WHERE PRIVACY BELONGS: PLACEMENT DIAGNOSIS AND CERTIFIED SELECTION FOR PRIVATE COUNTERFACTUAL EXPLANATIONS ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations for graph neural networks (GNNs) find the minimal intervention that flips a node's prediction—the gold standard of recourse—but computing one requires reading sensitive graph structure, and releasing it discloses that structure. The two existing placements both fail. Privatizing the graph *before* explaining corrupts the explanation target on exactly the borderline nodes that need recourse, manufacturing *spurious flips*: interventions that flip the privatized graph but not the true one. Explaining on the clean graph and perturbing the released explanation resists certification: our re-audit of the standard heuristic shows that its implied full-release budget reaches 573–753 on Cora and 256 on CiteSeer—orders of magnitude beyond its advertised per-entry budget—with worst-case single-entry leakage at AUC 1.0. We propose PRIVCFS, which replaces certification-by-optimization with certification-by-construction: counterfactual *selection* over a fixed, data-independent candidate universe—edge interventions from a public prior graph, feature interventions from a public schema, with no-op semantics so that every neighboring graph shares the same output support. A validity-gated, clipped utility whose *global* sensitivity bound $\Delta u \leq 1$ requires no optimizer sensitivity analysis is released through the exponential mechanism, giving *pure ε-DP for the complete released object*, composable over queries—to our knowledge the first such guarantee for counterfactual explanations on graphs. Empirically, privacy noise is the cheapest stage of the pipeline: the sampled release retains 94–97% of its support-restricted non-private optimum on the recourse population and 83–95% on the general population at $\varepsilon{=}8$; the optimal likelihood-ratio edge-inference audit attains AUC 0.50 on average and 0.59 in the worst pair, against the heuristic baseline's worst entry of 1.0 under the same attack; and the mechanism transfers to a 15K-node graph with a 0.96 valid rate at $\varepsilon{=}8$. The dominant cost is instead a *measurable, monotone price* in public disclosure, readable off one table before any budget is spent—turning explanation privacy from an accounting risk into a purchasable decision for the data owner.

1 INTRODUCTION

Graph neural networks (GNNs) have become the default machinery for learning on relational data—social networks, molecular graphs, financial transaction graphs—much of which is inherently sensitive (Kipf & Welling, 2017). When such models support consequential decisions, counterfactual explanations are the gold standard of recourse: they answer "what minimal change to the input would have flipped the prediction?" (Wachter et al., 2018). On graphs, counterfactual explainers search for a minimal intervention—deleting a few edges or masking a few features of the target node—that flips the GNN's prediction (Lucic et al., 2022; Bajaj et al., 2021). A counterfactual explanation, however, is only as private as the data it reads: to certify which edges matter, the explainer must inspect the target node's neighborhood, and the released explanation itself (which edges were deleted, which features were masked) then discloses structure the data owner considers sensitive. Graph-structure leakage is a demonstrated, practical threat: influence-based attacks such as LinkTeller recover private edges from model behavior (Wu et al., 2022), and released explanations amplify model-inversion attacks (Zhao et al., 2021).

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.

The natural fix suggested by the differential-privacy literature (Dwork & Roth, 2014) is to privatize the data *before* explaining—add calibrated noise to the adjacency, then run an off-the-shelf explainer. This *graph-level placement* fails in a diagnosable way: DP noise changes the predictions of exactly the low-margin (*borderline*) nodes on which recourse is sought, and the explainer then optimizes a flip against a corrupted target, producing *spurious flips*—interventions that flip the prediction on the privatized graph but *not* on the true one. The explanation is privately "correct" and publicly wrong (Sec. 3.2).

The mirror-image placement—*explain on the clean graph, privatize the released explanation*—keeps the target intact but resists certification: the released object is the output of a non-convex optimization whose sensitivity to a one-edge change admits no tractable global bound. The standard fix, a *per-entry* sensitivity assumption, yields an attractive per-entry budget ($\varepsilon \approx 7.2$) but does *not* certify the release: the mechanism publishes the whole vector, and our re-audit of this exact design measures full-release budgets of 573–753 (Cora) / 256 (CiteSeer), with the removed edge's own entry fully informative on the worst adjacent pair (entry AUC 1.0; App. A). DP is a worst-case guarantee, and $\approx$8–9% of adjacent pairs land in different optimizer basins: heuristic output perturbation is, at best, a *randomized release without a guarantee*.

In this paper we propose a placement that is certifiable *by construction*: instead of explaining through an optimizer and perturbing its output, we *select* the explanation from a fixed, data-independent candidate universe and release the selection through the exponential mechanism (McSherry & Talwar, 2007). PRIVCFS (Private CounterFactual Selection) draws edge candidates from a public prior graph and feature candidates from a public schema, scores them with a validity-gated, clipped utility of *global* sensitivity $\Delta u \leq 1$, and releases one intervention set sampled with probability $\propto \exp(\varepsilon\, u / 2\Delta u)$; deleting a nonexistent edge is a no-op, so every neighboring graph shares the same output support, and the release is *pure ε-DP for the complete output*, composable over queries. Our main contributions are:

- **A placement principle for private graph explanations.** We formalize *graph-level DP* (privatize the adjacency, then explain) versus *explanation-level DP* (explain on the clean graph, privatize the release), identify prediction pollution and spurious flips as the two channels through which the former corrupts counterfactual reasoning, and show that heuristic output perturbation is not certifiable at its advertised budget (App. A); at equal operation space the placement gap is the entire utility (Sec. 4.4).
- **PRIVCFS.** A certified private counterfactual selector on a data-independent output support: public-prior edge candidates plus public-schema feature candidates, a validity-gated utility with a provable global sensitivity bound, and an exact exponential-mechanism release that is pure ε-DP for the complete released object, with a utility guarantee and composition over queries (Secs. 3.3–3.5).
- **The privacy–utility frontier as a purchasable price list.** At $\varepsilon = 8$, PRIVCFS retains 83–97% of its support-restricted non-private optimum—privacy noise is the cheapest stage of the pipeline—while the dominant cost is a *measurable, monotone price* in public disclosure, readable off one table before any budget is spent. A likelihood-ratio edge-inference audit attains AUC 0.50 (max 0.59) against the heuristic baseline's worst entry of 1.0, and the mechanism transfers to a 15K-node core with a 0.96 valid rate (Sec. 4).

2 RELATED WORK

Counterfactual explanations on graphs. The explainability literature for GNNs offers two families of methods. *Factual* explainers, exemplified by GNNExplainer (Ying et al., 2019), identify the subgraph most responsible for *preserving* the prediction (Yuan et al., 2023). *Counterfactual* explainers instead search for the minimal perturbation that *flips* the prediction: CF-GNNExplainer optimizes a continuous relaxation of edge deletions (Lucic et al., 2022), RCExplainer adds robustness under distribution shift (Bajaj et al., 2021), GCFExplainer shifts the unit of explanation to the whole graph (Huang et al., 2023), and COMBINEX integrates feature- and structure-level search in one objective (Giorgi et al., 2025); causal evaluation through the probability of necessity and sufficiency (Cai et al., 2025) and the broader counterfactual-learning literature (Guo et al., 2025) complete the picture. These methods universally assume full access to a clean, non-private graph, and—as we show—naïvely adding privacy upstream destroys exactly the borderline explanations

they are designed to produce. Our selector is compatible with any of their search objectives *as a scoring oracle*: whatever objective produced the intervention set, the set itself can be scored and released through the private-selection layer of Sec. 3.5.

Privacy in graph learning. DP-SGD (Abadi et al., 2016) with RDP composition (Mironov, 2017) and its Opacus implementation (Yousefpour et al., 2021) is the standard recipe for private deep learning; on graphs, locally private GNNs (Sajadmanesh & Gatica-Perez, 2021) and node-level private GNN training (Daigavane et al., 2022; Zhang et al., 2024) protect the *training of the model* or the *publication of the graph*. None studies what happens when an explanation is computed on top of the private artifact. This matters because GNN predictions are fragile to structural perturbation (Zügner et al., 2018): an adversary who perturbs edges to flip predictions and a privacy mechanism that perturbs edges to hide them are, on borderline nodes, the same operation. Explanation-level DP is *complementary* to model-level DP: after a privately trained model, the explanation release still needs its own budget, and our release layer composes with either.

Privacy of explanations and counterfactual fairness. The GDPR-motivated right to explanation (Wachter et al., 2018) made counterfactuals a legal concept, and counterfactual fairness (Kusner et al., 2017) ties them to causal reasoning over sensitive attributes. The closest prior work combines privacy with counterfactual generation on *tabular* data: privately trained class prototypes (Yang et al., 2022), DP counterfactual retrieval (Nomeir et al., 2024), diverse counterfactuals under DP training (Vo et al., 2023), and DP attribution maps released alongside private classifiers (Harder et al., 2020). Attack studies show that explanation outputs themselves leak private information (Zhao et al., 2021), and membership (Shokri et al., 2017) and edge inference (Wu et al., 2022) quantify the stakes of graph leakage. None addresses the graph-specific channel we diagnose—*structure leakage through released interventions*—or provides a pure-ε guarantee over graph structure while preserving the validity of minimal interventions. Our contribution—selecting from a data-independent support with a globally bounded utility—is the graph-specific instance of the classical private-selection design (McSherry & Talwar, 2007).

