BEYOND UPDATE GEOMETRY: POST-COMMITMENT AUDITING OF INDIVIDUAL DECISION INFLUENCE FOR PROPOSAL ADMISSION IN DECENTRALIZED FEDERATED GRAPH-LEVEL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized federated graph-level clustering must admit peer proposals while preserving a shared clustering state. Updates that appear ordinary in parameter space can change assignments near small decision margins, so nominal authority need not reflect decision influence. To audit these effects while preserving execution continuity, we introduce Proof-Carrying Operator Consensus (PCOC). Peers commit aligned operator and claim pairs before an unlabeled audit challenge is sampled. Validators measure each proposal's leave-one-out effect and compute simultaneous upper bounds for individual influence in the frozen audit pool. The policy screens proposals using current and stored upper scores, then ranks passing proposals to select a fresh quorum. Stored pairs fill the remaining positions. Across five datasets with matched budgets, mean individual influence was 45.8% and 45.6% lower than for an internal reference under fixed and adaptive attacks. In a separate study with two encoders and 20 paired seeds, attack-induced boundary-flip rates (AI-BFR) were 6.3% and 5.2% lower, corresponding to approximately 0.052 and 0.047 percentage points. Neither difference passed Holm adjustment, with p values of 0.0625 in both attacks. Final clean normalized mutual information (NMI) was 0.1199 versus 0.1217. A transfer record covering 50 contexts found lower influence in the execution pool for PCOC in 43 of 50 contexts.

1 INTRODUCTION

Federated graph-level clustering seeks common cluster structure across private, heterogeneous graph collections (Wang et al., 2026). In decentralized coordination, peers must decide which submitted operators can modify the shared structure in the presence of adversarial peers (Gaucher et al., 2025). In a permissioned network, peers maintain a replicated transcript of proposal admission decisions. Each submitted cluster operator maps a graph to a soft assignment. The transcript identifies the current and stored operator and claim pairs used in each round. Proposals that appear ordinary in parameter space can still shift assignments for graphs near a cluster boundary.

Geometric screening can accept a proposal that still degrades model predictions (Zhao et al., 2026). Nominal weights likewise need not reflect decision influence. When the top two cluster scores are close, small output changes can switch an assignment. Each proposal's influence also depends on the other operators in the pool because removing it changes claim weights and the resulting consensus. Decision authority mismatch denotes the gap between a proposal's nominal authority and its counterfactual effect on cluster decisions. Replacing rejected proposals with stored pairs changes this context. Influence in the frozen audit pool may therefore differ from influence in the execution pool.

Figure 1 shows this dependence for graphs near a cluster boundary. Admission requires measuring how a submitted operator changes the consensus assignments. Sampling probes after observing proposals can compromise statistical validity (Dwork et al., 2015). The audit fixes proposals before sampling and compares counterfactuals in the same frozen audit pool. Cluster coordinate alignment separates label permutations from decision changes. Stored pairs preserve execution after admission failures. The resulting execution pool is audited separately.

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.

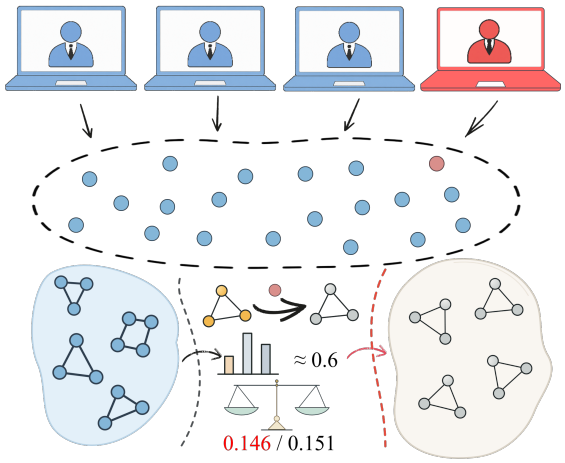

Figure 1: Nominal share and cluster assignments near a decision boundary. Peer icons are schematic. Numerical annotations correspond to the setting with the smallest budget in Section 4.2.

To measure decision influence before granting fresh authority, we introduce Proof-Carrying Operator Consensus (PCOC). Graph encoders supply the representations used by submitted cluster operators, and aligned soft assignments expose their effects on unlabeled graphs. The protocol commits operator and claim pairs, freezes the audit pool, and samples a common challenge to obtain simultaneous upper bounds for individual influence. A deterministic gate compares predeclared current and stored scores, and a ranker selects a fresh quorum from the passing proposals. Evaluation across five graph datasets examines individual influence, boundary stability, attacked clustering quality, and fresh quorum availability.

