable structure *with the wrong sign* both memorizes the label (large residual) and leaves it recoverable (the adversary knows to flip). The tradeoff is therefore a statement about aligned learners, and Assumption 1 marks its exact boundary. The flip attack presumes the sign error is *systematic and known to the adversary*; Remark 6(iii) shows such systematic error is certifiably expensive in the population alignment ε_v, whereas per-realization violations caused by an atypical label draw are not exploitable at all.

A.7 Population statements: the residual is pinned to unpredictability

Treat graph, features and labels as random, keep the target v fixed, and write Y_v for its random label. Let $\hat{Y}_v^-$ be the post-deletion score at v of *any* learner (on the label scale), measurable with respect to W, the information retained after deletion plus any adversary-archived context. Define the *population residual* $R_v := Y_v - \hat{Y}_v^-$ and the *excess error* $\varepsilon_v := \mathbb{E}[(\hat{Y}_v^- - \mathbb{E}[Y_v \mid W])^2]$ over the Bayes predictor $\mathbb{E}[Y_v \mid W]$; a learner is $\varepsilon_{\mathrm{al}}$-*aligned* if $\varepsilon_v \le \varepsilon_{\mathrm{al}}$. Alignment is mean-square closeness to the Bayes predictor — a population, metric notion, distinct from the realization-level, directional Assumption 1; Remark 6 bridges the two. Throughout this and the following subsections, write $\bar{Y}_v := \mathbb{E}[Y_v \mid W]$ and the *predictability* $P_v := 1 - \mathbb{E}\operatorname{Var}(Y_v \mid W) = \mathbb{E}[\bar{Y}_v^2] \in [0, 1]$: a node is *predictable* when P_v is near one.

Proposition 3 (Residual is pinned to unpredictability)**.** *(i)* Any learner, any data law: $\mathbb{E}[R_v^2] = (1 - P_v) + \varepsilon_v$, *and for every $t > 0$, $\mathbb{P}[|R_v| \ge t] \le \mathbb{E}[R_v^2]/t^2$ while $\mathbb{P}[|R_v| \ge t] \ge \frac{1}{4}\big((1 - P_v) - t^2\big)$.* *(ii)* CSBM, structure explains the label: *conditionally on $D_v = D \ge 1$ in the model of Definition 1, if W contains $(D_v, \phi_v^{\mathrm{struct}})$ and the learner is conditionally $\varepsilon_{\mathrm{al}}$-aligned, then $\mathbb{E}[R_v^2 \mid D] \le 4e^{-D\gamma^2/8} + 4e^{-\mu^2\gamma^2 D/8\nu^2} + \varepsilon_{\mathrm{al}}$.* *(iii)* CSBM, converse: *if $W = (D_v, \phi_v^{\mathrm{struct}})$ (structure-only observation: a self-masked aggregator, or full node deletion), then for any learner $\mathbb{E}[R_v^2 \mid D] \ge 2\,\mathbb{E}\,F_{\mathcal{N}}\big(-\frac{\mu\sqrt{D}}{\nu}\big|\frac{2S}{D} - 1\big|\big)$, where $S \sim \mathrm{Bin}(D, \frac{1+\gamma}{2})$ is the same-community neighbor count — a lower bound matching (ii) in the feature term, not in the composition term; separately, at $\gamma = 0$ the channel is independent of Y_v and $\mathbb{E}[R_v^2 \mid D] \ge 1$: the label is entirely unexplained.*

Proof of Proposition 3(i). Since $\hat{Y}_v^-$ and $\bar{Y}_v$ are W-measurable, $Y_v - \bar{Y}_v$ is orthogonal to $\bar{Y}_v - \hat{Y}_v^-$, so $\mathbb{E}[R_v^2] = \mathbb{E}[(Y_v - \bar{Y}_v)^2] + \mathbb{E}[(\bar{Y}_v - \hat{Y}_v^-)^2] = \mathbb{E}\operatorname{Var}(Y_v \mid W) + \varepsilon_v$, the stated identity. The first

tail bound is Markov's inequality for R_v^2. For the second, define the clipped residual $\tilde{R}_v := Y_v - \mathrm{clip}(\hat{Y}_v^-)$. Since $\mathrm{clip}(\hat{Y}_v^-)$ is W-measurable, the identity applies to $\tilde{R}_v$ as well (with its own excess error), giving $\mathbb{E}[\tilde{R}_v^2] \geq 1 - P_v$. Since $Y_v \in [-1, 1]$ and projection onto $[-1, 1]$ is nonexpansive, $|\tilde{R}_v| \leq |R_v|$; also $|\tilde{R}_v| \leq 2$ always, and $|\tilde{R}_v| \leq 1$ when Assumption 5 holds on the realization (which gives $Y_v \mathrm{clip}(\hat{Y}_v^-) = \mathrm{clip}(m_v) \in [0, 1]$). Hence $\tilde{R}_v^2 \leq t^2 + 4\,\mathbf{1}\{|\tilde{R}_v| \geq t\}$, with 4 replaced by 1 under the assumption; taking expectations gives $\mathbb{P}[|R_v| \geq t] \geq \mathbb{P}[|\tilde{R}_v| \geq t] \geq \frac{1}{4}((1-P_v) - t^2)$, with constant 1 if the assumption holds on every realization. □

Clipping is compatible with every part of Proposition 3: $\mathrm{clip}(\hat{Y}_v^-)$ is W-measurable, so the part-(i) identity holds for $\tilde{R}_v$ too, and because the Bayes predictor $\mathbb{E}[Y_v \mid W]$ itself lies in $[-1, 1]$, clipping can only move the prediction toward it — the clipped excess error satisfies $\tilde{\varepsilon}_v := \mathbb{E}[(\mathrm{clip}(\hat{Y}_v^-) - \mathbb{E}[Y_v \mid W])^2] \leq \varepsilon_v$, so the release-side fix never worsens alignment. The weight-space quantities (Δw_v, Lemma 1) are untouched: clipping lives on the observation channel, not in training.

Remark 6 (Bridging realization-level and population alignment). Write $\bar{Y}_v := \mathbb{E}[Y_v \mid W]$ and decompose the retained margin as $m_v = Y_v \bar{Y}_v + Y_v(\hat{Y}_v^- - \bar{Y}_v)$: a *label-luck* term and a *misalignment* term. Three consequences. (i) *Expectation link.* $\mathbb{E}[Y_v \bar{Y}_v] = \mathbb{E}[\bar{Y}_v^2] = P_v$, and by the tower property $\mathbb{E}[Y_v(\hat{Y}_v^- - \bar{Y}_v)] = \mathbb{E}[\bar{Y}_v(\hat{Y}_v^- - \bar{Y}_v)]$, so Cauchy–Schwarz gives

$$\mathbb{E}[m_v] \;\geq\; P_v - \sqrt{P_v\,\varepsilon_v}.$$

(ii) *Probability link.* For every $\tau > 0$,

$$\mathbb{P}[m_v < 0] \;\leq\; \tfrac{1}{2}\,(1 - P_v) \;+\; \mathbb{P}\big[|\bar{Y}_v| < \tau\big] \;+\; \varepsilon_v/\tau^2,$$

since $\{m_v < 0\} \subseteq \{Y_v \bar{Y}_v < 0\} \cup \{|\bar{Y}_v| < \tau\} \cup \{|\hat{Y}_v^- - \bar{Y}_v| \geq \tau\}$, with $\mathbb{P}[Y_v \bar{Y}_v < 0] = \mathbb{E}\big[\frac{1-|\bar{Y}_v|}{2}\,\mathbf{1}\{|\bar{Y}_v| > 0\}\big] \leq (1 - P_v)/2$ (because $|\bar{Y}_v| \geq \bar{Y}_v^2$; the indicator matters at $\bar{Y}_v = 0$, where the event has probability zero) and Chebyshev on the last event. A decomposition-free alternative: $m_v < 0$ forces $|R_v| > 1$, so Markov applied to $\mathbb{E}[R_v^2] = (1 - P_v) + \varepsilon_v$ gives $\mathbb{P}[m_v < 0] \leq \min\{1,\ (1-P_v)+\varepsilon_v\}$. These are *sufficient* conditions for high-probability alignment on predictable nodes fit by a population-aligned learner — the regime in which Theorem 1 invokes it — not a characterization. (iii) *Taxonomy of violations.* The two failure causes differ in kind. *Label luck* ($\hat{Y}_v^- \approx \bar{Y}_v$, drawn label atypical) is not exploitable: flipping the sign rule is wrong on the majority of draws, so the attack of Proposition 2 is unavailable; on such draws reconstruction fails and the surprising label is memorized — the frontier operating at realization level, not a counterexample. *Systematic sign error* (the learner anti-predicts $\bar{Y}_v$ on an event A, so $|\hat{Y}_v^- - \bar{Y}_v| \geq |\bar{Y}_v|$ on A) is exploitable and certifiably expensive: $\varepsilon_v \geq \mathbb{E}[\bar{Y}_v^2 \mathbf{1}_A]$. Assumption 1 is therefore best read as excluding *systematic* anti-prediction; benign per-draw violations are quantified by (ii) and absorbed by the population statements of this section. A draw with $m_v < 0$ nonetheless falls outside the hypotheses of Theorem 1 even when it is an ordinary Bayes error; population applications must budget for such draws via (ii) rather than reinterpret the assumption.