3 Methodology

3.1 Problem Formulation and Threat Model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ be an attributed graph with $N = |\mathcal{V}|$ nodes, $M = |\mathcal{E}|$ edges, and feature matrix $X \in \mathbb{R}^{N \times F}$, and let f be a GNN classifier, pretrained and *frozen* thereafter, that outputs class logits $f(\mathcal{G}) \in \mathbb{R}^{N \times C}$. For a target node v with original prediction $\hat{y}_{\text{orig}} = \arg\max_y f(\mathcal{G})[v]_y$, a *counterfactual explanation* is an intervention

$$S = (S_E, S_F), \qquad S_E \subseteq \mathcal{E}, \quad S_F \subseteq [F], \tag{1}$$

that deletes the edges in S_E and masks the feature dimensions in S_F of node v, such that the counterfactual graph $\mathcal{G}' = \mathcal{G} \ominus S$ satisfies $f(\mathcal{G}')[v] \neq \hat{y}_{\text{orig}}$. Among valid interventions, the explanation should be *minimal*. The flip target we use throughout is the *runner-up* class—the class other than $\hat{y}_{\text{orig}}$ with the highest predicted probability, the standard counterfactual target (Wachter et al., 2018; Lucic et al., 2022).

Threat model. The data owner holds a graph whose *edges* are sensitive (social ties, molecular bonds, transactions). She releases counterfactual explanations to a data analyst. The adversary observes the complete released output and aims to infer whether a specific edge is present in $\mathcal{G}$. A mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if, for any two *neighboring* graphs $\mathcal{G}, \mathcal{G}'$ that differ in exactly one undirected edge (replace-one adjacency) and any output set O,

$$P\big(\mathcal{M}(\mathcal{G}) \in O\big) \;\leq\; e^{\varepsilon}\, P\big(\mathcal{M}(\mathcal{G}') \in O\big) + \delta. \tag{2}$$

We make the primary guarantee *edge-level* with $\delta = 0$ (pure DP). *Public* objects—part of the release interface, fixed before the mechanism runs—are: the node set and IDs, the feature matrix X (a feature-entry guarantee is a separate accounting left to future work), the class set, the target v (named by an external query), the candidate universe U_v and public snapshot $\mathcal{G}_{\text{pub}}$ (Sec. 3.3), and the frozen backbone f. The private object is the edge set $\mathcal{E}$: the guarantee bounds the *incremental* leakage of $\mathcal{E}$ relative to the snapshot—what the snapshot discloses is public prior knowledge and is never claimed hidden. If the backbone was trained on $\mathcal{E}$, model release is an additional, separately accounted channel (Sec. 4.4); our theorems condition on a public, fixed f. The adversary may query adaptively; by basic composition, k releases consume budget $k\varepsilon$ (Cor. 1).

3.2 Where Privacy Fails Upstream: The Placement Diagnosis

Graph-level placement privatizes the adjacency before any explanation is computed: with two symmetric entries per undirected edge, the L2 sensitivity of the full-matrix Gaussian release is $\sqrt{2}$:

$$\tilde{A} = \mathrm{clamp}\big(A + \mathcal{N}(0, \sigma^2 I), 0, 1\big), \qquad \varepsilon = \sqrt{2}\,\sqrt{2\ln(1.25/\delta)}\,/\,\sigma. \tag{3}$$

The counterfactual explainer then runs unchanged on $\tilde{\mathcal{G}} = (\tilde{A}, X)$. Two failure modes follow, both rooted in a corrupted *explanation target*. *Prediction pollution*: on borderline nodes—smallest top-two margin $\mu(v) = \max_y p_y(v) - \max_{y \neq \hat{y}(v)} p_y(v)$, exactly where recourse is sought—$f(\tilde{\mathcal{G}})[v]$ may differ from $f(\mathcal{G})[v]$ before the explainer does anything. *Spurious flips*: the minimal intervention found on $\tilde{\mathcal{G}}$ may flip $\tilde{\mathcal{G}}$ but not $\mathcal{G}$. At the operating budget, graph-level DP pollutes 60–92% of targets and reports 23–37% spurious flips, 100% of them on polluted nodes (Sec. 4.4).

Explanation-level placement—explain on the clean graph, privatize the release—removes the failure structurally: the explanation target is never touched, so spurious flips are impossible by construction. The remaining question is entirely about *certification*: can the release be given a valid (ε, δ) guarantee? The standard heuristic is Gaussian perturbation of the released mask logits under a *unit per-entry sensitivity* assumption. We show in App. A that this assumption does not certify the actual release: it is a per-entry budget, not a full-release one, and the measured from-scratch sensitivity exceeds the assumption by $\sim 20\times$ on basin-jump pairs.

3.3 Data-Independent Candidate Support

For a public target node v, the release mechanism selects from a fixed candidate universe

$$C_v \;=\; \Big\{ S = (S_E, S_F) \;:\; S_E \subseteq U_v^E,\; |S_E| \le k_E^{\max},\; S_F \subseteq U_v^F,\; |S_F| \le k_F^{\max} \Big\}, \tag{4}$$

with two public, data-independent ingredient sets:

Edge candidates from a public prior graph. The data owner discloses a public snapshot $\mathcal{G}_{\mathrm{pub}}$ of the graph, which must be an *externally fixed* disclosure—fixed before, and independent of, the current private edge set (e.g., a historical snapshot or public metadata). Our experiments instantiate it as a nested random ρ-fraction of the edges to *simulate* such a disclosure ("what the platform discloses or what is inferable from public metadata"); the $\rho<1$ arms are therefore *information-condition* experiments, and an actual snapshot release would carry its own, separately accounted budget. The edge universe is

$$U_v^E = \big\{\, (v, u) : u \in N_{\mathrm{pub}}(v) \,\big\} \tag{5}$$

truncated to the first $k_E = 12$ pairs (a two-hop extension is possible and is ablated in the released results). Crucially, U_v^E is a set of *node pairs*, not of edges: whether the pair is an actual edge of the private graph is unknown to the analyst, and

$$\mathcal{G} \ominus (u, w) \;=\; \mathcal{G} \quad \text{if } (u, w) \notin \mathcal{E}, \tag{6}$$

i.e., deleting a nonexistent edge is a no-op. Features are public in the primary model, so feature candidates are likewise data-independent: the top-k_F nonzero dimensions of x_v ranked by the public first-order influence $|x_{v,d}| \cdot \left\| W^{(1)}_{:,d} \right\|_2$ of the frozen backbone,

$$U_v^F = \text{top-}k_F \big\{ d : x_{v,d} > 0, \text{ ranked by } |x_{v,d}| \, \left\| W^{(1)}_{:,d} \right\|_2 \big\}. \tag{7}$$

Masking a dimension is a public operation with zero sensitivity in the edge-adjacency model.

Lemma 1 (Support independence)**.** *Given the fixed public context—the snapshot $\mathcal{G}_{\mathrm{pub}}$, the features, the backbone, and the target—for neighboring graphs $\mathcal{G}, \mathcal{G}'$ that differ in one edge, the candidate universe C_v is identical. The release support—the set of outputs with positive probability—is therefore data-independent.*

Proof. U_v^E and U_v^F depend only on the public snapshot, the public features, and the public backbone; edge deletions act through the no-op semantics of Eq. (6), which changes only whether a deletion has an effect, never the set of candidate outputs. □

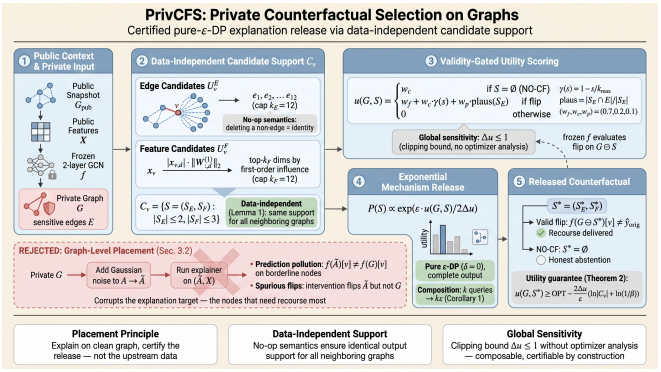

Figure 1: Overview of PRIVCFS. The release selects from a fixed, data-independent candidate universe (public prior snapshot + public feature schema); the validity-gated utility carries a global sensitivity bound; the exponential mechanism releases the complete intervention set with pure ε-DP. The crossed-out graph-level placement corrupts the explanation target and induces spurious flips (Sec. 3.2).