Our contributions are as follows:

- The decision authority mismatch diagnostic separates nominal authority from counterfactual cluster movement near decision boundaries.
- Auditing after commitment provides simultaneous bounds on individual influence within one round in the frozen audit pool. It relates changes in soft outputs to hard assignments through decision margins.
- Studies with matched budgets and paired seeds evaluate influence, boundary stability, attacked clustering quality, and fresh quorum availability across five graph datasets.

2 Related Work

Federated graph-level clustering. FedGCN combines local structural prototypes to construct consensus representations for graph-level clustering (Liu et al., 2025). PERFECT personalizes aggregation under client heterogeneity (Liu et al., 2026a). FGCN-DKS separates shared cluster structure from graph patterns specific to each client (Wang et al., 2026), and FedAI addresses missing node attributes (Han et al., 2026). These methods learn graph representations under server coordination. Robust federated clustering has also been studied with heterogeneous and adversarial clients using a central aggregator (Bravo & Dalleiger, 2026). PCOC keeps the encoder and clustering objective fixed and audits committed operator and claim pairs, rather than changing representation learning. Node clustering on incomplete federated graphs (Liu et al., 2026b) and clustering from private structural graphs (He et al., 2026) address distinct tasks.

Robust admission and decentralized learning. Krum and Multi-Krum use distances among submissions (Blanchard et al., 2017), coordinate-wise trimmed mean suppresses extreme coordinates (Yin et al., 2018), and RFA estimates a geometric median (Pillutla et al., 2022). These aggregators act on update values rather than assignments. DSpodFL analyzes convergence under sporadic local computation and peer exchanges (Zehtabi et al., 2025). Remove-then-Clip filters and clips neighbor messages (Yang & Ghaderi, 2024). Robust gossip analyses characterize Byzantine breakdown in

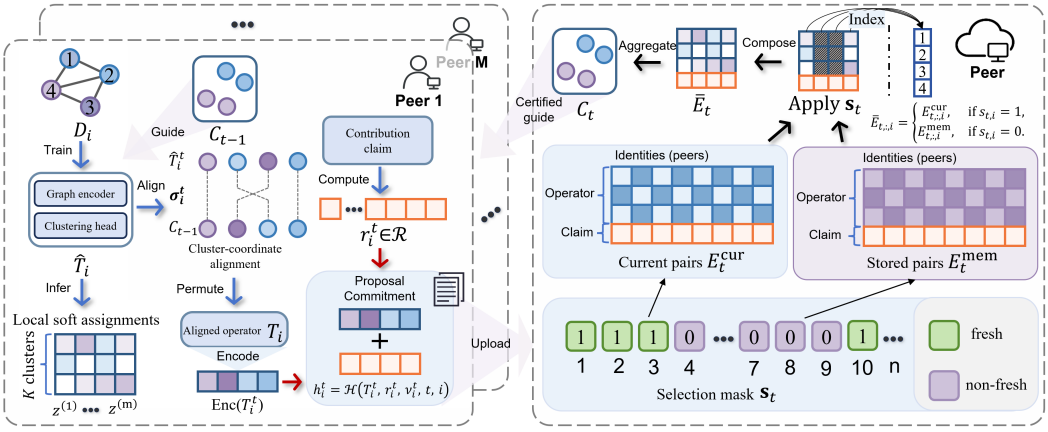

Figure 2: Current pairs E_t^{cur} and stored pairs E_t^{mem}. Selection mask $\mathbf{s}_t$ chooses the pair for each peer, composing tensor $\bar{E}_t$ for aggregation into C_t. The "Certified guide" label marks the recorded output for the mixed execution pool.

peer-to-peer aggregation (Gaucher et al., 2025). FedTGE selects clients using graph energy in a federated graph backdoor defense (Wan et al., 2025). PCOC tests unlabeled cluster decisions within a frozen audit pool. Its permissioned validators replicate one admission transcript and execution state rather than independently filtering neighbors.

Influence and verifiability. Influence functions and Data Shapley attribute outcomes to training examples (Koh & Liang, 2017; Ghorbani & Zou, 2019). Fed-Influence estimates client influence on federated model parameters (Xue et al., 2021). Ripple Shapley attributes data influence within one training run (Zeng et al., 2026). DICE estimates influence cascades across communication hops, supporting contribution attribution and collaborator selection (Zhu et al., 2025). PCOC measures changes in cluster outputs after removing one committed proposal within a round. Admission uses a challenge sampled after commitment and simultaneous individual bounds in the frozen audit pool. Janus cryptographically verifies aggregation correctness using two servers (Pu et al., 2025). PCOC provides statistical coverage for decision influence over the declared probe distribution.