The pieces above assemble into a population version of the tradeoff that dispenses with per-realization alignment altogether.

Theorem 2 (Population frontier; no per-realization alignment)**.** *Define the population exposures by averaging over data and channel: $a_{\mathrm{rec}}^{\mathrm{pop}} := \mathbb{P}[\mathrm{sign}(z_-) = Y_v]$, and $\mathrm{Adv}_{\mathrm{MIA}}^{\mathrm{pop}} :=$ the total variation between the two membership-world joint laws of the release together with the world-shared data — which includes Y_v and ϕ_v: the population analogue of the label-informed attacker of Assumption 7. The label side information is essential, not cosmetic: without it the advantage can be strictly smaller — an anti-correlated design can make the trained and retained release laws coincide as mixtures over Y_v (release-only total variation zero) while the label-informed lower bound of (iii) below stays bounded away from zero. Write $\beta_v := \mathbb{E}[R_v^2] = (1 - P_v) + \varepsilon_v$ and let $h_{\max}$ bound the leverage (one may take $h_{\max} = 1$). Under Assumptions 2–4 and the setting of this subsection:*

(i) (Reconstructible ⇒ presence-private.) $\mathrm{Adv}_{\mathrm{MIA}}^{\mathrm{pop}} \;\leq\; \dfrac{\mathbb{E}[h_v|R_v|]}{\sigma\sqrt{2\pi}} \;\leq\; \dfrac{h_{\max}\sqrt{\beta_v}}{\sigma\sqrt{2\pi}}.$

(ii) (Predictable ⇒ reconstructible.) *For every $c \in (0, 1)$,* $a_{\mathrm{rec}}^{\mathrm{pop}} \;\geq\; F_{\mathcal{N}}(c/\sigma)\Big(1 - \dfrac{\beta_v}{(1-c)^2}\Big).$

*(iii) (*Unpredictable ⇒ reconstruction at chance, presence exposed.*)* $a_{\mathrm{rec}}^{\mathrm{pop}} \le \frac{1}{2} + \dfrac{\sqrt{P_v} + \sqrt{\varepsilon_v}}{\sigma\sqrt{2\pi}}$*;*

and under Assumption 6 with parameter $\eta \le 1$*,* $\mathrm{Adv}_{\mathrm{MIA}}^{\mathrm{pop}} \ge F_{\mathcal{N}}\Big(\dfrac{1-2\eta}{\sigma}\Big)\Big(1 - \dfrac{\beta_v}{4}\Big) - a_{\mathrm{rec}}^{\mathrm{pop}}$.

Assumption 1 is not used: anti-prediction and overshoot are absorbed by β_v.

Proof. (i) The total variation between mixtures is at most the mean conditional total variation, so by Lemma 2 and the linear bound, $\mathrm{Adv}_{\mathrm{MIA}}^{\mathrm{pop}} \le \mathbb{E}[2F_{\mathcal{N}}(\tilde{\Delta}_v/2\sigma) - 1] \le \mathbb{E}[\tilde{\Delta}_v]/(\sigma\sqrt{2\pi})$; $\tilde{\Delta}_v \le h_v|R_v|$ always, and Cauchy–Schwarz gives $\mathbb{E}|R_v| \le \sqrt{\beta_v}$. (ii) $a_{\mathrm{rec}}^{\mathrm{pop}} = \mathbb{E}[F_{\mathcal{N}}(\tilde{m}_v/\sigma)] \ge F_{\mathcal{N}}(c/\sigma)\,\mathbb{P}[m_v \ge c]$, and $m_v < c \le 1$ forces $|R_v| = |1 - Y_v\hat{Y}_v^-| > 1 - c$, so Markov gives $\mathbb{P}[m_v < c] \le \beta_v/(1-c)^2$. (iii) $F_{\mathcal{N}}(x) \le \frac{1}{2} + x_+/\sqrt{2\pi}$ gives $a_{\mathrm{rec}}^{\mathrm{pop}} \le \frac{1}{2} + \mathbb{E}[\tilde{m}_{v,+}]/(\sigma\sqrt{2\pi})$, and $\mathbb{E}[\tilde{m}_{v,+}] \le \mathbb{E}[(Y_v\bar{Y}_v)_+] + \mathbb{E}|\hat{Y}_v^- - \bar{Y}_v| \le \sqrt{P_v} + \sqrt{\varepsilon_v}$ by Cauchy–Schwarz. For the advantage, the correctness attacker — guess $b = +$ iff $\mathrm{sign}(z) = Y_v$, a function of the release and the world-shared Y_v, hence admissible — gives $\mathrm{Adv}_{\mathrm{MIA}}^{\mathrm{pop}} \ge \alpha_+ - \alpha_-$ with $\alpha_- = a_{\mathrm{rec}}^{\mathrm{pop}}$. By Lemma 1(iv), $Y_v f(v; w) = 1 - (1-h_v)Y_v R_v \ge 1 - \eta|R_v|$, so on $\{|R_v| \le 2\}$ the released trained margin $\mathrm{clip}(Y_v f(v; w))$ is at least $1 - 2\eta \ge -1$; splitting on that event with $\mathbb{P}[|R_v| > 2] \le \beta_v/4$ (Markov) yields $\alpha_+ \ge F_{\mathcal{N}}((1-2\eta)/\sigma)(1 - \beta_v/4)$. □

Homophily, structured heterophily, degree, and feature informativeness are interchangeable *mechanisms* that raise P_v; none appears in the statements above, which is what permits the tradeoff to be read as a statement about the two privacy notions themselves.

A.8 CSBM: PROOFS AND DISCUSSION

Remark 7 (Balanced in expectation, not by constraint). In Definition 1 labels are i.i.d. uniform, so the two communities are balanced in expectation only. Exact balance is a different model: the conditional neighbor composition below becomes noncentral hypergeometric rather than binomial, and the label-sum constraint is itself informative even at $\gamma = 0$. The two models agree asymptotically only in sparse regimes such as $p = a/N$, $q = b/N$; at fixed dense p, q the conditional laws differ even as $N \to \infty$. Conditionally on y_v and $D_v = D$, each neighbor is same-community independently with probability $p/(p+q) = \frac{1+\gamma}{2}$ (this uses the i.i.d. label law; edges are conditionally independent given labels), so the same-community count is $S \sim \mathrm{Bin}\big(D, \frac{1+\gamma}{2}\big)$ and

$$\big\langle \phi_v^{\mathrm{struct}}, v_\star \big\rangle \;\big|\; (y_v, S) \;\sim\; \mathcal{N}\Big(\mu\big(\tfrac{2S}{D} - 1\big)\, y_v,\; \tfrac{\nu^2}{D}\Big), \qquad \mathbb{E}\big[\tfrac{2S}{D} - 1\big] = \gamma.$$

The natural structural signal-to-noise ratio, accounting for both feature noise and composition noise, is $\kappa_v := \mu\,|\gamma|\,\sqrt{D_v}/\sqrt{\nu^2 + \mu^2(1-\gamma^2)}$ — the exact form of the order-$\mu|\gamma|\sqrt{D_v}/\nu$ rate quoted in the main text.

Proof of Proposition 3(ii)–(iii). Both parts use the pointwise sandwich $2\,e(W) \le \mathrm{Var}(Y_v \mid W) \le 4\,e(W)$, where $e(W) := \frac{1}{2}\big(1 - |\mathbb{E}[Y_v \mid W]|\big)$ is the conditional Bayes error: indeed $\mathrm{Var}(Y_v \mid W) = 1 - \mathbb{E}[Y_v \mid W]^2 = (1 - |\mathbb{E}[Y_v \mid W]|)(1 + |\mathbb{E}[Y_v \mid W]|)$ and the second factor lies in $[1, 2]$.