3.4 A VALIDITY-GATED UTILITY WITH A GLOBAL SENSITIVITY BOUND

The utility of a candidate S on the private graph $\mathcal{G}$ is

$$u(\mathcal{G}, S) = \begin{cases} w_c, & S = \varnothing, \\ w_f + w_c\gamma(s) + w_p \mathrm{plaus}(S_E), & \text{flip}, \\ 0, & \text{else}, \end{cases} \tag{8}$$

where $s = |S_E| + |S_F|$, $\gamma(s) = 1 - \frac{s}{k_{\max}}$ with $k_{\max} = k_E^{\max} + k_F^{\max}$, the "flip" case applies iff $f(\mathcal{G} \ominus S)[v] \neq \hat{y}_{\text{orig}}$, $\mathrm{plaus}(S_E) = |S_E \cap \mathcal{E}|/|S_E|$ (a function of $\mathcal{G}$) rewards interventions on edges that actually exist, and $w_f + w_c + w_p = 1$ (we use $0.7, 0.2, 0.1$). $S = \varnothing$ is the "no intervention sampled" output. The utility is *validity-gated*: only interventions that actually flip the prediction receive the flip reward, and non-flipping interventions are worth strictly less than the $\varnothing$ output. Gating the reward does *not* gate the support: a non-flipping candidate has utility 0 and exponential weight 1, so it remains releasable. The release is therefore a privately scored *candidate suggestion* with a measured validity rate (Sec. 3.6); filtering invalid candidates per graph would break support independence and is deliberately not done. A *soft* variant giving partial credit to margin reduction retains the same sensitivity structure and is evaluated as an ablation (App. G).

Lemma 2 (Global sensitivity)**.** *For neighboring graphs $\mathcal{G}, \mathcal{G}'$, $\sup_{S \in C_v} |u(\mathcal{G}, S) - u(\mathcal{G}', S)| \le \Delta u$, with $\Delta u = w_f + w_c + w_p = 1$ for the validity-gated utility of Eq. (8) and $\Delta u = w_f + w_p$ for the soft variant.*

Proof. Each term lies in $[0, 1]$: the flip indicator is binary; the size term is data-independent; $\mathrm{plaus}_{\mathcal{G}}(S_E)$ changes only through the single edge that differs between $\mathcal{G}, \mathcal{G}'$, and by at most $1/|S_E| \le 1$. Hence the validity-gated utility is a convex combination of terms whose individual changes are each bounded by 1, giving $\Delta u \le w_f + w_c + w_p$. The bound follows from the public clipping range *alone*—no optimizer sensitivity analysis is involved—which is precisely the property the optimizer-based designs of App. A lack. Tighter model-Lipschitz bounds are a direction for further budget efficiency. □

3.5 MECHANISM AND GUARANTEES

The release mechanism is the exponential mechanism over the fixed support C_v (Algorithm 1):

$$P\big(\mathcal{M}(\mathcal{G}) = S\big) \;=\; \frac{\exp\big(\varepsilon\, u(\mathcal{G}, S)/(2\Delta u)\big)}{\sum_{S' \in C_v} \exp\big(\varepsilon\, u(\mathcal{G}, S')/(2\Delta u)\big)}. \tag{9}$$

Theorem 1 (Pure DP for the complete release)**.** *$\mathcal{M}$ of Eq. (9) is ε-DP with $\delta = 0$ for replace-one edge adjacency, for the complete released object S—not per coordinate.*

Proof. Immediate from Lem. 2 and the standard exponential-mechanism guarantee (McSherry & Talwar, 2007): for any output S, the ratio $P(\mathcal{M}(\mathcal{G})=S)/P(\mathcal{M}(\mathcal{G}')=S)$ is bounded by $\exp\big(\varepsilon \cdot \sup_S |u(\mathcal{G},S) - u(\mathcal{G}',S)|/(2\Delta u)\big) \cdot \exp(\cdots) \le e^{\varepsilon}$, using $|u(\mathcal{G},S) - u(\mathcal{G}',S)| \le \Delta u$ for both directions of the ratio. □

To our knowledge, this is the first pure-ε guarantee for a *complete* counterfactual explanation release on graphs: prior explanation-privacy mechanisms either protect a privatized artifact upstream (graph- or model-level, Sec. 3.2) or carry per-entry budgets that do not certify the released object (App. A).

Corollary 1 (Composition)**.** *For k adaptive queries—any sequence of target nodes, with the universe and budget fixed per query—the composed release is $k\varepsilon$-DP under basic composition.*

Theorem 2 (Utility guarantee)**.** *Let* $\mathrm{OPT} = \max_{S \in C_v} u(\mathcal{G}, S)$*. With probability at least* $1-\beta$ *over the draw of Eq.* (9)*,*

$$u(\mathcal{G}, S^\star) \;\ge\; \mathrm{OPT} - \frac{2\Delta u}{\varepsilon}\Big(\ln|C_v| + \ln\tfrac{1}{\beta}\Big). \tag{10}$$

Theorem 2 is the standard exponential-mechanism utility bound (McSherry & Talwar, 2007); it is worst-case and conservative at the operating parameters (at ε=8 the penalty term exceeds the utility range), while the empirical concentration is far tighter (Sec. 4.3). The release semantics are *probabilistic*: $\varnothing$ means "no intervention sampled," and the valid counterfactual rate of Sec. 3.6 is the reported quality quantity, with non-empty invalid releases counted as failures, never filtered.

Frozen backbone: a two-layer GCN. Following CF-GNNExplainer (Lucic et al., 2022), the model being explained is a two-layer GCN (Kipf & Welling, 2017) with the symmetric renormalized adjacency and forward pass

$$\hat{A} = D^{-1/2}(A{+}I)D^{-1/2}, \quad D_{ii} = \textstyle\sum_j (A{+}I)_{ij}, \qquad H^{(1)} = \mathrm{ReLU}\big(\hat{A}XW^{(1)}\big), \quad Z = \hat{A}H^{(1)}W^{(2)}. \tag{11}$$

The backbone is pretrained and frozen; all explanations treat it as a black box. Its two-hop receptive field makes candidate evaluation exact on the induced subgraph (verified bit-identical to full-graph forward, App. B), and its fragility to structural perturbation (Zügner et al., 2018) is precisely the channel through which graph-level DP corrupts borderline targets.

3.6 Evaluation Metrics

For a target node v with clean prediction $y_v = f(\mathcal{G})[v]$, and a released intervention S: **valid counterfactual rate (valid-CF)** $\Pr_S\big[f(\mathcal{G}\ominus S)[v] \neq y_v\big]$—the probability that the released candidate truly flips the clean-graph prediction (we report the deterministic ceiling *starCF* and the sampled rate; non-empty non-flipping releases count as failures, never filtered); **targeted success**: the valid flip lands on the runner-up class $\hat{y}_{\mathrm{cf}}$; **intervention size** $\mathrm{Size}(S) = (|S_E| + |S_F|)/(M+F)$; **empty-output rate** $\Pr[\mathcal{M}$ releases $\varnothing]$ (not a certificate that no counterfactual exists), with the *non-empty conditional validity* separating the two failure modes; **utility regret** $\mathrm{OPT} - \mathbb{E}\,u(\mathcal{G}, S^\star)$ against the support-restricted optimum; **release collision** $\sum_{S \in C_v} P(S)^2$ (repeated-query stability); and **spurious-flip/prediction-pollution rates** (diagnosis only, Sec. 3.2).

4 Experiments

4.1 Setup

Datasets and backbone. Cora (2,708 nodes) and CiteSeer (3,327 nodes) are the main benchmarks, and a 15,000-node degree-core subset of ogbn-arxiv (160,944 edges, 40 classes) provides the scale-transfer benchmark (Table 4, appendix). The backbone is a two-layer GCN (hidden 32) trained 200 steps with Adam (lr 0.01, weight decay 5×10^{-4}) on the Planetoid training split and frozen; test accuracy 0.809 (Cora) / 0.682 (CiteSeer).

Mechanism. Public snapshot strength $\rho \in \{0.3, 0.5, 0.7, 1.0\}$ (nested; the ρ<1 arms are the certified arms, ρ=1.0† marks the data-dependent upper reference); edge universe k_E=12, $k_E^{\max}$=2; feature universe k_F=12, $k_F^{\max}$=3; utility weights $(0.7, 0.2, 0.1)$, validity-gated (Δu=1.0); budget grid $\varepsilon \in \{0.1, 0.5, 1, 2, 4, 8\}$; exact enumeration with 100 Monte-Carlo release draws per cell. All results are mean ± std over 3 seeds (pooled over targets).

Table 1: Primary: the privacy–utility frontier on the *general* population (16 publicly drawn targets per seed, mean ± std over n=48, 3 seeds). starCF: support-restricted non-private ceiling; P(flip): valid rate of the sampled release (empty-output rate 0.1–0.4% throughout). ρ: public-prior strength (ρ<1 certified; †: data-dependent upper reference). Retention: P(flip)@ε=8/starCF.