3 Proof-Carrying Operator Consensus

Admission measures how a submitted operator changes cluster assignments near a decision boundary. In round t, each of n permissioned peers may submit a cluster operator T_i^t and nonnegative contribution claim r_i^t. The operator maps an unlabeled graph probe z to a soft assignment in Δ^{K-1}. For a pool A indexed by peer identity, consensus map $C_t(A)$ aligns cluster coordinates, transforms claims into normalized authority, and aggregates the operators. Its hard assignment is $h_t(A, z) = \arg\max_k [C_t(A)(z)]_k$, with a fixed tie rule. Every counterfactual uses a common cluster ordering and retains positive total authority. Alignment rules and reference data are fixed before challenge sampling. Graph structure enters through the encoder and graph probes. The encoder and clustering objective are fixed across admission methods within each comparison.

Valid current submissions form the frozen audit pool A_t^{audit}, where individual influence is measured. Selection combines current pairs with stored pairs for the remaining peers to form the execution pool A_t^{eff} (Figure 2).

Peers finalize a commitment root and a bitmap of valid openings before the round beacon is released. The ledger assumption and the influence guarantee concern distinct properties. The ledger layer assumes at most $f < n/3$ Byzantine peers and provides the assumed finality and verification of commitment openings. The statistical certificate conditions on the frozen audit pool and simultaneously bounds the individual influence of every valid proposal. Joint coalition influence is a separate counterfactual. The protocol requires a fresh quorum of $q = n - f$. Invalid or withheld openings and admission failures reduce fresh quorum availability. An adversary may prepare proposals using

Algorithm 1 One PCOC admission round

Input: Current and stored pairs, quorum q, tolerance β, probe count M_t, ranker $\mathcal{R}_t$, and public round configuration
Output: Execution consensus, fresh identities, and updated stored pairs

1: Align operators and commit canonical operator and claim pairs
2: Close commitments, verify openings, and freeze A_t^{audit}
3: **if** $|A_t^{\mathrm{audit}}| < q$ **then**
4: Record round failure and **return** C_{t-1}^{exec}
5: **end if**
6: Draw M_t probes using the transcript and round beacon
7: Score current and stored counterfactuals on the same probes using (2), (5), and the predeclared score configuration
8: Form passing set F_t by (6)
9: **if** $|F_t| < q$ **then**
10: Record round failure and **return** C_{t-1}^{exec}
11: **end if**
12: Rank passing proposals by (ρ_i^t, i) and select q identities as G_t^{fresh}
13: Form A_t^{eff} by (7)
14: Update stored pairs for G_t^{fresh} and record the transition
15: **return** $C_t(A_t^{\mathrm{eff}})$

public configuration, stored history, and a separate search sample. Once the beacon is released, no peer can revise a committed opening.

Probe selection uses reference output $\pi_t(z) = C_{t-1}^{\mathrm{exec}}(z)$ and margin $\mu_t(z) = \pi_{t,(1)}(z) - \pi_{t,(2)}(z)$. We draw probes from a predeclared unlabeled distribution P_t^{bdry}. It favors graphs with low decision margins, graphs between regions induced by stored operators, and graphs on which previously admitted operators disagree. Its score combines the reference margin, historical disagreement, and a bridge score with weights fixed before the round. The distribution and its support are specified in the Supplementary Material.

A fixed nonnegative transformation g_t maps each declared claim to authority, which is then normalized over the pool under evaluation.

$$w_i^t(A) = \frac{g_t(r_i)}{\sum_{j \in A} g_t(r_j)}, \qquad i \in A. \tag{1}$$

Removing a proposal changes the denominator for every surviving peer. The consensus map also reruns cluster coordinate alignment and aggregation on the remaining operators. Influence must therefore be evaluated from two full consensus outputs, rather than from the removed operator alone.

For $K \geq 2$ and $u, v \in \Delta^{K-1}$, let $d(u, v) = K^{-1}\|u - v\|_1 \in [0, 2/K]$ measure distance between soft outputs. Pointwise and population leave-one-out influence for proposal i are defined as

$$\begin{aligned} Y_i^t(z, A) &= d(C_t(A)(z), C_t(A \setminus \{i\})(z)), \\ I_i^t(A) &= \mathbb{E}_{z \sim P_t^{\mathrm{bdry}}} Y_i^t(z, A). \end{aligned} \tag{2}$$

For a pool with positive total individual influence, decision authority mismatch is the difference between a proposal's share of output movement and its share of nominal authority.

$$\Delta_{\mathrm{DA}}^t(i, A) = \frac{I_i^t(A)}{\sum_{j \in A} I_j^t(A)} - w_i^t(A). \tag{3}$$

The diagnostic is set to zero when total influence is zero. The gate uses upper influence scores. Δ_{DA}^t is reported as a diagnostic.