(ii) The Bayes error is dominated by the error of any test; take $\hat{y} := \mathrm{sign}(\gamma) \cdot \mathrm{sign}\langle \phi_v^{\mathrm{struct}}, v_\star \rangle$ (admissible by Assumption 8) and let $\gamma > 0$ (for $\gamma < 0$ replace S by $D - S$). By Hoeffding, $\mathbb{P}\big[\frac{2S}{D} - 1 \le \frac{\gamma}{2}\big] \le e^{-D\gamma^2/8}$; on the complementary event the conditional error is at most $F_{\mathcal{N}}\big(-\mu(\gamma/2)\sqrt{D}/\nu\big)$. Hence

$$1 - P_v(D) = \mathbb{E}[\mathrm{Var}(Y_v \mid W) \mid D] \le 4\,\mathbb{E}[e(W) \mid D] \le 4e^{-D\gamma^2/8} + 4F_{\mathcal{N}}\Big(-\frac{\mu|\gamma|\sqrt{D}}{2\nu}\Big),$$

a form sharper than the stated exponential bound, which follows via $F_{\mathcal{N}}(-t) \le e^{-t^2/2}$. Transfer to $\mathbb{E}[R_v^2 \mid D]$ is the part-(i) identity applied conditionally on $\{D_v = D\}$ (a W-measurable event, so the conditional predictability $P_v(D)$ is well defined); a globally aligned learner recovers the unconditional statement after averaging over D_v.

(iii) Couple S across the two hypotheses $y_v = \pm 1$. Given S, the conditional laws of ϕ_v^{struct} are Gaussians with means $\pm\mu(\frac{2S}{D}-1)v_\star$ and equal covariance, at total variation $2F_{\mathcal{N}}\big(\frac{\mu\sqrt{D}}{\nu}|\frac{2S}{D}-1|\big)-1$. Hence the unconditional laws satisfy $\mathrm{TV} \le \mathbb{E}_S\big[2F_{\mathcal{N}}(\cdot)-1\big]$, the Bayes error is at least $\frac{1}{2}(1-\mathrm{TV}) \ge \mathbb{E}_S\, F_{\mathcal{N}}\big(-\frac{\mu\sqrt{D}}{\nu}|\frac{2S}{D}-1|\big)$, and the sandwich lower-bounds $1-P_v(D)$ by twice this quantity; the part-(i) identity, applied conditionally as in (ii), gives $\mathbb{E}[R_v^2 \mid D] \ge 1 - P_v(D)$ and hence the claim (the degree is included in W so that conditioning on $\{D_v = D\}$ remains W-measurable). At $\gamma = 0$ the map $S \mapsto D - S$ preserves $\mathrm{Bin}(D, \frac{1}{2})$ (this is where the i.i.d. label law is used; under exact balance the label-sum constraint breaks the symmetry) and flips the mean, so the two mixture laws coincide, $\mathbb{E}[Y_v \mid W] = 0$, $P_v(D) = 0$, and $\mathbb{E}[R_v^2 \mid D] \ge 1$. □

Remark 8 (Own features). Including x_v in the representation adds an independent private channel of strength μ/ν on top of κ_v; it changes P_v but not the logic. Note the scope this imposes on Proposition 3(iii): supervision deletion retains x_v, and a message-passing ϕ_v typically includes it, so the structure-only hypothesis models a self-masked aggregator or full node deletion. With self-features included, "low $|\gamma|\sqrt{D_v}$ forces memorization" additionally requires the own-feature channel to be weak (μ/ν small); otherwise $x_v \in W$ explains the label and the residual is small regardless of structure. The masked channel is isolated deliberately: it is exactly what survives node deletion (Section 3.3).

Proof of Proposition 1. Immediate from the test-error bound in the proof of Proposition 3(ii): the test is a function of retained data and the archived edge list only. The $\frac{1}{2}$ term is a fair guess; the displayed bound is vacuous at weak signal. □

Remark 9 (Joint adversaries and conditioning bases). Under node deletion the post-deletion reconstruction exposure is at least $\max\big\{F_{\mathcal{N}}(\tilde{m}_v/\sigma),\ a_{\mathrm{rec}}^{\mathrm{struct}}\big\}$ — each attack is separately available. This max is a lower bound, not the total exposure: a joint adversary combining the two observations can do strictly better when they carry complementary evidence (two conditionally independent Gaussian channels with signals α_1, α_2 give joint accuracy $F_{\mathcal{N}}(\sqrt{\alpha_1^2+\alpha_2^2})$), and adds nothing only when the query is redundant given the structural information ($Y \to W \to z_-$). The two terms are also computed on different conditioning bases (fixed-dataset channel versus degree-conditioned population), so the max should be read as a summary of available attacks, not as a single well-defined optimal accuracy. The frontier of Theorem 1 governs only the model-mediated term.

A.9 Scope of the analysis

Remark 10 (Node deletion). Full node deletion also removes v's incident edges, perturbing retained representations: $\phi_u \mapsto \phi_u + \delta_u$ for u near v. The comparison is then no longer a rank-one update, and the identities of Lemma 1 do *not* carry over with a generic additive $O(\max_u \|\delta_u\|)$ term. What is available, for fixed $\lambda > 0$, is a design-dependent perturbation budget: writing E for the stacked row perturbations ($\|E\| \le \sqrt{k}\,\max_u \|\delta_u\|$ when k rows change) and $B = \Phi_{-v}$, the refit satisfies $\|w' - w_{-v}\| \le \frac{1}{\lambda}\big[\|E\|\|y_{-v}\| + (2\|B\|\|E\| + \|E\|^2)\|w_{-v}\|\big]$, an error of at most e in one released mean shifts the corresponding advantage or sign accuracy by at most $e/(\sigma\sqrt{2\pi})$, and querying with a perturbed representation adds a further $\|\phi_v\|$-scaled term. The resulting $\varepsilon_{\mathrm{repr}}$ is therefore a placeholder for this explicit, constant-laden budget — which need not remain uniform as $\lambda \to 0$ (interpolation) or $\sigma \to 0$ — not a universal small term. The reconstruction query then requires the adversary to hold archived context (edges, and features if used) to pose the ghost query; Section 3.3 develops the structure-only leakage picture. Note also that a post-deletion history test (unlearned versus never-trained) is a *different experiment* from this section's membership game (pre-deletion model versus refit), and conclusions do not transfer between them without restating the game.

Remark 11 (Single-query release). Assumption 4 restricts the release to one query at v. A multi-query or full-weight release can leak presence through *other* nodes' predictions (under node deletion their representations change too), so the MIA expressions become lower bounds and the frontier's upper (impossibility) branch requires an additional on-average stability bound on the released transcript. Repeated queries at v itself average the channel noise down as $\sigma/\sqrt{k}$, improving *both* attackers: $\mathrm{Adv}_{\mathrm{MIA},k} = 2F_{\mathcal{N}}(\sqrt{k}\,\tilde{\Delta}_v/2\sigma) - 1$ and $a_{\mathrm{rec},k} = F_{\mathcal{N}}(\sqrt{k}\,\tilde{m}_v/\sigma)$, each tending to one when its signal is positive — so neither axis is invariant to the release interface, and the single-query expressions are baselines, not optimal enlarged-transcript accuracies. Remark 4 quantifies the

consequence and the release-time randomization (one noise draw, queries as functions of it) that neutralizes the averaging; what survives any interface is the signal-budget identity of Remark 4 and the weight-space quantities of Lemma 1. The multi-query frontier itself is deferred to Section A.10.

Remark 12 (Where the randomness lives). Ridge is deterministic, so the Gaussian channel supplies all in-game randomness. Substituting training stochasticity (SGD noise, random initialization of a nonlinear model) for the channel preserves the structure — the mean separation is still $h_v r_v$ in the linearized regime — but the exact Gaussian forms become approximations; we prefer to keep the noise explicit and named.

Remark 13 (Nonlinear and multi-layer GNNs). Two separate extensions. *Depth.* An ℓ-layer network reads the ℓ-hop ball around v. In the CSBM of Section 3.3 the label correlation along a length-k path decays like γ^k while the number of k-hop neighbors grows like D^k, so the k-hop aggregate carries a structural signal-to-noise of order $(\mu/\nu)\,(|\gamma|\sqrt{D})^k$: depth amplifies the structural channel when $\gamma^2 D > 1$ and attenuates it otherwise — the residual-level face of the familiar depth/oversmoothing dial. Oversmoothing itself homogenizes representations, flattening both the leverage profile h_v and the residual spectrum, so nodes' positions on the frontier of Theorem 1 cluster. *Nonlinearity.* The rank-one leave-one-out identity (Lemma 1) is exact only for ridge. In the linearized/NTK regime the results apply verbatim with ϕ_u the NTK features (Assumption 2). Beyond it, Δw_v is replaced by its influence-function first-order approximation — still "residual scaled by (kernel) leverage" — and the frontier holds to first order in the residual, with curvature and feature-learning corrections. Feature learning cuts both ways: it can *restore* alignment under heterophily by learning the sign flip, shrinking the failure mode of Proposition 2; but it also makes deletion inexact — practical unlearning leaves a gap δ_{unl} between the returned model and the true refit, which adds to the membership advantage (extra mean separation) and can add to reconstruction (residual traces of y_v), so both exposures acquire an additive $O(\delta_{\mathrm{unl}})$ slack.