Dataset	ρ	starCF	@0.5	@1	@2	@4	@8	Ret.@8
Cora	0.5	0.167±.373	0.078±.194	0.089±.215	0.098±.242	0.116±.278	0.139±.320	84%
Cora	0.7	0.250±.433	0.117±.235	0.123±.248	0.142±.282	0.172±.323	0.213±.377	85%
Cora	1.0†	0.312±.464	0.136±.237	0.146±.248	0.166±.272	0.211±.336	0.260±.397	83%
CiteSeer	0.5	0.146±.353	0.076±.185	0.085±.210	0.091±.224	0.118±.287	0.139±.336	95%
CiteSeer	0.7	0.229±.420	0.114±.217	0.128±.239	0.152±.282	0.184±.340	0.219±.401	95%
CiteSeer	1.0†	0.292±.455	0.145±.235	0.157±.259	0.181±.295	0.223±.352	0.268±.421	92%

Table 2: Diagnostic: the same frontier on *borderline* targets (10 lowest-margin nodes per seed, n=30, 3 seeds). Columns as in Table 1.

Dataset	ρ	starCF	@0.5	@1	@2	@4	@8	Ret.@8
Cora	0.5	0.500±.500	0.295±.315	0.328±.344	0.369±.380	0.420±.424	0.484±.484	97%
Cora	0.7	0.633±.482	0.373±.331	0.392±.340	0.448±.364	0.525±.411	0.614±.467	97%
Cora	1.0†	0.900±.300	0.513±.235	0.544±.243	0.631±.252	0.742±.266	0.866±.291	96%
CiteSeer	0.5	0.367±.482	0.172±.235	0.199±.275	0.235±.315	0.288±.381	0.348±.458	95%
CiteSeer	0.7	0.567±.496	0.287±.287	0.314±.303	0.357±.334	0.447±.400	0.534±.468	94%
CiteSeer	1.0†	0.767±.423	0.397±.263	0.439±.279	0.500±.307	0.611±.355	0.738±.408	96%

Targets. Three protocols over test nodes: *random* (16 uniformly sampled nodes, fixed by a public draw—the primary, deployment-fair *general population*), *stratified* (16 nodes by margin quartiles), and *borderline* (the 10 lowest-margin nodes, the recourse population on which the placement diagnosis bites). The margin-based protocols are *diagnostics* of population heterogeneity, not a routing mechanism: who is explained is chosen from public information in deployment.

Implementation. Candidate evaluation is exact via the two-hop induced subgraph (bit-identical to full-graph forward; App. B). Hardware: one RTX 4090, PyTorch 2.8, PyTorch Geometric 2.6. Code and raw results accompany the submission.

4.2 Certification Checks: Exhaustive Small-Graph Unit Tests

Before any result is read as certified, we verify the guarantee of Sec. 3.5 at the implementation level by exhaustive small-graph checks (protocol and numbers in App. C): 40 random graphs on 8 nodes against up to 40 neighbors each, checking (i) identical release supports, (ii) $\max_S |u(\mathcal{G}, S) - u(\mathcal{G}', S)| \leq \Delta u$, and (iii) $\max_S P(\mathcal{M}(\mathcal{G})=S)/P(\mathcal{M}(\mathcal{G}')=S) \leq e^{\varepsilon}$. Every check passes on all 3,200 pairs at ε=4 (measured sensitivity $0.93 \leq 1.0$; worst release ratio 10× below the bound), and three boundary configurations at ε=8 pass the same checks.

4.3 Main Results: The Privacy–Utility Frontier

This section carries the paper's core empirical claim: privacy is the cheapest stage of the pipeline, and the remaining costs form a purchasable price list. Table 1 reports the primary frontier on the general population (fixed, publicly drawn targets—the deployment-fair table) and Table 2 the same grid on borderline targets (the recourse-population diagnostic), all at three seeds; the frontier curves and the deployment-guidance view are plotted in App. H (Figs. 4 and 5). Seven findings stand out.

(i) Privacy is the cheapest stage of the pipeline. The sampled release retains 83–95% of its support-restricted non-private ceiling at ε=8 on the general population (Table 1) and 94–97% on the borderline diagnostic (Table 2), degrading gracefully at smaller budgets. The mechanism costs almost nothing *where the candidate support suffices*—on the general population the binding constraint is the support itself (ceiling 0.15–0.31), the honest price of enumerable, certifiable candidates rather than of noise.

(ii) The public prior is a purchasable, monotone price list. The borderline ceiling moves monotonically with disclosure strength—0.37–0.50 at the certified arm ρ=0.5, 0.57–0.63 at ρ=0.7, 0.77–0.90 at the data-dependent upper reference—with the general-population ceiling moving in step (0.15–0.25/0.23–0.25/0.29–0.31; since the snapshots are nested, monotonicity holds by construction). Every additional

fraction of disclosed structure buys a measurable slice of ceiling *without* spending privacy budget, readable off Tables 1 and 2 before committing anything—a capability no prior mechanism offers. The price list extends to the support's *breadth*: the public influence ranking concentrates the flipping candidates in its head, so a $27\times$ smaller universe leaves the borderline frontier unchanged, while a two-hop extension buys $+0.067/+0.100$ of ceiling at $\sim 4\times$ candidates (Ablation (i)).

(iii) Who to explain is decided from public information—not private margins. On uniformly sampled nodes the sampled rate reaches 0.14–0.22 at $\varepsilon{=}8$; on borderline nodes the same mechanism reaches 0.35–0.87 (Fig. 5). The ceiling is a one-number function of the target's margin (≈ 0.008 vs. 0.69–0.75), which is exactly why margin-based routing is *not* free: announcing a margin-selected population discloses information about private edges. Margin stratification is used here only to *diagnose* heterogeneity; deployments may instead target a fixed, publicly drawn list at no privacy cost, or spend a separate selection budget $\varepsilon_{\text{select}}$ composed with the release budget.

(iv) Validity is the stable, budget-priced release quantity. Two independent releases rarely agree (collision 0.2–0.3%), while the *validity* of the release is stable and priced by the budget. The *empty-output rate* is 0.1–0.4% across all rows: the $\varnothing$ candidate is swamped by the mass of near-tied non-flipping candidates, so the non-empty conditional validity equals P(flip) to within rounding, and non-empty invalid releases are counted as failures (never filtered). Each query spends budget (Cor. 1), so the frontier tables are also the per-query price schedule.

(v) The budget–validity curve has the shape the theory predicts. Utility regret decays smoothly with ε (borderline Cora $\rho{=}0.7$: 0.368 at $\varepsilon{=}0.5$ down to 0.164 at $\varepsilon{=}8$), and P(flip) tracks the ceiling with a gap that is largest where near-flipping candidates are numerous (low ε, low ρ)—the regret term of Theorem 2 realized empirically. At $\varepsilon{=}0.1$ the release approaches uniform sampling over C_v, the mechanism's information-free floor. *(vi) Any-flip versus runner-up recourse is a deployment choice*: on borderline targets the targeted (runner-up) rate reads $0.44/0.53/0.69$ (Cora) and $0.29/0.38/0.41$ (CiteSeer) at $\varepsilon{=}8$, the gap set by the backbone's class structure, not by the mechanism. *(vii) Releases spend the size budget on validity*: sampled releases use 3.1–3.9 of the $k_{\max}{=}5$ budget elements, so the size penalty of Eq. (8) mainly orders candidates *within* the flipping set—validity first, minimality second, tunable through $k_{\max}$ and w_c.

A four-stage loss decomposition (App. D) orders the three prices a deployment pays: on the borderline chain the support restriction costs $-0.07/-0.17$, the public prior is the dominant purchase ($-0.40/-0.40$), and the privacy noise is the cheapest element ($-0.02/-0.02$); on the stratified chain the support restriction dominates instead—*who* is explained selects which price dominates. An edge–feature decomposition of the certified arm (App. D) shows its utility is carried by public-prior *edge* interventions: the feature-only arm is exactly zero.

4.4 Mechanism Comparison

Table 3 places PrivCFS on the two axes that decide private-explanation design—placement and certification—against the baseline suite; the decisive column is *certification*: which methods actually carry a guarantee for their complete release at the quoted budget.

Three readings. *First, the certification column is the decisive ranking dimension.* PrivCF-out's $1.000/0.958$ CSR—identical to the non-private ceiling—comes from a mechanism whose advertised budget does not certify its release (App. A). Among the mechanisms that *do* certify their complete release, the certified league on the recourse population is PrivCF-SGD ($0.500/0.533$) and PrivCFS (0.484–$0.614/0.348$–0.534 across $\rho{=}0.5/0.7$): comparable validity, with PrivCFS offering pure ε-DP (vs. (ε, δ)), one joint draw with set-level minimality, and a ρ-priced support; on the general population PrivCFS's CSR is set by its own support ceiling—the purchasable price of certifiable enumeration, not of noise. *Second, the placement diagnosis survives under certification.* The graph-level rows pollute 60–92% of targets and report 23–37% spurious flips at the same budget; the explanation-level rows are spurious-free *by construction*, and at *equal operation space* the two placements differ by the entire utility (0.000 vs. $0.485/0.349$, Table 8). *Third, the certified league is small*: only PrivCF-SGD, PrivCFS, and the per-edge randomized-response variant (App. G) certify their complete release at the quoted budget, and only the explanation-level designs remain valid by construction. Size currencies differ across arms—edges only, edge-plus-feature, or a fraction of all edges removed—and deployments should state the currency explicitly.