Committed encodings bind each operator and claim to a peer identity, round, and nonce. The commitment root, opening bitmap, and beacon determine the challenge seed used to draw M_t independent probes from P_t^{bdry}. Attack search, calibration, and evaluation use separate samples, and private graph labels are excluded from the challenge. For each valid proposal i, validators evaluate

$C_t(A_t^{\mathrm{audit}})$ and $C_t(A_t^{\mathrm{audit}} \setminus \{i\})$ on the same probes. These paired outputs yield sample mean $\widehat{I}_i^t$ and unbiased sample variance $\widehat{V}_i^t$ of $Y_i^t(z_j, A_t^{\mathrm{audit}})$.

For $M_t \geq 5$, $\delta_t \in (0,1)$, and $L_t = \log(3|A_t^{\mathrm{audit}}|/\delta_t)$, validators compute an empirical Bernstein upper score (Maurer & Pontil, 2009) as

$$U_i^t = \min\left\{2/K,\ \widehat{I}_i^t + \sqrt{\frac{2\widehat{V}_i^t L_t}{M_t}} + \frac{3L_t}{M_t}\right\}. \tag{4}$$

The cap $2/K$ is the largest possible value of d. The logarithmic term allocates the round's failure probability across all valid openings. Validators can recompute each score without private graph records. Required inputs are committed openings, the frozen bitmap, challenge seed, and public consensus configuration recorded in the round transcript.

The stored comparison replaces only peer i's current pair. With stored pairs initialized at the common warmup state, define

$$\begin{aligned} A_{t,i}^{\mathrm{mem}} &= \big(A_t^{\mathrm{audit}} \setminus \{i\}\big) \cup \{(i, T_i^{\mathrm{last}}, r_i^{\mathrm{last}})\}, \\ \widetilde{Y}_i^t(z) &= d\big(C_t(A_{t,i}^{\mathrm{mem}})(z), C_t(A_t^{\mathrm{audit}} \setminus \{i\})(z)\big)\,. \end{aligned} \tag{5}$$

Current and stored scores use the same probes and leave-one-out background, with only the version at position i changed. All other current openings remain fixed. Validators apply Equation (4) to Y_i^t and $\widetilde{Y}_i^t$ to obtain the formal current and stored scores. The empirical study uses a predeclared calibrated current score $\overline{U}_i^t$ and stored reference score $\overline{L}_i^{t-1\to t}$ from held-out validation. The current pair passes if the calibrated current score is at most $\beta \geq 0$ above the stored reference score, with β fixed before the challenge.

$$F_t = \{i \in A_t^{\mathrm{audit}} \mid \overline{U}_i^t \leq \overline{L}_i^{t-1\to t} + \beta\}. \tag{6}$$

Once the transcript is fixed, the gate is deterministic. Proposition 1 provides coverage for the formal score U_i^t, whereas the empirical gate uses calibrated scores. A bound on current minus stored population influence would require paired differences or joint coverage of a current upper score and a stored lower score. The failure budget δ_t controls simultaneous coverage and β sets the permitted excess over the stored reference.

The predeclared ranker assigns $\rho_i^t = \mathcal{R}_t(i; \overline{\mathbf{U}}_t, \overline{\mathbf{L}}_t, H_{t-1})$, where H_{t-1} is the recorded history. Passing proposals are ordered by increasing (ρ_i^t, i), with peer identity resolving ties. The first q form G_t^{fresh}. The mask $s_{t,i}$ selects their current pairs and stored pairs elsewhere, forming the execution pool.

$$\begin{aligned} A_t^{\mathrm{eff}} = &\{(i, T_i^t, r_i^t) \mid i \in G_t^{\mathrm{fresh}}\} \\ &\cup \{(i, T_i^{\mathrm{last}}, r_i^{\mathrm{last}}) \mid i \in \mathcal{P} \setminus G_t^{\mathrm{fresh}}\}. \end{aligned} \tag{7}$$

Only selected current pairs update storage for the next round.

If fewer than q peers have valid openings or pass the gate, the protocol records failure and retains the previous execution state. The transcript records the opening root and bitmap, challenge digest, scores, gate outcome, selected identities, and execution pool. It binds each certificate to A_t^{audit} and separately identifies the execution pool A_t^{eff}. Exact state transitions are specified in the Supplementary Material. The ranker is a public deterministic protocol input. Its serialized score fields, history inputs, and tie rule are fixed before challenge sampling and included in the round configuration hash. The coverage statement is invariant to the particular deterministic ordering within the frozen pool.

Sampling after pool closure gives simultaneous coverage within one round for individual influence in the frozen audit pool. Decision margins then connect movement of soft outputs to assignment changes after removing one proposal.