A.10 Extensions not proved here

Each item below is a plausible extension whose missing ingredient we can name; none is used by the results above.

Multi-query transcripts. For releases answering queries at several nodes, the membership expressions of Lemma 2 become lower bounds, and the exclusion branch of Theorem 1 requires an on-average stability bound on the released transcript — under node deletion the perturbed neighbor representations enter here. Missing: the stability bound.

One-draw weight release. Releasing $w + \mathcal{N}(0, \sigma_w^2 I_d)$ once caps all query strategies at $2F_{\mathcal{N}}(\|\Delta w_v\|_2/2\sigma_w) - 1$ by data processing, neutralizing noise averaging (Remark 4). The query noise induced at the target is correlated across queries, with variance $\sigma_w^2\|\phi_v\|^2$, so the clipped scalar frontier is not automatically a theorem for this release; a differential-privacy reading further requires an adjacency relation and a sensitivity bound. Missing: the frontier for the correlated-noise channel.

A structural presence channel under node deletion. Removing v's edges shifts each neighbor's mean aggregate by exactly $(\phi_u - x_v)/(D_u - 1)$ for a mean aggregator with $D_u > 1$ — of order $1/D_u$ only under bounded features. A multi-query attacker probing all D_v neighbors under independent equal-variance noise sees a combined signal of Euclidean order $\sqrt{D_v}/D_u$, which may grow, stagnate, or vanish with degree (identical neighbor features produce zero shift). The label-deletion and representation-change effects add algebraically but can reinforce or cancel. Missing: distributional control of the shifts and of their interaction with the label channel.

The frontier under full node deletion. That the residual–leverage tradeoff survives node deletion is supported by the perturbation budget of Remark 10, and the structural floor of Proposition 1 matches the frontier's endpoints, but the correspondence at the frontier's interior is not verified formally. Missing: a node-deletion analogue of Lemma 1.

B EXPERIMENTAL DETAILS

B.1 DATA AND TERMS OF USE

Table 3 summarizes both graphs and the audited cohort. Both labels are near-balanced and are predicted from profile features that exclude the label and every sensitive column. The graphs differ in how a label can be read from the neighborhood: a Pokec-n user has a median of one friend with a known label, which agrees with the user's own 64% of the time (a sparse, homophilous label signal); a Cornell5 user has 61, but gender agreement among friends is near chance (0.56), so the neighborhood speaks mainly through friends' features.

Table 3: Graphs and cohort. *Candidates* are the users who request deletion and the only users measured; *anchors* are same-pool users held out of every model and take no part in the results. "Friends with a known label" counts a candidate's retained neighbors whose label is in the training set; "agreement" is the share of those friends with the candidate's label (chance 0.50).

	Pokec-n	Cornell5
Users / friendships	66,569 / 517,047	18,660 / 790,777
Label the model predicts	field of work	gender
Supervised training users	5,278	10,093
Majority label in the cohort	52.7%	50.3%
Label agreement among friends	0.64	0.56
Candidates / anchors	1,000 / 1,000	1,000 / 1,000
Median candidate degree	10	66
Median friends with a known label	1	61

Pokec-n. Pokec is a Slovak online social network (Takáč & Zábovský, 2012); Pokec-n is the regional subgraph introduced by Dai & Wang (2021) and distributed through the PyGDebias collection, with $66{,}569$ users and $517{,}047$ friendships. Profile fields are the model's input features after removing the label column and every sensitive column (age, gender, region); the label is the binary field "I am working in field", known for $8{,}797$ users, of whom 60% are assigned to training by a stratified split ($5{,}278$ users), the rest to validation and test. Missing profile values are imputed with the retained-graph mean and every feature is standardized on the retained graph, so the served preprocessing is identical for every release.

Cornell5. Cornell5 is one of the Facebook100 networks collected in 2005 from public profiles (Traud et al., 2012), taken from the redistribution of Lim et al. (2021) ($18{,}660$ users, $790{,}777$ friendships). The input features are one-hot encodings of student/faculty status, major, second major and high school (codes carried by fewer than 20 users merged; 235 columns); the label is gender ($10{,}093$ training users). Class year and dormitory are withheld from the features as sensitive attributes.

Terms. Neither dataset carries an explicit data license: Pokec-n is used under the terms of the research code that distributes it and Facebook100 is redistributed by Lim et al. (2021) without one. We infer no permission beyond research use, run every experiment locally, and release aggregate results only; no user identifier appears in any artifact.

B.2 WORLDS, MODEL AND RELEASES

Worlds. From the training users with a known label we draw $2{,}000$ users uniformly at random and split them into $1{,}000$ candidates and $1{,}000$ anchors. The *retained graph* is the full graph with all $2{,}000$ removed. A world adds a random half of the candidates back — with their features, labels and friendships — and trains on the result; the two worlds of a pair use complementary halves, so each candidate is a training member in one world of every pair. There are 32 pairs per dataset: 16 shadow, 8 calibration and 8 evaluation pairs. Every released model, of every world, is served on the retained graph, and a query about a deleted candidate reaches the model only through the neighborhood the querier supplies.

Model. A two-layer GCN (Kipf & Welling, 2017) with 64 hidden units, dropout 0.5, Adam with weight decay 5×10^{-4} and the validation-loss checkpoint; on Cornell5 a plateau learning-rate schedule (lr 5×10^{-3}, 10 warm-up epochs, at most 600 epochs). Test accuracy on the retained graph is 0.69–0.71 on Pokec-n and 0.75–0.76 on Cornell5 for every release except CGU on Cornell5 (0.71; Table 5).

Releases. Each world serves seven models. **No-op** is the trained model, the theorem's w. **GOLD** retrains from scratch on the world's data without the deleted candidates: exact deletion, a draw from the law of w_{-v}, and the reference every unlearner is meant to match. **Never** is an independent second retrain without the deleted candidates; since GOLD and Never differ only in training randomness, the GOLD–Never gap is the noise floor of every number in the paper. The four unlearners below are applied to No-op and served in its place.

Releases and tuning. GIF (Wu et al., 2023a) applies an influence-function correction with a damped Hessian-vector recurrence; MEGU (Li et al., 2024) co-trains a forgetting and a retention objective on the deleted users' neighborhood; GNNDelete (Cheng et al., 2023) freezes the trained weights and learns per-layer deletion operators on the deleted users and their neighbors; CGU (Chien et al., 2023) is a certified method that replaces the GCN by a convex linear model on two-step propagated features and removes users by sequential Newton updates with a certification budget — a different model class from the other releases, which is why its utility differs. All four are our OpenGU adaptations (Fan et al., 2025) to node deletion on the retained-graph contract. Published defaults were not usable as-is on Pokec-n (GIF's update was inert, MEGU's nearly so, GNNDelete's operators over-corrected, CGU's optimizer left its model at the majority class), so each method was tuned per dataset on eight development worlds disjoint from the audit, against the paired membership advantage of a raw-score attacker (target 0.5) under a utility constraint — a validation-accuracy drop of at most 2 pp for GIF and MEGU, a cross-entropy guard for GNNDelete and CGU, relaxed for CGU on Cornell5 where chance-level membership was reachable only at a utility cost. MEGU and GNNDelete were searched with Optuna's TPE sampler (Akiba et al., 2019) (35 and 21 trials); GIF and CGU with one-dimensional sweeps of the quantity that governs their numerical stability (the effective damping $\lambda_{\text{eff}} = \text{damp} \times \text{scale}$ of GIF's recurrence; CGU's regularization λ). Settings were frozen before any audit world was trained (Table 4). With these settings GIF and CGU reach the pass band on Pokec-n and CGU alone on Cornell5 (Table 5); GIF's update is numerically inert on Cornell5, MEGU saturates near AUROC 0.56–0.60 at any acceptable utility, and GNNDelete's operators leave deleted users *more* traceable than No-op on Pokec-n (0.73).

Table 4: Selected unlearner settings. GIF on Cornell5 is inert at every numerically stable λ_{eff} (its Hessian is ten times stiffer than on Pokec-n and the Pokec-n setting diverges); CGU on Cornell5 reaches chance-level membership only with the utility loss visible in Table 5.