Table 3: Same-population mechanism comparison on the recourse population (borderline targets; the general-population rows mirror Table 1, and the full baseline suite—PNS, RCExplainer, COMBINEX, NaiveDP, LapDP—is Table 7 in App. H). Every row shares the same targets, the same public prior (ρ=0.5), and clean-graph validation. CSR: valid counterfactual rate on the clean graph (Ours-EM: sampled rate at ε=8). Spur./Poll.: spurious-flip / prediction-pollution rates. Certified: does the quoted ε certify the complete released object? PrivCF-out: heuristic output perturbation—full-release budget 573–753/256 (App. A).

Dataset	Method	CSR↑	Spur.↓	Poll↓	ε	Certified?
Borderline (diagnostic); support ceiling: starCF of Table 2						
Cora	CF (no privacy)	1.000	0.000	0.000	∞	—
Cora	RR-LDP (graph-level)	0.633±.125	0.367±.125	0.600±.141	7.2	✓(graph)
Cora	DP-SGD backbone	0.967±.047	0.433±.047	0.767±.047	7.16	✓(model)
Cora	PrivCF-out (heur.)	1.000	0.000	0.000	7.2	×
Cora	PrivCF-SGD	0.500±.141	0.000	0.000	7.16	✓
Cora	**Ours-EM** (ρ=0.5/0.7)	0.484±.484/0.614±.467	0.000	0.000	8	✓**(pure)**
Borderline (diagnostic); support ceiling: starCF of Table 2						
CiteSeer	CF (no privacy)	1.000	0.000	0.000	∞	—
CiteSeer	RR-LDP (graph-level)	0.733±.047	0.267±.047	0.500±.082	7.2	✓(graph)
CiteSeer	DP-SGD backbone	1.000	0.400±.141	0.633±.094	7.16	✓(model)
CiteSeer	PrivCF-out (heur.)	1.000	0.000	0.000	7.2	×
CiteSeer	PrivCF-SGD	0.533±.125	0.000	0.000	7.16	✓
CiteSeer	**Ours-EM** (ρ=0.5/0.7)	0.348±.458/0.534±.468	0.000	0.000	8	✓**(pure)**

4.5 Audit of the Released Selection

We measure the *actual* leakage of the certified release with the strongest feasible auditor on real adjacent pairs (80 pairs per dataset, ε=8, ρ=0.5; optimal likelihood-ratio attack on the exact release distributions of Eq. (9); protocol and full analysis in App. E). Three findings (Fig. 3). *(i) The measured sensitivity respects the bound with headroom:* $\max|\Delta u| = 0.96 \leq 1.0$ on every pair, and the maximum release ratio (35.3) sits two orders of magnitude below the e^8 ceiling. *(ii) The release is piecewise constant:* the utility vector—and hence the entire release distribution—is identical on 91–96% of adjacent pairs. *(iii) The optimal attack gains almost nothing:* the likelihood-ratio AUC is 0.50 on average and 0.59 in the worst pair, far below the heuristic release's worst entry (1.0) under the same attack and protocol (App. A). The audit is supplementary evidence for Theorem 1, not a substitute for it. The mechanism also transfers to a 15K-node ogbn-arxiv core at full ceiling (0.963 valid at ε=8; Table 6), with per-query certification cost independent of N under a bounded-degree condition (App. F).

5 Conclusion

We have shown that the privacy–explainability conflict in graph counterfactual explanations is decided by *placement*, and that the explanation-level placement can be made *certifiable by construction*: selecting the intervention from a fixed, data-independent candidate universe with a validity-gated, clipped utility gives pure ε-DP for the complete released object, composition over queries, and a utility bound that requires no optimizer sensitivity analysis. Three quantitative facts anchor the result. *Privacy is the cheapest stage of the pipeline*: the sampled release retains 83–97% of its support-restricted optimum at ε=8, and the dominant cost is a *purchasable price list* in public disclosure strength. *The guarantee is borne out empirically*: the optimal likelihood-ratio edge-inference attack attains AUC 0.50 (max 0.59) on the certified release, against the heuristic baseline's worst entry of 1.0 under the same attack and protocol (App. A). *The placement gap is total at equal operation space*: graph-level DP retains 0.000 validity inside the identical support (Sec. 4.4). Population targeting from public information is free, while margin-based gating would require its own selection budget (Sec. 4.3). Extensions that inherit the same guarantee unchanged—thresholded "no intervention" releases, tighter model-Lipschitz bounds, feature-entry DP, two-hop public-prior universes, graph-level tasks with motif ground truth—are direct specializations of the support–utility–mechanism interface, not redesigns.

REPRODUCIBILITY STATEMENT

All components needed to reproduce our results are specified in the paper and its appendices: the candidate-universe construction with its boundary rules (Sec. 3.3, App. B), the utility and mechanism with their proofs (Secs. 3.4–3.5), the exact two-hop evaluation procedure (App. B), the backbone and training hyperparameters (Sec. 4.1), and the exhaustive small-graph certification checks (Sec. 4.2). All reported numbers are mean $\pm$ std over 3 seeds (pooled over targets) unless stated otherwise. An anonymized code bundle reproducing every table and figure is available at `https://anonymous.4open.science/r/wherePrivacyBelong-25F1/`.

ETHICS STATEMENT

This work studies how counterfactual explanations can be released without disclosing private graph structure; its goal is to *reduce* privacy harm in deployed explanation services. All experiments use publicly available benchmark datasets (Cora, CiteSeer, ogbn-arxiv); no human subjects and no newly collected private data are involved. The edge-inference audit of Sec. 4.5 is an evaluation instrument run exclusively on our own releases to measure their leakage, not an attack deployed against any third party. The mechanism requires a public snapshot of part of the graph structure, so deploying it involves a disclosure decision that must remain with the data owner under the consent of the affected parties; the price-list analysis of Sec. 4.3 is intended to make that trade-off explicit, not to encourage disclosure. We are not aware of other ethical concerns raised by this work.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure: the research ideas, theoretical models and conceptual frameworks, mathematical claims and their proofs, hypotheses, research methodology and experiment design, method implementation, data collection and processing, and interpretation of results are entirely the authors' own work. Additionally, we used generative AI tools to aid and polish the writing of this paper (a task with recommended disclosure). We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318.

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 5644–5655, 2021.

Ruichu Cai, Yuxuan Zhu, Xuexin Chen, Yuan Fang, Min Wu, Jie Qiao, and Zhifeng Hao. On the probability of necessity and sufficiency of explaining graph neural networks: A lower bound optimization approach. *Neural Networks*, 184:107065, 2025. doi: 10.1016/j.neunet.2024.107065. arXiv:2212.07056.

Ameya Daigavane, Gagan Madan, Aditya Sinha, Abhradeep Guha Thakurta, Gaurav Aggarwal, and Prateek Jain. Node-level differentially private graph neural networks. In *ICLR 2022 Workshop on PAIR^{2}Struct*, 2022. arXiv:2111.15521.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Flavio Giorgi, Fabrizio Silvestri, and Gabriele Tolomei. COMBINEX: A unified counterfactual explainer for graph neural networks via node feature and structural perturbations. *arXiv preprint arXiv:2502.10111*, 2025.

Zhimeng Guo, Zongyu Wu, Teng Xiao, Charu Aggarwal, Hui Liu, and Suhang Wang. Counterfactual learning on graphs: A survey. *Machine Intelligence Research*, 22(1):17–59, 2025. doi: 10.1007/s11633-024-1519-z.

Frederik Harder, Matthias Bauer, and Mijung Park. Interpretable and differentially private predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4083–4090, 2020.

Zexi Huang, Mert Kosan, Sourav Medya, Sayan Ranu, and Ambuj Singh. Global counterfactual explainer for graph neural networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 141–149, 2023. doi: 10.1145/3539597.3570376.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4069–4079, 2017.

Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511. PMLR, 2022.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 94–103, 2007.

Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Mohamed Nomeir, Pasan Dissanayake, Shreya Meel, Sanghamitra Dutta, and Sennur Ulukus. Private counterfactual retrieval. *arXiv preprint arXiv:2410.13812*, 2024.

Sina Sajadmanesh and Daniel Gatica-Perez. Locally private graph neural networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2130–2145, 2021. doi: 10.1145/3460120.3484565.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. Feature-based learning for diverse and privacy-preserving counterfactual explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2211–2222, 2023. doi: 10.1145/3580305.3599343.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018.

Fan Wu, Yunhui Long, Ce Zhang, and Bo Li. Linkteller: Recovering private edges from graph neural networks via influence analysis. In *43rd IEEE Symposium on Security and Privacy (S&P)*, pp. 2005–2024, 2022.

Fan Yang, Qizhang Feng, Kaixiong Zhou, Jiahao Chen, and Xia Hu. Differentially private counterfactuals via functional mechanism. *arXiv preprint arXiv:2208.02878*, 2022.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Ashkan Yousefpour, Igor Shilov, Alex Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5782–5799, 2023. doi: 10.1109/TPAMI.2022.3204236.

Qiuchen Zhang, Hong Kyu Lee, Jing Ma, Jian Lou, Carl Yang, and Li Xiong. DPAR: Decoupled graph neural networks with node-level differential privacy. In *Proceedings of the ACM Web Conference (WWW)*, pp. 1170–1181, 2024. doi: 10.1145/3589334.3645531.

Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Y. Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 682–692, 2021.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2847–2856, 2018. doi: 10.1145/3219819.3220078.

A The Certification Gap of Heuristic Output Perturbation

This appendix analyzes the accounting of the heuristic output-perturbation baseline: the mechanism perturbs trained mask logits with Gaussian noise calibrated to a *unit per-entry sensitivity* assumption ($\Delta_1{=}1$), advertising $\varepsilon{=}7.2$ per released entry. The actual release is the full logit vector, and its true budgets, measured on the deployed from-scratch pipeline ($1{,}500$ adjacent pairs, 3 seeds), are:

- **Full-release (vector) budget:** $\varepsilon \approx 753$ (Cora) / 256 (CiteSeer) at the operating noise—the effective full-release budget implied by the deployed mechanism. The adjacent-optima (warm-started) certificate is far smaller ($\Delta_\infty \leq 0.51$, vector $\varepsilon \approx 300$), but warm-start is not the deployed mechanism.
- **Basin jumps:** on $\approx$8–9$\%$ of adjacent pairs, from-scratch retraining lands in a different optimizer basin: the removed edge's own logit moves by up to $10.3/9.9$ (per-entry) and the full vector distance reaches $104.6/35.5$ (L_2)—$\sim$20$\times$ the unit assumption. Local Lipschitz regularization (gradient-norm and consistency penalties) does not close the gap, which is optimizer-basin-level, not curvature-level.
- **Average-safe, worst-case-leaky:** the omniscient likelihood-ratio attacker attains mean AUC 0.54—near chance, because 91–93$\%$ of adjacent pairs produce identical releases—but on the jump pairs the removed edge's own entry attains AUC 1.0 at $\varepsilon{=}7.2$ (0.93 at $\varepsilon{=}1.5$), and $\approx$8$\%$ of pairs exceed AUC 0.8. Average-case attack success cannot repair a worst-case guarantee, and DP *is* a worst-case guarantee.

The conclusion is not that explanation-level privacy is impossible—it is that *certification through the optimizer* is the wrong instrument. The selection layer of Secs. 3.3–3.5 replaces a sensitivity analysis that does not exist with one that is global by construction, and the audit of Sec. 4.5 shows the certified mechanism's worst pair (AUC 0.59) is far below the heuristic's worst entry (AUC 1.0), both measured by the same likelihood-ratio attack on the same adjacent-pair protocol. The three placements thus form a leakage–utility spectrum: the heuristic's worst entry leaks at AUC 1.0 with full (uncertified) utility; the certified selector's worst pair sits at AUC 0.59 with support-priced utility; and graph-level DP at the same operation space retains no utility at all (0.000, Sec. 4.4).

B Implementation Notes

Candidate evaluation uses the two-hop induced subgraph of the target with explicit degree renormalization $\hat{A}' = (A_{\text{sub}} + I \ominus S_E) \oslash \sqrt{d'_i d'_j}$, where d' is the full-graph degree minus deletion decrements; for a two-layer GCN this is exact for the target's output (verified against full-graph forward: max logit difference $2{\times}10^{-7}$ over 60 random candidates). Support sizes: $|C_v|$ ranges from 2,923 ($k_F{=}8$, $k_F^{\max}{=}2$) to 55,063 ($k_F{=}16$, $k_F^{\max}{=}3$), and every result reports the *actual per-node* support size rather than a fixed nominal value; on the borderline protocol the actual support is far smaller than the nominal bound (mean 877 Cora / 947 CiteSeer vs. nominal $23{,}621$ at $k_F{=}12$—the nominal bound binds only on high-degree nodes), and per-target evaluation takes seconds (2.1s / 4.1s) at the nominal

config, scaling with the actual support and the two-hop neighborhood size ($|N_2(v)|$ mean 33.3 / 11.2), consistent with the conditional complexity statement of App. F. Exact enumeration plus 100 release draws per cell runs in seconds per node on one RTX 4090. Candidate-construction boundary rules (Sec. 3.3): if fewer than k public-prior pairs exist, the universe is the actual set (k=0 yields the empty support, i.e., only $\varnothing$); deleting a nonexistent edge is a no-op; the plausibility denominator for an empty edge set is defined as 1.0 (no division by zero); zero-valued feature dimensions are excluded from the feature universe (masking them is a no-op); and pair/feature lists are deduplicated. Utilities carry runtime range checks (finite, no NaN, within the public range), and the exponential mechanism normalizes via log-sum-exp; any violation raises an error rather than silently releasing. The boundary unit tests of Sec. 4.2 run against these exact rules.

Algorithm 1 Private counterfactual selection (PRIVCFS)

Require: Frozen backbone f; public snapshot $\mathcal{G}_{\text{pub}}$; public features X; target v; budget ε; weights w_f, w_c, w_p; caps $k_E^{\max}, k_F^{\max}$

Ensure: Private candidate S ($\varnothing$ = no intervention sampled); utility oracle retained for repeated queries

1: Build U_v^E (Eq. (5)) and U_v^F (Eq. (7)); enumerate C_v (Eq. (4))
2: $\hat{y}_{\text{orig}} \leftarrow \arg\max_y f(\mathcal{G})[v]_y$
3: **for** each candidate $S = (S_E, S_F) \in C_v$ **do**
4: Apply S to $\mathcal{G}$ (no-op semantics, Eq. (6)); evaluate $u(\mathcal{G}, S)$ via Eq. (8)
5: **end for**
6: Sample $S^\star \sim P(S) \propto \exp\big(\varepsilon\, u(\mathcal{G}, S)/(2\Delta u)\big)$ ▷ exact exponential mechanism, Δu of Lem. 2
7: **return** $S^\star$ ▷ pure ε-DP; k queries compose to $k\varepsilon$

Table 4 summarizes the datasets and their roles in the evaluation (Sec. 4.1).

Table 4: Datasets and roles.

Dataset	#Nodes	#Edges	Feat.	#Classes	Acc.
Cora	2,708	5,278	1,433	7	0.809
CiteSeer	3,327	4,552	3,703	6	0.682
ogbn-arxiv-core	15,000	160,944	128	40	0.792

C CERTIFICATION CHECKS: FULL PROTOCOL

Before any result is read as certified, we establish that the guarantee of Sec. 3.5 holds in the implementation. The DP guarantee rests on Lem. 2 and the no-op support construction. Both are verified at the implementation level by exhaustive small-graph checks—a consistency test, not a proof (the proof is the lemma): 40 random graphs on 8 nodes, each paired against up to 40 neighboring graphs (every flippable edge), enumerate the full support and check (i) the support is identical on both graphs; (ii) $\max_S |u(\mathcal{G}, S) - u(\mathcal{G}', S)| \le \Delta u$; (iii) the empirical release ratio $\max_S P(\mathcal{M}(\mathcal{G})=S)/P(\mathcal{M}(\mathcal{G}')=S) \le e^\varepsilon$. At ε=4: measured maximum sensitivity $0.93 \le 1.0$ (validity-gated) and $0.70 \le 0.8$ (soft); maximum release ratio 4.75 (valid) and 9.22 (soft) against the bound $e^4 = 54.6$. Every check passes on all pairs. Two readings beyond the pass/fail: the measured sensitivities concentrate below the bound (0.93 vs. 1.0; 0.70 vs. 0.8), confirming the clipping bound captures the true sensitivity while leaving headroom; and the release ratios are $10\times$ below e^ε even in the worst of 3,200 pairs, showing the exponential mechanism converts the utility bound into release-level protection without an extra constant factor. The same three checks on *real* adjacent pairs (Sec. 4.5) reproduce the pattern at scale. We additionally stress the interface with three *boundary configurations* at ε=8 (App. B): a no-feasible-flip graph, graphs with exactly one valid candidate, and an empty support (only $\varnothing$). All pass the same checks—candidate-ID invariance, sensitivity within Δu (max 0.95), release ratios below e^8 (max 190.6), exact normalization—and the no-flip configuration matches the closed-form invalid-release probability of the probabilistic semantics (finding (iv) of Sec. 4.3): 97.2% measured at $|C_v|$=77.

D Loss Decomposition and Edge–Feature Contributions

Fig. 2 isolates the three prices a deployment pays, decomposing the valid rate into four stages: the free counterfactual explainer (CF-GNNExplainer, the non-private upper bound), the support restriction (the same explainer's problem restricted to the fixed candidate universe, ρ=1), the public-prior restriction (ρ<1), and the privacy noise (sampling at ε=8). Each stage transition is one loss channel.