Proposition 1 (Simultaneous influence coverage within one round). Condition on the transcript after valid openings are frozen and before beacon release. It fixes A_t^{audit}, the consensus maps, P_t^{bdry}, and $M_t \geq 5$. If the M_t challenge probes are drawn independently from P_t^{bdry}, then

$$\Pr\big[\forall i \in A_t^{\mathrm{audit}},\quad I_i^t(A_t^{\mathrm{audit}}) \leq U_i^t\big] \geq 1 - \delta_t. \tag{8}$$

The probability is over challenge sampling conditional on this transcript. Simultaneous coverage permits selecting proposals using their observed scores while retaining coverage for their individual influence in the same frozen audit pool. Joint removals and substitution of stored pairs change the counterfactual and require separate bounds. Proofs and the precise counterfactual definitions appear in the Supplementary Material.

Proposition 2 (Hard assignments and decision margins). Let $m(p) = p_{(1)} - p_{(2)}$ be the margin between the two largest coordinates of p. For any fixed pool A, proposal $i \in A$, and $\tau > 0$,

$$\begin{aligned} &\Pr_{z \sim P_t^{\mathrm{bdry}}}[\arg\max C_t(A)(z) \neq \arg\max C_t(A \setminus \{i\})(z)] \\ &\leq \Pr_{z \sim P_t^{\mathrm{bdry}}}[m(C_t(A)(z)) \leq \tau] + \frac{K I_i^t(A)}{\tau}. \end{aligned} \tag{9}$$

The first term measures the probe probability near a decision boundary. A numerical bound on assignment flips requires estimating this term alongside influence.

4 Experiments

4.1 Experimental Setup

Using the benchmarks and sources in Table 1, a Dirichlet partition with concentration 0.5 distributes graphs across six honest and two malicious peers. Each run has eight warmup, twelve attack, and ten recovery rounds. The paired comparison uses a Weisfeiler–Lehman encoder (Shervashidze et al., 2011) and a graph isomorphism network (GIN) encoder (Xu et al., 2019). The study with matched budgets and the ablations use GIN. Methods within each study share partitions, initializations, attack schedules, and evaluation probes.

Table 1: Graph benchmarks from TUDataset (Morris et al., 2020).

Dataset	Graphs	Classes	Domain	Source
MUTAG	188	2	Molecular	(Debnath et al., 1991)
PROTEINS	1113	2	Protein	(Borgwardt et al., 2005)
NCI1	4110	2	Molecular	(Wale & Karypis, 2006)
ENZYMES	600	6	Protein	(Borgwardt et al., 2005; Schomburg et al., 2004)
IMDB-BINARY	1000	2	Social	(Yanardag & Vishwanathan, 2015)

Both attacks perturb the aligned operator and its claim. Sixteen settings cross four budgets for operator displacement with four claim budgets. The fixed attack uses a predeclared perturbation. The adaptive attack searches within the same constraints using past transcripts and an independent sample. Proposals are committed before challenge probes are drawn.

Comparators include no defense, coordinate-wise trimmed mean (Yin et al., 2018), geometric median aggregation (Pillutla et al., 2022), and Multi-Krum (Blanchard et al., 2017). These update geometry methods test whether geometric screening controls the measured decision effect. Two internal controls isolate admission and stabilization. The internal reference keeps the operator and claim interface, budgets, partitions, schedule, and probes while following its reference admission path. The fresh quorum predecessor represents selection before the full stabilization composition. The internal reference is the controlled comparator for the audit effect. The primary score configuration uses an audit budget of 0.55. The internal reference uses the same budget and held-out validation protocol. Remove-then-Clip uses the adapted graph-level configuration with a trim ratio of 0.2.

The exported probe records separate search, calibration, audit, and evaluation probes. Search and audit, audit and evaluation, and calibration and evaluation have zero recorded overlap, and methods share evaluation hashes within each matched context. Task outcomes use the same held-out evaluation probes for all methods.

Individual influence measures leave-one-out movement in the frozen audit pool (Equation (2)). Attack-induced boundary-flip rate (AI-BFR) measures the fraction of evaluation probes whose as-