Method	Pokec-n	Cornell5
GIF	scale 16,500, damp 7.05×10^{-4}, 500 it. (λ_{eff}=11.6)	scale 139,360, damp 5×10^{-4}, 500 it. (λ_{eff}=69.7; inert)
MEGU	30 ep., κ 0.847, lr 3.43×10^{-3}, 1 hop, T 0.89, wd 8.4×10^{-4}	same (best trial of the Cornell5 search)
GNNDelete	lr 0.021, α 0.20, 100 ep., MSE (mean)	same
CGU	$\lambda\, 10^{-2}$, Adam 0.01×500, σ 0.01, ϵ 1, K=2	$\lambda\, 10^{-3}$, otherwise same

B.3 The LiRA-style membership test

Score. For candidate v, world ω and release R the auditor submits v's true record — its features and all of its archived friendships — to the released model through the exact two-hop neighborhood of v on the served graph, reads the posterior q of v's true label and records the score $s_R(v,\omega) = \log \frac{q}{1-q}$, the logit scaling of Carlini et al. (2022). Nothing else about the model is observed.

Per-candidate likelihood ratio. The 32 shadow worlds contain v as a member in 16 and as a non-member in 16. For each (v, R) the auditor fits $\mathcal{N}(\mu_v^{\text{in}}, \sigma_v^{\text{in}\,2})$ and $\mathcal{N}(\mu_v^{\text{out}}, \sigma_v^{\text{out}\,2})$ to the shadow scores (population MLE, standard deviations floored at 0.01) and tests with the log-likelihood ratio

$\Lambda_R(v, \omega)$. The threshold for a 1% false-positive rate is fixed on the 16 calibration worlds (16,000 candidate–world rows) and the 16 evaluation worlds are scored exactly once: the pooled AUROC of Λ_R over their 16,000 rows (Table 2), with a 95% percentile-bootstrap interval of the per-world AUROC over the 8 evaluation pairs, and the true-positive rate at the calibrated threshold with the realised false-positive rate (Table 5). Two node-agnostic baselines — the raw confidence and the negative entropy of the posterior — stay at AUROC ≤ 0.54 for every release on both graphs: the membership signal lives entirely in per-node offsets, as the residual picture of Proposition 3 predicts.

Table 5: Membership audit of every release on the evaluation worlds. TPR (%) at the threshold calibrated for a 1% FPR on the calibration worlds, with the realised FPR (%) in parentheses; utility is the retained-graph test accuracy of the served release. Bold: AUROC within 0.05 of GOLD.

	Pokec-n / field of work			Cornell5 / gender		
Release	AUROC	TPR (FPR)	Utility	AUROC	TPR (FPR)	Utility
No-op	0.619	2.8 (0.7)	0.702	0.680	9.1 (0.9)	0.755
GIF	**0.519**	1.0 (0.7)	0.701	0.666	8.3 (0.8)	0.756
MEGU	0.560	5.6 (3.7)	0.692	0.599	4.4 (1.0)	0.747
GNNDelete	0.729	7.8 (0.9)	0.697	0.697	10.0 (1.3)	0.750
CGU	**0.493**	0.9 (0.8)	0.707	**0.506**	0.9 (1.0)	0.709
GOLD	0.498	0.6 (1.0)	0.704	0.506	0.7 (0.8)	0.754
Never	**0.500**	1.2 (1.1)	0.706	**0.497**	1.3 (1.1)	0.756

What the test does and does not certify. Worlds are complete train–unlearn pipelines on one fixed graph with a resampled half of the cohort, so a candidate's shadow scores are not independent across worlds: they share the retained graph, and the Gaussian family is a modelling choice. Both can only lower the test's power: an AUROC at chance is evidence that this auditor detected no trace under this query contract, not a certificate of removal. It is nevertheless the strongest per-node membership evidence available to a user, and the evidence Section 4 shows to be silent where the label is most exposed.

B.4 Per-node quantities

All quantities below are computed for each of the 1,000 candidates from saved model outputs; no model is queried more than the audit already did, except for the ghost queries.

Membership advantage. AUROC_v^R is the Mann–Whitney AUROC of the score $s_R(v, \cdot)$ over the 24 shadow and calibration worlds in which v was a training member against the 24 in which it was not, and $\mathrm{Adv}_v^R = 2\,\mathrm{AUROC}_v^R - 1$ (signed; Figure 2 and the pass band use the magnitude $|\mathrm{Adv}_v^R|$). Under an exact null this statistic has standard deviation 0.17, so a genuinely untraceable user lands in $|\mathrm{Adv}_v| \leq 0.1$ with probability 0.45; the GOLD and Never bands have this size (44% and 43% of users on Pokec-n, 42% and 42% on Cornell5), which is why Figure 2 uses the *trained* model's advantage $|\mathrm{Adv}_v^{\text{No-op}}|$ as its x-axis: it compares w with w_{-v} exactly as Theorem 1 does, it is the same quantity for every release, and it is not noise for the releases that pass. The evaluation worlds are never used for it, so it shares no model with the label measurements.

Ghost accuracy and margin. A candidate is re-inserted into the served graph with every feature equal to the retained-graph mean (zero after standardization) and its archived friendships, and the release is evaluated once through the resulting two-hop neighborhood. With $\hat{p}_v$ the posterior of the label $+1$, the released score on the label scale is $2\hat{p}_v - 1$ (Appendix A.7), the predicted label is its sign and the margin is $m_v^{\text{ghost}} = y_v(2\hat{p}_v - 1)$. Over the 8 evaluation worlds in which v was deleted, $a_{\text{rec}}(v)$ is the fraction with a correct sign and $\bar{m}_v^{\text{ghost}}$ the mean margin; the posterior on the true label used in Figure 2 is $\frac{1}{2}(1 + \bar{m}_v^{\text{ghost}})$. The ghost queries on the evaluation worlds reproduce the audit's saved posteriors to 3×10^{-7}.

Model-free attacker. The dotted curve of Figure 2 and the x-axis of Figure 5 come from an attacker that never queries any model: a logistic decoder over eleven label statistics of the candidate's retained one- and two-hop neighborhood (degree, number and share of neighbors with a training

label, and the label histograms at both hops), fit on the anchors — users held out of every model — with all retained training labels disclosed, its regularization chosen on a held-out 30% of the anchors. Per candidate it yields a posterior on the true label and an accuracy over the same 8 deleted worlds as the ghost. On the pass-band users of Figure 2 it is right in at least six of eight worlds for 80% (Pokec-n) and 82% (Cornell5) of users, against the ghost's 88% and 85%; the class prior is 52.7% and 50.3%.

Unexplained share. m_v is the margin of GOLD queried with v's *true* record, averaged over the 48 shadow and calibration worlds — the retained margin $y_v f(v; w_{-v})$ of Section 3.2, since GOLD is the law of w_{-v} (it agrees with No-op's margin on the worlds that did not train on v to within 0.01 on average). Since a classifier's score lies in $[-1, 1]$, clipping is vacuous and $\tilde{s}_v = 1 - m_v = y_v r_v \in [0, 2]$: it is the theorem's $\max\{y_v r_v, 0\}$, and exceeds 1 exactly on the anti-predicted users; users with $m_v \geq 0$ are aligned (Assumption 1), users with $m_v < 0$ anti-predicted. The ghost's own unexplained share, $1 - \bar{m}_v^{\text{ghost}}$, tracks $\tilde{s}_v$ bin by bin and is about 0.1 larger: a user's own features explain roughly a tenth of the label that the neighborhood does not.

Stored shift and channel. $\Delta_v = \bar{m}_v^{\text{IN}} - \bar{m}_v^{\text{OUT}}$ is the difference between No-op's mean true-record margin over the worlds that trained on v and over those that did not; by Lemma 1 it equals $h_v \tilde{s}_v$, and $\hat{h}_v = \Delta_v / \tilde{s}_v$ is the effective leverage. σ_v is the pooled seed-to-seed standard deviation of the same margin.

B.5 Both blades for every release

Figure 4 shows the two blades of Figure 3 for all seven releases; the numbers quoted in Section 4.2 (mean own advantage per release, the sign change of GNNDelete's advantage, and the residue of each release over GOLD at $\tilde{s}_v \geq 0.8$) are read from it.