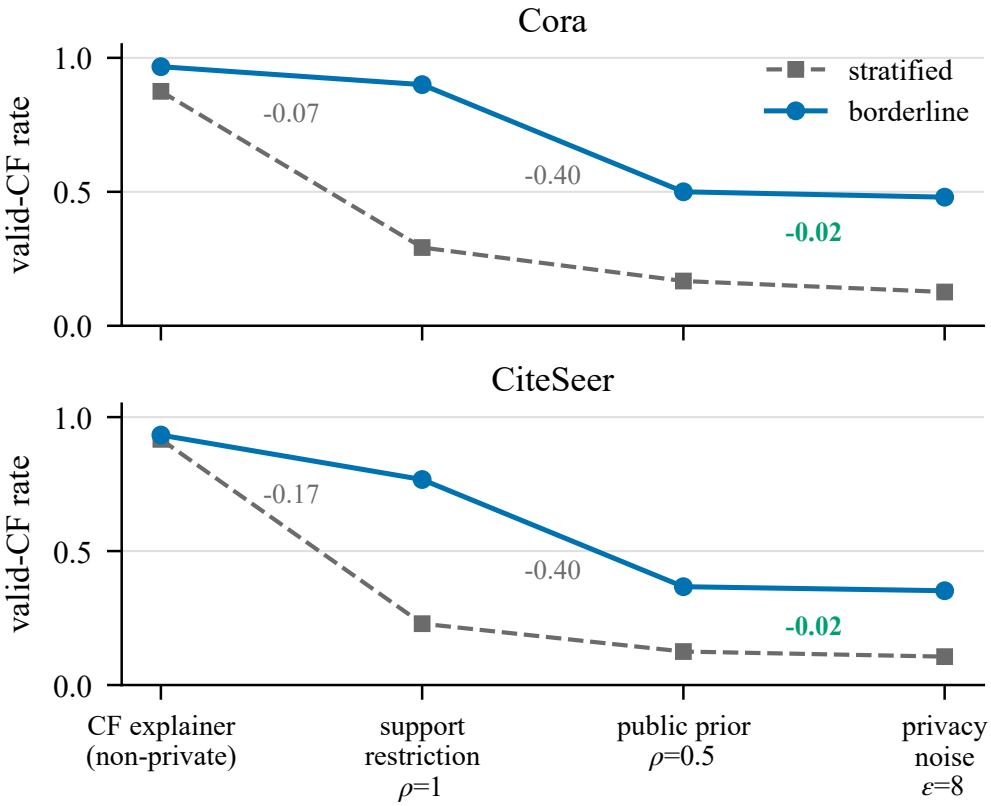

Figure 2: Loss decomposition on both protocols (3 seeds; borderline: n=30 targets, stratified: n=48). CF: non-private CF-GNNExplainer CSR (upper bound, Sec. 4.4); ρ=1: valid rate of the support-restricted non-private optimum; ρ=0.5: the public-prior support (certified under the fixed-snapshot premise); ε=8: the sampled release (soft utility, Δu=0.8; the validity-gated variant reproduces every number within $\pm$0.01). Annotated drops on the borderline chain mark each stage's price: the privacy price (-0.02) is the smallest, and the public prior is the dominant purchase.

The decomposition shows the order of the three prices: the *privacy* price is the smallest—the sampled release retains 76%/85% of the prior-restricted optimum at ε=8—while the larger gaps (-0.58/-0.69 to a fixed support, then -0.13/-0.10 to a public prior) are the *price of certifiability*: the difference between an unconstrained optimizer with full graph access and a fixed, public, enumerable support. This is the paper's answer to "where does private counterfactual quality actually go?": into the shape of the support, not into the noise the release adds—and the support's shape is exactly what the price list of Sec. 4.3 prices. The borderline chain shows the same mechanism in its target regime: on the recourse population the support restriction is nearly free (-0.07/-0.17), the public prior is the main purchase (-0.40/-0.40), and the privacy noise remains the cheapest element (-0.02/-0.02). The two protocols are the two regimes of one mechanism: *who* is explained selects which price dominates—on average nodes, the candidate budget; on borderline nodes, the public prior.

Table 5: Edge vs. feature contribution of the certified arm (borderline, ρ=0.5, ε=8, 3 seeds): starCF / P(flip)@ε=8. The feature-only arm is exactly zero—the certified arm's utility is carried by public-prior *edge* interventions, not by the (privacy-free) feature masking.

Arm	Cora	CiteSeer
edge-only (12/2, 0)	0.500 / 0.478±.478	0.367 / 0.341±.449
feature-only (0, 12/3)	0.000 / 0.000±0.000	0.000 / 0.000±0.000
joint (12/2, 12/3)	0.500 / 0.485±.486	0.367 / 0.349±.459

The certified arm's utility is carried by the public-prior *edge* interventions—edge-only is within noise of joint, and feature-only is exactly zero—so the released graph-structure contribution is substantive rather than an artifact of the (privacy-free) feature masking.

E AUDIT OF THE RELEASED SELECTION: FULL FINDINGS

This section measures the *actual* leakage of the certified release rather than its proxies, complementing Theorem 1 with the strongest feasible auditor on real adjacent pairs: for each of 80 pairs per dataset (deletions and additions within the target's two-hop neighborhood, ρ=0.5 support, ε=8), we evaluate the full utility vector on both graphs, the two exact release distributions $P_{\mathcal{G}}, P_{\mathcal{G}'}$ of Eq. (9), and the optimal likelihood-ratio attack—predict $\mathcal{G}$ iff $P_{\mathcal{G}}(S)/P_{\mathcal{G}'}(S) > 1$—with 400 Monte-Carlo draws per pair.

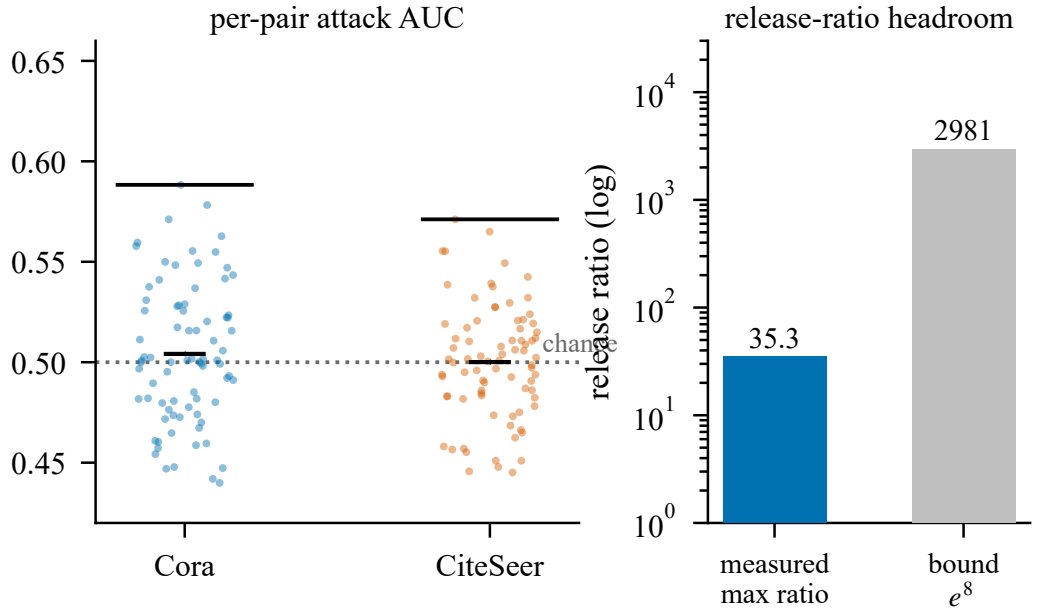

Figure 3: Audit of the certified release (80 real adjacent pairs per dataset, ε=8, Δu=1). Left: per-pair likelihood-ratio attack AUC (black ticks: mean and max). The release sits at chance on average and below 0.59 on every pair; measured max $|\Delta u|$=0.96 (bound 1.0), and the support-optimum changes on only 8.75%/3.75% of pairs (Cora/CiteSeer). Right: maximum release-probability ratio versus the e^8=2981 bound (log scale)—the measured worst case, 35.3, sits two orders of magnitude below the guarantee. The audit is supplementary evidence for Theorem 1, not a substitute for it.

Three findings (Fig. 3). *(i) The measured sensitivity respects the bound with room to spare:* $\max|\Delta u| = 0.96 \leq 1.0$ on every pair, and the maximum release ratio is 35.3—two orders of

magnitude below the $e^8 = 2981$ ceiling, i.e., the mechanism sits far from saturating its guarantee on real data. The gap is an empirical observation on the sampled pairs, *not* a re-certification: the mechanism's guarantee remains the proven ε=8 of Theorem 1, and the measured maximum is not an upper bound on all inputs—it indicates the headroom a tighter model-Lipschitz bound (Lem. 2, last remark) could convert into budget. *(ii) The release is piecewise constant:* on 91.25%/96.25% of adjacent pairs the utility vector—and hence the entire release distribution—is identical, and the argmax changes on only 8.75%/3.75%. The asymmetry between the two datasets is itself informative: CiteSeer's larger feature dimension dilutes the influence of any single edge, so fewer adjacent pairs reach the decision—support geometry, not mechanism tuning, sets the piecewise-constant fraction. *(iii) The optimal attack gains almost nothing:* the likelihood-ratio AUC is 0.50 on average and 0.59 in the worst pair. Both statistics use the same attack, the same adjacent-pair protocol, and the same likelihood-ratio statistic; contrasted like for like, the heuristic release of App. A attains mean 0.54 and worst-entry AUC 1.0: the certified mechanism's mean (0.50) is at chance, and its worst pair (0.59) is far below the heuristic's worst entry (1.0). The residual worst-pair AUC (0.59) sits on the argmax-changing pairs, and its size is exactly what the DP bound permits—the audit therefore measures the bound's tightness on real data, not a failure of the guarantee.