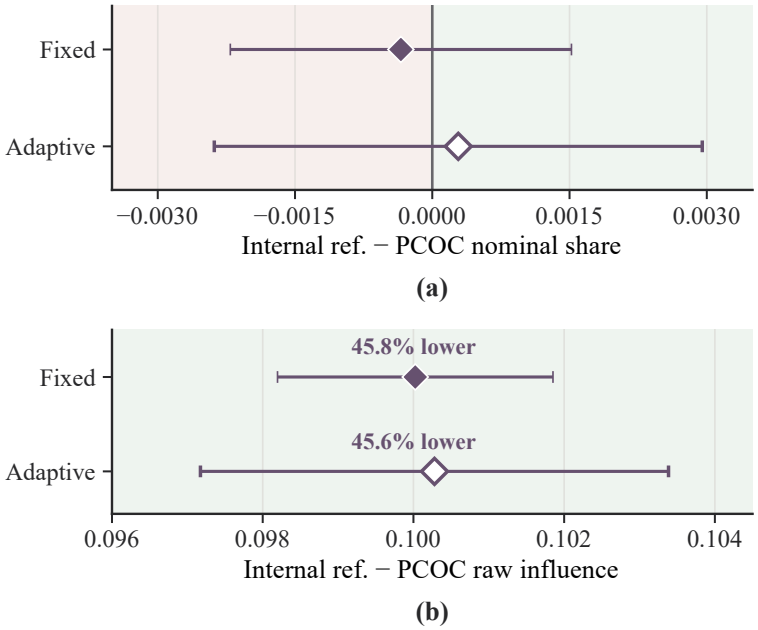

Figure 3: Effects under matched budgets. Values show the internal reference minus PCOC across five datasets, with 95% t_4 intervals. Panels (a) and (b) show nominal share and individual influence, respectively. "Raw influence" denotes individual influence before normalization into shares. Panel (b) uses an expanded horizontal scale.

signment differs between paired attacked and unattacked runs with the same partition, initialization, and seed. Conditional fresh quorum availability counts complete current quorums among eligible attack rounds. Malicious fresh admission counts selected malicious proposals. Final clean normalized mutual information (NMI) measures clustering quality outside the attacked trajectory. Metric formulas and denominators appear in the Supplementary Material. Effects under matched budgets average three seeds and 16 budgets within each dataset. Main comparison effects average two encoders and 20 paired seeds. In both studies, the five dataset effects are the units for the t_4 intervals. Two prespecified one-sided exact Wilcoxon tests form one family for Holm adjustment.

4.2 Comparison Study

Matching operator and claim budgets isolates the admission decision from differences in adversary resources. Under these conditions, mean individual influence is 45.8% lower for PCOC under fixed attack and 45.6% lower under adaptive attack (Figure 3). These relative reductions use the declared scale d_K, which averages output movement over cluster coordinates. Paired contrasts are computed within each dataset before equal dataset aggregation. All five dataset effects and all 16 budget effects favor PCOC. At the lowest budget, detector area under the curve remains near 0.60, while malicious nominal share changes little under either attack. Influence therefore varies more than nominal authority in this setting. Absolute effects and values by budget appear in the Supplementary Material. The influence scale is K^{-1} times the output ℓ_1 distance, with range $[0, 2/K]$. We compute contrasts within each dataset before aggregation. The Supplementary Material gives the total variation conversion and applies it to the execution transfer record.

Across paired trajectories, PCOC has the lowest mean AI-BFR (Table 2). The three update geometry baselines produce higher AI-BFR than no defense on this task. Ungated methods reach 100% conditional fresh quorum availability by construction.

Every dataset shows a lower mean AI-BFR for PCOC under both attacks (Figure 4). The absolute mean differences are 0.000521 and 0.000469, and neither prespecified test passes Holm adjustment ($p = 0.0625$ for each attack). Fresh quorums form in 98.68% of eligible attack rounds. Malicious fresh admission is 1.18% across 400 combinations of dataset, encoder, seed, and attack.

Table 2: Paired comparison. Fixed and adaptive columns report AI-BFR in $\times 10^{-3}$. Fresh denotes conditional fresh quorum availability.

Method	Fixed ↓	Adapt. ↓	Clean NMI ↑	Fresh (%) ↑
No defense	15.218	16.031	.1190	100.00
Trimmed mean	17.452	18.241	.1203	100.00
Geo. median	25.406	26.170	.1207	100.00
Multi-Krum	23.242	24.116	.1198	100.00
Fresh quorum pred.	38.617	39.603	.1176	100.00
Internal ref.	8.245	9.018	**.1217**	100.00
PCOC	**7.725**	**8.548**	.1199	98.68
Δ vs. internal ref.	-6.3%	-5.2%	–	–