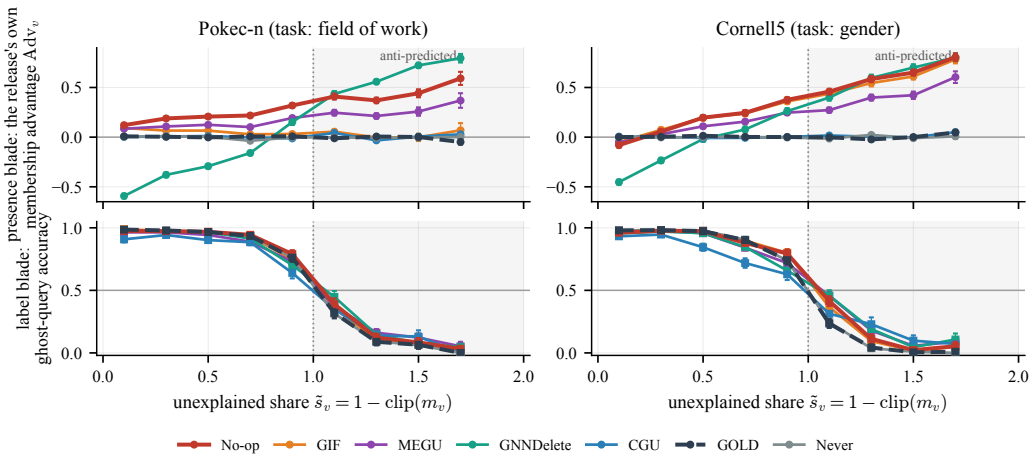


Figure 4: **Unlearning closes the presence blade and cannot move the label blade.** Same x-axis as Figure 3; binned means $\pm$ SE per release. *Top:* each release's *own* membership advantage, the presence it left behind. *Bottom:* accuracy of one ghost query on each release. GOLD (dashed): exact deletion; grey: anti-predicted users.

B.6 A model-independent ordering of the users

The unexplained share $\tilde{s}_v$ of Figure 3 is read from GOLD's output on the user's true record, and the label blade is read from GOLD's output on a blank record, so a reader may ask how much of the red curve is the definition of $\tilde{s}_v$. Figure 5 therefore orders the same users by a quantity computed without any model: the posterior that the model-free attacker of Appendix B.4 assigns to the true label from the labels of retained friends and friends-of-friends. Both directions persist. Across the populated bins the trained model's advantage falls from 0.47 to 0.20 on Pokec-n and from 0.48 to 0.26 on

Cornell5, while the ghost's accuracy after exact deletion rises from 0.24 to 0.98 and from 0.22 to 0.97; the model-free posterior is itself strongly ordered with $\tilde{s}_v$ (Spearman -0.60 and -0.61), so the unexplained share is largely a property of the neighborhood, not of the model that reads it. The presence blade is flatter on this axis than on $\tilde{s}_v$; the model-free posterior sees neither the user's own features nor the friends' features, both of which enter how much the retained data explain.

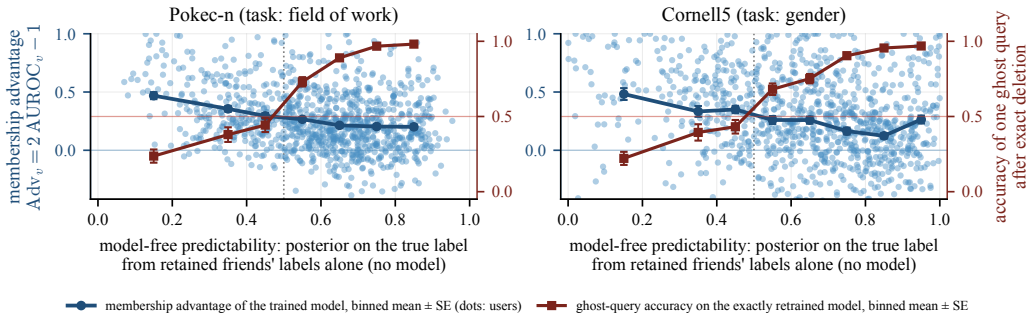


Figure 5: **The two directions on a model-free axis.** Each deleted user is placed at the posterior on its true label assigned by the model-free attacker (retained friends' labels only, no model). Blue, left: signed membership advantage of the trained model (No-op), binned mean ± SE (dots: users). Red, right: accuracy of one ghost query on the exactly retrained model (GOLD). Bins with fewer than 15 users are not drawn.

B.7 Exposure planes for every release

Figures 6–9 show the plane of Figure 2 for all seven releases. With the trained model's advantage on the x-axis (Figures 6 and 7) every release falls along the same frontier: the mean posterior on the true label drops from 0.64–0.69 in the first bin to 0.41–0.46 in the last on Pokec-n and from 0.64–0.73 to 0.41–0.46 on Cornell5, and 20% (16%) of users are not memorized by the original model yet inferable in at least six of eight deleted worlds for 80–88% of them under every release. With each release's *own* advantage on the x-axis (Figures 8 and 9) the curve falls only where presence was retained — No-op ($0.69 \to 0.45$; $0.73 \to 0.45$), MEGU ($0.65 \to 0.45$; $0.71 \to 0.46$), GNNDelete and GIF on Cornell5 ($0.65 \to 0.46$; $0.74 \to 0.44$, where GIF is inert) — and is flat for the releases that pass (GIF on Pokec-n $0.61 \to 0.58$; CGU, GOLD and Never within ± 0.04 across their populated bins, Never on Cornell5 excepted, $0.62 \to 0.73$ on a sparse last bin), for which 41–44% of users fall in the pass band and 67–74% of those are inferable.

B.8 The stored shift and the channel

Two further checks of Theorem 1, both on the original model. Figure 10 tests Lemma 1: the stored shift Δ_v lies below $\tilde{s}_v$ for every aligned user and grows with it, with an effective leverage of median 0.038 (Pokec-n) and 0.033 (Cornell5) — the trained GCN is far from interpolation, which is why the trained model's advantages are moderate and why the near-interpolation assumption of Appendix A.1 is needed only for the direction "unreconstructible ⇒ presence-exposed". Figure 11 tests the channel: with training randomness as the observation noise, the measured advantage of every user is the Gaussian-channel reading $2F_{\mathcal{N}}(\Delta_v/\sqrt{2}\sigma_v) - 1$ of its stored shift, to within an RMSE of 0.036 on both graphs (mean measured 0.263 vs. predicted 0.263 on Pokec-n, 0.256 vs. 0.255 on Cornell5); the theorem's total-variation form is the same curve under a monotone rescaling.

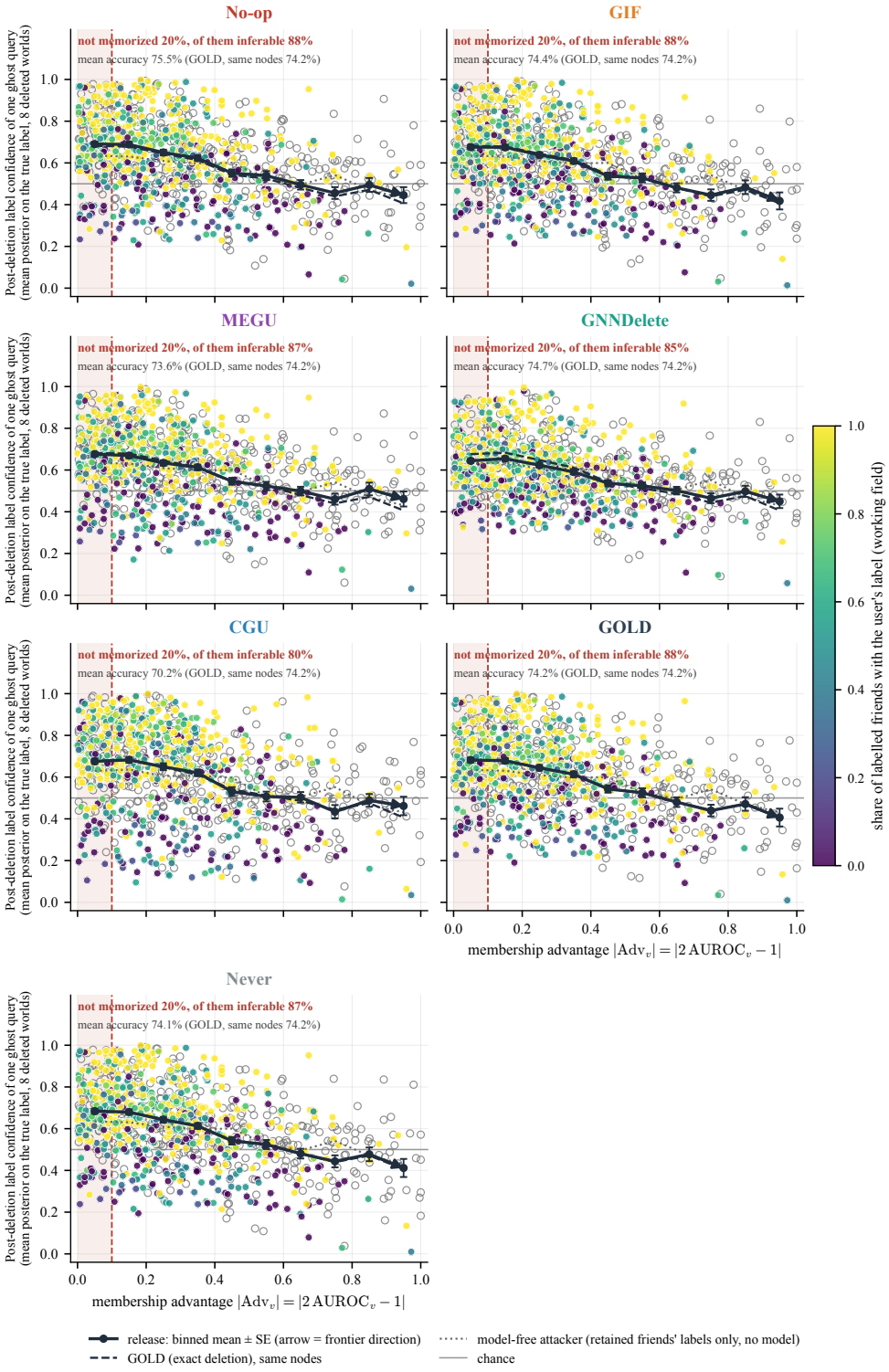


Figure 6: Exposure planes of all seven releases on Pokec-n with the *trained* model's membership advantage $|\mathrm{Adv}_v^{\text{No-op}}|$ on the x-axis. Axes, colors and lines as in Figure 2; the annotation gives the share of users with $|\mathrm{Adv}_v^{\text{No-op}}| \leq 0.1$ and the share of those whose label the ghost recovers in at least six of eight deleted worlds.

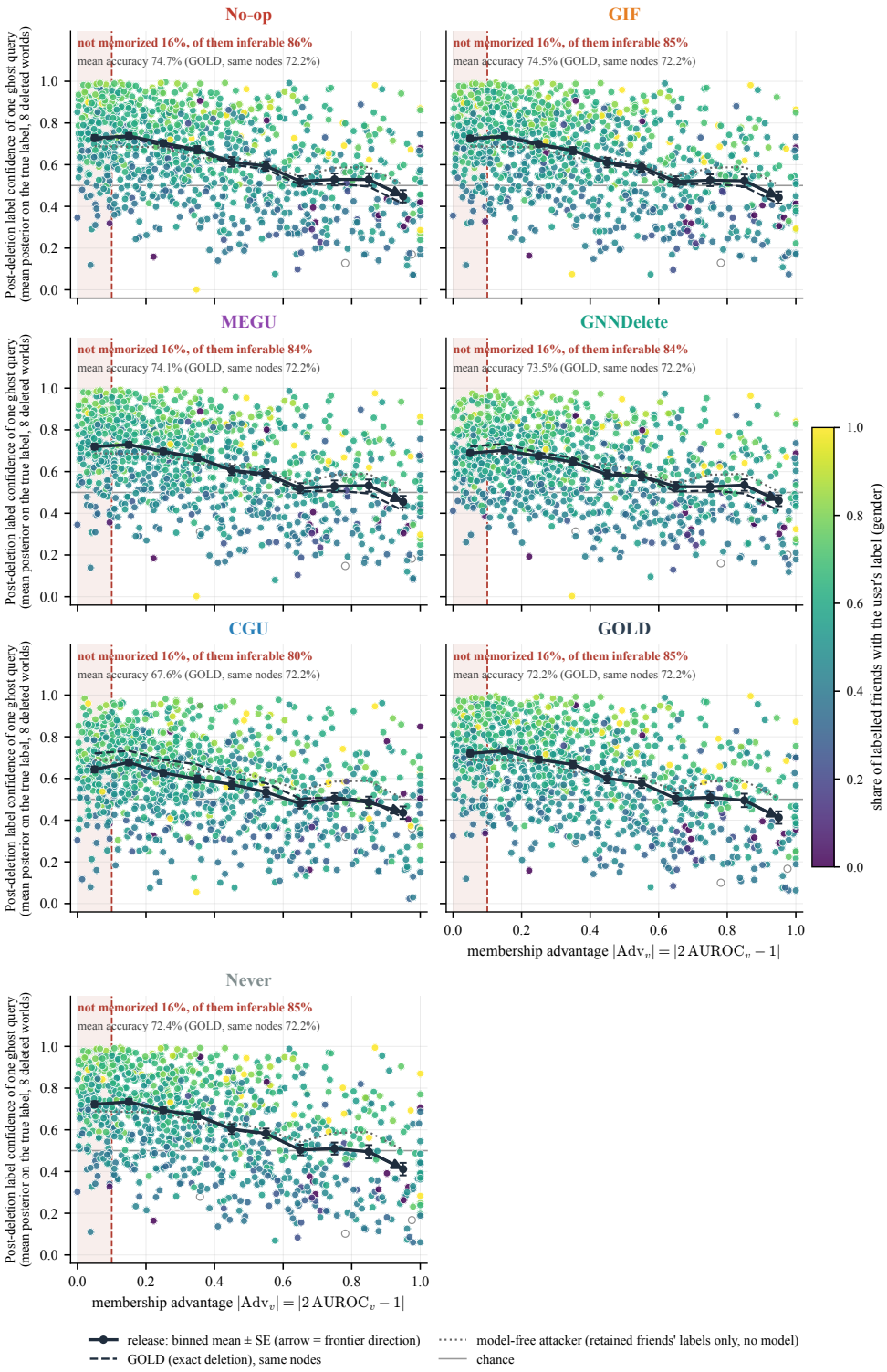


Figure 7: As Figure 6, for Cornell5.

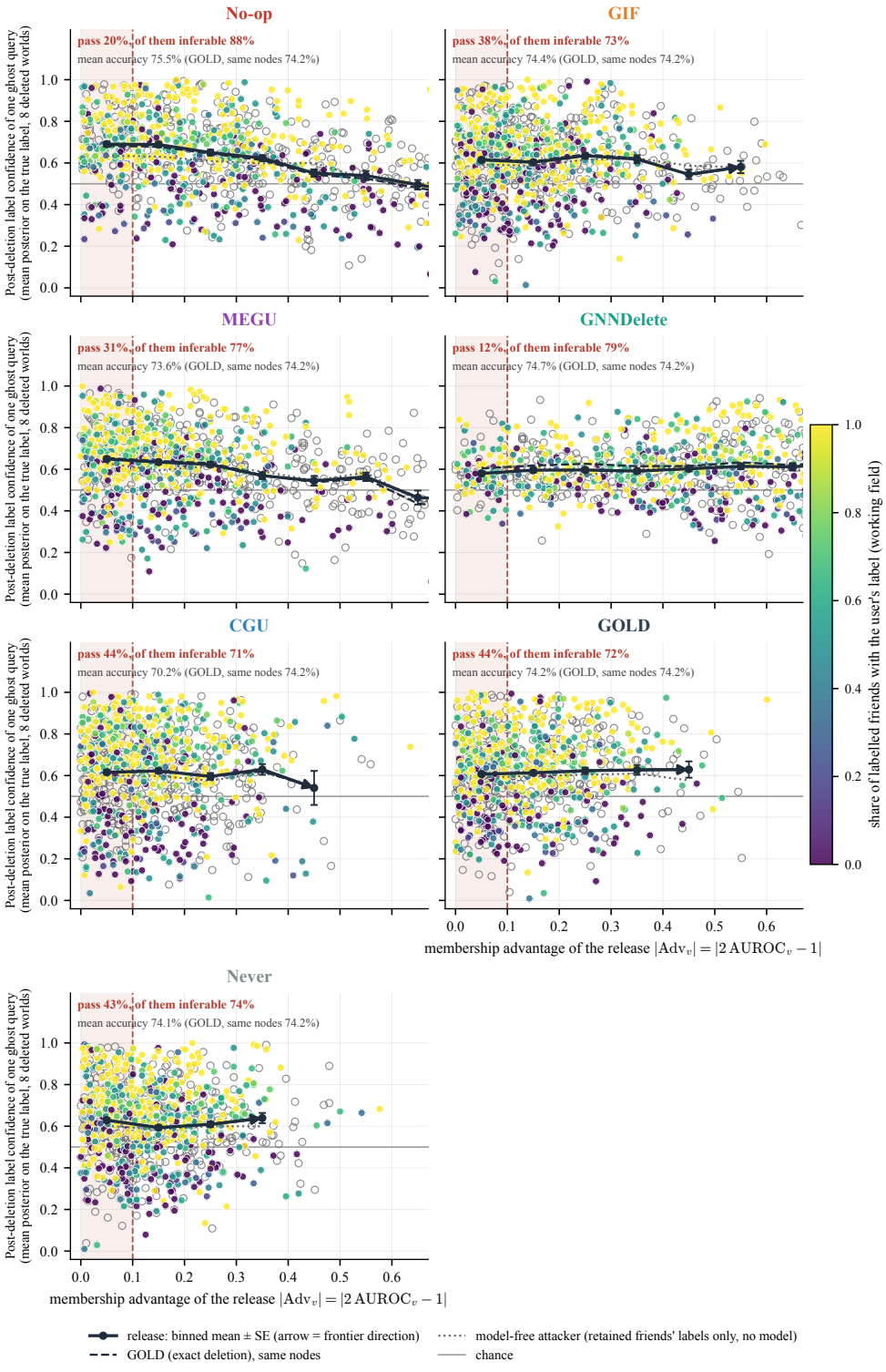


Figure 8: Exposure planes on Pokec-n with each release's *own* membership advantage on the x-axis — the audit's view. For the releases that pass, x is noise and the curve is flat: the audit places the most and the least inferable users in the same band.