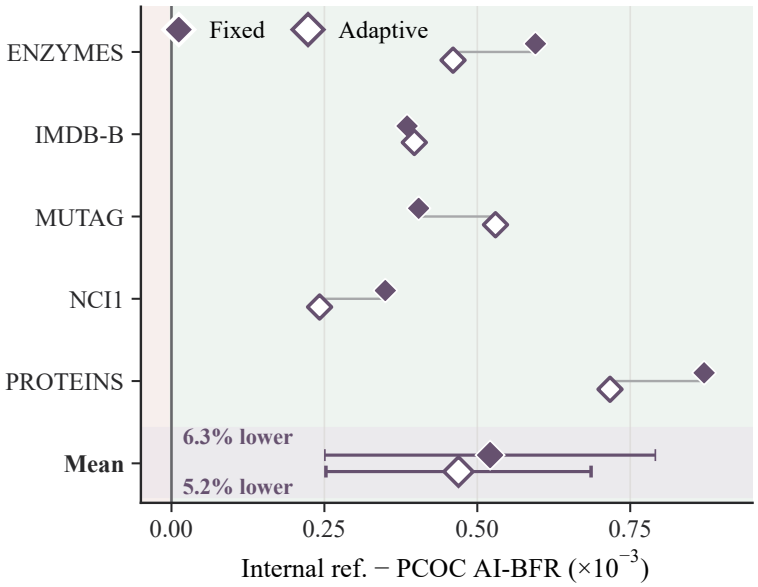

Figure 4: Paired AI-BFR effects by dataset. Markers show the internal reference minus PCOC after averaging two encoders and 20 seeds. Diamonds show mean effects with 95% t_4 intervals. IMDB-B denotes IMDB-BINARY.

The lower AI-BFR also holds on ring, random regular, and scale-free communication graphs with dropout rates up to 30%. At 30% dropout, mean AI-BFR is 0.00926, and fresh quorums form in 97.2% of eligible attack rounds. Relative to no defense, runtime, peak memory, and communication increase by factors of 1.55, 1.18, and 2.08. The Supplementary Material reports these topology and resource factors together with the five-dataset transfer record.

4.3 Ablation Study

Component removals test whether scoring, commitment, and ledger enforcement contribute to boundary stability. In the GIN ablation, removing influence scoring produces the largest AI-BFR increase relative to the complete protocol (Table 3). Removing commitment or ledger enforcement also increases AI-BFR under both attacks.

Removing stabilization lowers AI-BFR but reduces fresh quorum availability by about 14 percentage points and changes clean NMI from approximately 0.1355 to 0.1316, exposing the stability and participation tradeoff.

Table 3: Relative AI-BFR increase after component removal.

Removed component	Fixed (%)	Adaptive (%)
Influence scoring	+31.6	+28.6
Commitment	+18.3	+16.2
Ledger enforcement	+13.7	+11.4

5 Discussion

Aggregation based on update distance (Blanchard et al., 2017) examines parameter geometry. The audit measures how removing one proposal changes soft cluster assignments on graph probes. Stored substitution changes the background, and the transfer record shows lower execution pool influence for PCOC in 43 of 50 paired contexts. The formal certificate covers individual influence in the frozen audit pool.

Commitment finality fixes the audit pool, the beacon fixes the challenge, and the transcript records both the audited and executed compositions. Permissioned validators therefore replicate the ordering of commitment, sampling, admission, and execution. Remove-then-Clip (Yang & Ghaderi, 2024) filters messages within neighbor aggregation, whereas this protocol audits graph-level cluster outputs.

6 Conclusion

This paper introduced PCOC, an admission protocol that audits individual changes in cluster decisions after commitment. Matched-budget studies reduce measured influence, and the transfer record extends this comparison to the executed composition. Paired task results lower attack-induced boundary flips, while the Holm-adjusted tests do not pass the 0.05 threshold. Decision influence therefore complements update geometry for proposal admission in decentralized graph-level clustering.

AI USE STATEMENT

Generative AI tools were used solely for language polishing. The authors take responsibility for the final content.

REFERENCES

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html.

Karsten M. Borgwardt, Cheng Soon Ong, Stefan Sch"onauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21 (suppl_1):i47–i56, 2005. doi: 10.1093/bioinformatics/bti1007. URL https://academic.oup.com/bioinformatics/article/21/suppl_1/i47/202991.

Martín Bravo and Sebastian Dalleiger. Robust federated clustering under heterogeneity and adversaries. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3718–3726, 2026. URL https://proceedings.mlr.press/v300/bravo26a.html.

Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991. doi: 10.1021/jm00106a046. URL https://pubs.acs.org/doi/10.1021/jm00106a046.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2350–2358, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/bad5f33780c42f2588878a9d07405083-Abstract.html.

Renaud Gaucher, Aymeric Dieuleveut, and Hadrien Hendrikx. Unified breakdown analysis for Byzantine robust gossip. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18868–18896, 2025. URL https://proceedings.mlr.press/v267/gaucher25a.html.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Renda Han, Junlong Wu, Wenxuan Tu, Jingxin Liu, Haotian Wang, and Jieren Cheng. Federated graph-level clustering network with attribute inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21585–21593, 2026. doi: 10.1609/aaai.v40i26.39307. URL https://ojs.aaai.org/index.php/AAAI/article/view/39307.