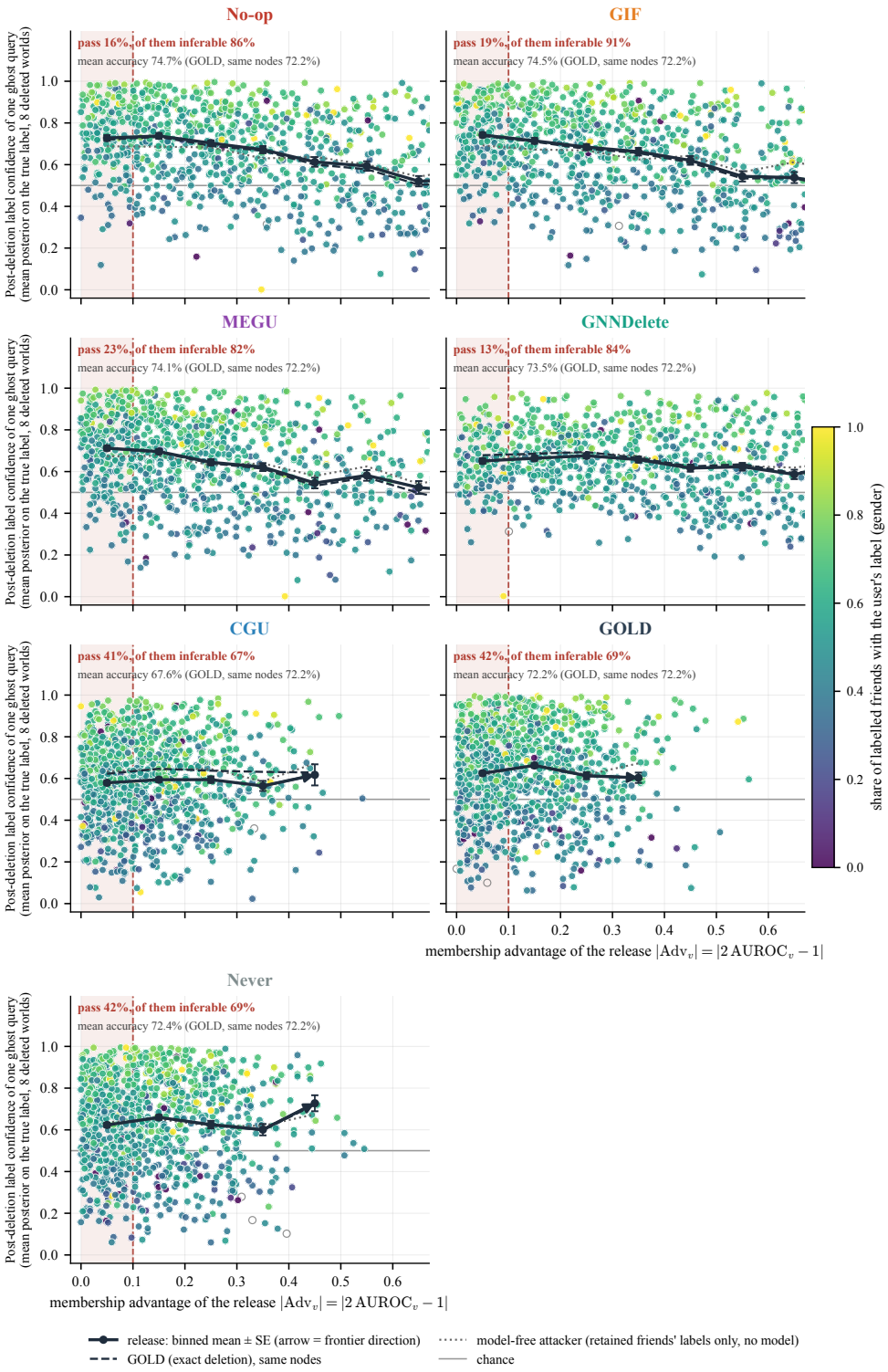


Figure 9: As Figure 8, for Cornell5.

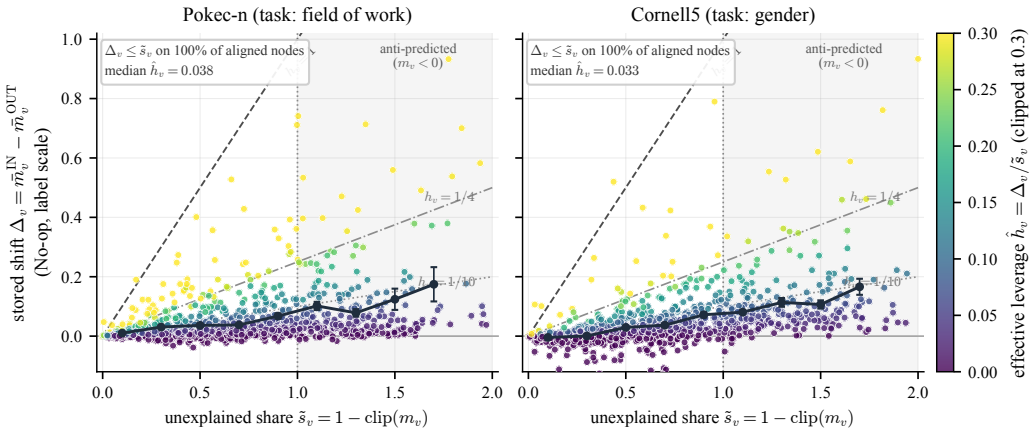


Figure 10: A training label moves the model in proportion to its unexplained part. The stored shift Δ_v against the unexplained share $\tilde{s}_v$; color: effective leverage $\hat{h}_v$; dashed: $h_v = 1$; navy: binned mean $\pm$ SE.

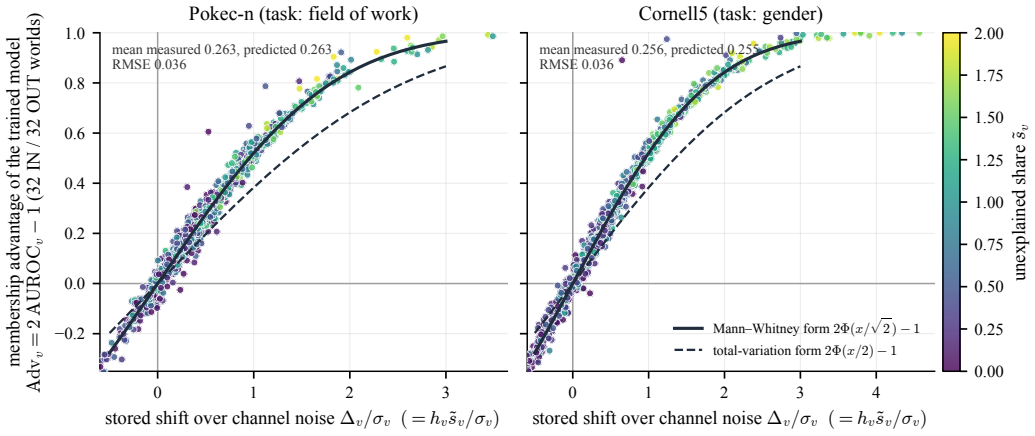


Figure 11: The membership attacker reads the stored shift through a Gaussian channel. Measured membership advantage of the trained model against Δ_v/σ_v; solid: $2F_{\mathcal{N}}(x/\sqrt{2}) - 1$; dashed: the total-variation form of Theorem 1; color: unexplained share.