Guanxiong He, Zheng Wang, Jie Wang, Liaoyuan Tang, Rong Wang, and Feiping Nie. Towards federated clustering: A client-wise private graph aggregation framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21619–21627, 2026. doi: 10.1609/aaai.v40i26.39311.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Jingxin Liu, Jieren Cheng, Renda Han, Wenxuan Tu, Jiaxin Wang, and Xin Peng. Federated graph-level clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18870–18878, 2025. doi: 10.1609/aaai.v39i18.34077.

Jingxin Liu, Wenxuan Tu, Renda Han, Junlong Wu, Haotian Wang, Guohui Liu, Xiangyan Tang, and Yue Yang. Personalized federated graph-level clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23721–23729, 2026a. doi: 10.1609/aaai.v40i28.39546.

Jingxin Liu, Wenxuan Tu, Haotian Wang, Renda Han, Haoyi Li, Junlong Wu, and Xiangyan Tang. Causally-aware attribute completion for incomplete federated graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23730–23738, 2026b. doi: 10.1609/aaai.v40i28.39547.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/012.pdf.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020. URL https://eprints.cs.univie.ac.at/6469/.

Krishna Pillutla, Sham M. Kakade, and Zaïd Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135. URL https://doi.org/10.1109/TSP.2022.3153135.

Lang Pu, Jingjing Gu, Chao Lin, and Xinyi Huang. Janus: Dual-server multi-round secure aggregation with verifiability for federated learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49927–49945, 2025. URL https://proceedings.mlr.press/v267/pu25a.html.

Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(suppl_1):D431–D433, 2004. doi: 10.1093/nar/gkh081. URL https://academic.oup.com/nar/article/32/suppl_1/D431/2505276.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011. URL https://jmlr.org/papers/v12/shervashidze11a.html.

Nikil Wale and George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. In *Sixth International Conference on Data Mining*, pp. 678–689. IEEE, 2006. doi: 10.1109/ICDM.2006.39. URL https://doi.org/10.1109/ICDM.2006.39.

Guancheng Wan, Zitong Shi, Wenke Huang, Guibin Zhang, Dacheng Tao, and Mang Ye. Energy-based backdoor defense against federated graph learning. In *International Conference on Learning Representations*, pp. 93913–93942, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea64883d500d31738cd39eb49a748fa4-Abstract-Conference.html.

Xiaobao Wang, Renda Han, Ronghao Fu, and Di Jin. Federated graph-level clustering network with dual knowledge separation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/971f1e59cd956cc094da4e2f78c6ea7c-Abstract-Conference.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Yihao Xue, Chaoyue Niu, Zhenzhe Zheng, Shaojie Tang, Chengfei Lyu, Fan Wu, and Guihai Chen. Toward understanding the influence of individual clients in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10560–10567, 2021. doi: 10.1609/aaai.v35i12.17263. URL https://ojs.aaai.org/index.php/AAAI/article/view/17263.

Pinar Yanardag and S. V. N. Vishwanathan. Deep graph kernels. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374. ACM, 2015. doi: 10.1145/2783258.2783417. URL https://doi.org/10.1145/2783258.2783417.

Caiyi Yang and Javad Ghaderi. Byzantine-robust decentralized learning via remove-then-clip aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21735–21743, 2024. doi: 10.1609/aaai.v38i19.30173. URL https://ojs.aaai.org/index.php/AAAI/article/view/30173.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659, 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *International Conference on Learning Representations*, pp. 59567–59636, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9521b6e7f33e039e7d92e23f5e37bbf4-Abstract-Conference.html.

Dewen Zeng, Wenlong Tian, Haozhao Wang, Jianfeng Lu, Weijun Xiao, and Zhiyong Xu. Ripple shapley: Data influence attribution in one federated training run. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28085–28093, 2026. doi: 10.1609/aaai.v40i33.40034. URL https://ojs.aaai.org/index.php/AAAI/article/view/40034.

Xiaojie Zhao, Jinqiao Shi, Yi Li, Junmin Huang, and Chongru Fan. Good gradients poison your model: Evading defenses in federated learning via boundary-adaptive perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13262–13270, 2026. doi: 10.1609/aaai.v40i16.38328. URL https://ojs.aaai.org/index.php/AAAI/article/view/38328.

Tongtian Zhu, Wenhao Li, Can Wang, and Fengxiang He. DICE: Data influence cascade in decentralized learning. In *International Conference on Learning Representations*, pp. 69007–69042, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ac8fbba029dadca99d6b8c3f913d3ed6-Abstract-Conference.html.