F Transfer and Scale: Full Analysis

The mechanism transfers to a 15K-node graph at full ceiling, with per-query certification cost independent of the node count *under a bounded-degree condition* (below). On a $15{,}000$-node degree-core subset of ogbn-arxiv with a two-layer GCN backbone (test accuracy 0.792), the borderline protocol (10 lowest-margin test nodes; Table 6) gives the support-restricted ceiling 1.000 at both public-prior strengths, and the sampled valid rate reaches 0.963 ± 0.026 at ε=8 on the *certified* arm (ρ=0.5). Three scale-specific readings follow.

Table 6: Transfer to ogbn-arxiv-core ($15{,}000$ nodes, $160{,}944$ edges, 40 classes): borderline protocol, 10 lowest-margin test nodes, mean $\pm$ std over targets (one seed; margins $\approx 8\times10^{-4}$—the GNN is near-tied on all targets). starCF: support-restricted non-private ceiling; P(flip): sampled valid rate; sampled releases average 4.24 ± 0.22 / 4.31 ± 0.13 budget elements at ε=8.

Dataset	ρ	starCF	@2	@4	@8
arxiv-core	0.5	$1.000\pm.000$	$0.734\pm.147$	$0.838\pm.110$	$0.963\pm.026$
arxiv-core	$1.0^\dagger$	$1.000\pm.000$	$0.704\pm.143$	$0.831\pm.107$	$0.957\pm.056$

First, the mechanism transfers at full ceiling: every one of the ten near-tied targets has a valid flip inside the candidate support, on both the public-prior arm (ρ=0.5, certified under the fixed-snapshot premise) and the upper reference—the data-independence restriction costs nothing on this population, because the public snapshot retains the (few) edges that matter for each ultra-borderline node. *Second, certification cost does not grow with the node count when the neighborhood is bounded:* the support is local to the target, and utility evaluation is exact on the two-hop induced subgraph regardless of N, so the per-query computation is $O(|C_v| \cdot |N_2(v)|^2)$—it grows with the candidate count, the two-hop neighborhood, and the backbone's forward cost, and it is independent of N *only if the degree (hence* $|N_2(v)|$*) is bounded*; no per-scale re-certification is needed. *Third, the concentration strengthens at scale:* the release-collision probability drops to 0.02–0.03% (vs. 0.2–0.3% on Cora/CiteSeer) because the 128-dimensional features make flipping candidates rarer and more decisive, so the exponential mechanism concentrates sharply once $\varepsilon \geq 4$—the sampled rate moves from 0.73 at ε=2 to 0.96 at ε=8 while the ceiling is 1.0 throughout. The same reading applies to the ε=2 dip (0.734 ± 0.147): at small budgets the many *near*-flipping candidates share probability mass, and the drop is exactly the regret term of Theorem 2 realized empirically. The transfer evidence is a scale probe over ten near-tied targets (one seed); uniform large-graph performance additionally depends on the bounded-neighborhood condition.

G Ablations

The ablations rule out the alternative explanations for the main results—staged sampling, utility tuning, the validity gate, data-dependent support, and the choice of public prior. *(a) Joint vs.*

sequential sampling (E10.4): a sequential two-step release—first edge by $\varepsilon/2$-budget EM over single-edge candidates, second edge by $\varepsilon/2$-budget EM conditioned on the first, budgets composed—reaches valid rate 0.688 ± 0.451 vs. 0.679 ± 0.445 for joint selection (Cora) and 0.274 ± 0.424 vs. 0.283 ± 0.432 (CiteSeer) at ε=8, ρ=0.5, borderline, one seed: statistically indistinguishable. The practical consequence is twofold: the simpler joint release loses nothing to the staged design, and staged search—if ever needed for larger supports—can be added without re-deriving the guarantee, because each stage composes exactly as Cor. 1 prescribes. *(b) Utility weights* (E10.5): on the Cora borderline protocol (n=10, seed 0), $(0.9, 0.05, 0.05)$ moves the ceiling to 0.700 ± 0.458 and the sampled rate to 0.679 ± 0.445 at ε=8 (ρ=0.5), $(0.5, 0.4, 0.1)$ gives 0.700 ± 0.458 and 0.669 ± 0.438—both within noise of the main configuration's behavior, and neither reverses any ordering. The mechanism is therefore not a delicate tuning artifact: the validity gate dominates the utility as long as w_f dominates w_c, which is the entire point of the gating design. *(c) Validity-gated vs. soft utility*: the soft variant (partial credit for margin reduction, Δu=0.8) reproduces the main borderline table within ±0.01 (e.g., Cora ρ=0.7: soft 0.602 ± 0.459 vs. gated 0.614 ± 0.467 at ε=8), and is the cheaper-sensitivity alternative where the $\varnothing$ semantics of Eq. (8) is not needed. *(d) Data-dependent support* (E10.1): ρ=1.0 rows throughout are the upper reference; the gap to the public-prior arms is the purchasable price of data independence, priced per dataset in Table 2 (0.13–0.40 of ceiling on borderline nodes). *(e) Public-prior necessity*: a purely feature-based universe (feature-kNN pairs, no public graph) overlaps the true neighborhood of citation-network targets by $\approx$0 pairs—data-independent universes must carry structural information; the ρ-snapshot is one concrete carrier, and other public priors (public metadata graphs, public link-prediction models) slot into the same interface unchanged.

Two further ablations sharpen the audit's interpretation rather than the mechanism's tuning. *(f) Budget-pair composition* (App. A material): the heuristic's 8% basin-jump pairs are precisely where its per-entry budget collapses; the certified mechanism has no such pairs by construction, because its sensitivity bound is global rather than local. *(g) Population sweep* (Table 2 and Fig. 5 read jointly): the margin distribution of the explained population (0.008 on borderline vs. 0.69–0.75 on random targets) is a one-number predictor of the mechanism's ceiling—borderline targets are flippable within a small budget, average targets are not. This is exactly why deployments must choose *who* is explained from public information (a fixed list or public attributes, at no privacy cost): margin-based gating would disclose private margin information and requires a separate selection budget. *(h) Joint vs. per-edge budget* (E5 design axis): an independent per-edge exponential mechanism—the randomized-response-style endpoint of the same placement, with budget $\varepsilon/|U_v^E|$ per edge—releases edges only and reaches 0.654 ± 0.434 (Cora, ρ=0.5, ε=8) against 0.484 ± 0.484 for joint selection, with the ordering reversing on CiteSeer at ρ=1.0 (0.620 ± 0.391 vs. 0.738 ± 0.408). The two arms are the two ends of one design axis—independent per-element budgets versus one joint set-level budget—and the joint arm additionally composes feature and edge candidates in a single draw with set-level minimality; we adopt the joint arm as the primary mechanism. *(i) Public support shrink* (B5): shrinking the public universe from k_E=12/k_F=12 down to 4/4 (candidate count $\frac{1}{27}\times$) leaves the borderline frontier unchanged—starCF $0.500/0.367$ throughout and P(flip)@ε=8 within 0.01—because the public influence ranking concentrates the flipping candidates in its top few entries; the empty-output rate rises (0.4% to 8.6–10.2%) as the $\varnothing$ candidate's relative mass grows, consistent with the probabilistic semantics of Sec. 4.3(iv). The two-hop universe extension raises the ceiling to $0.567/0.467$ and P(flip)@ε=8 to $0.527/0.440$ at $\sim4\times$ candidate count—both are public, data-independent rules, never private screening. *(j) Same-operation-space graph-level control* (B5): with the *identical* candidate support, privatizing the graph first (Gaussian, ε=7.2) collapses the support-restricted optimum to CSR 0.000 on both datasets (pollution $0.73/0.70$: no intervention in the support flips the privatized graph), against the clean same-space ceiling $0.500/0.367$ and our $0.485/0.349$ at ε=8. The placement diagnosis therefore holds at equal operation space: the two placements differ by the *entire* utility, not by a constant factor. *(k) Edge vs. feature contribution* (B4): on the borderline protocol the certified arm restricted to edges only reaches $0.478/0.341$—within noise of the joint arm ($0.485/0.349$)—while the feature-only arm is $0.000/0.000$: the mechanism's utility comes from public-prior *edge* interventions, not from the (privacy-free) feature masking, which substantiates the graph-structure contribution rather than limiting it.

H ADDITIONAL TABLES AND FIGURES

This appendix collects the tables and figures referenced from the main text that were omitted for space: the privacy–utility frontier curves (Fig. 4) underlying Tables 1 and 2, the population-gating view (Fig. 5) underlying finding (iii) of Sec. 4.3, and the full mechanism-comparison suite (Table 7) with the same-operation-space control (Table 8) supporting Sec. 4.4.

Figure 4 plots the frontier curves behind Tables 1 and 2 (Sec. 4.3).

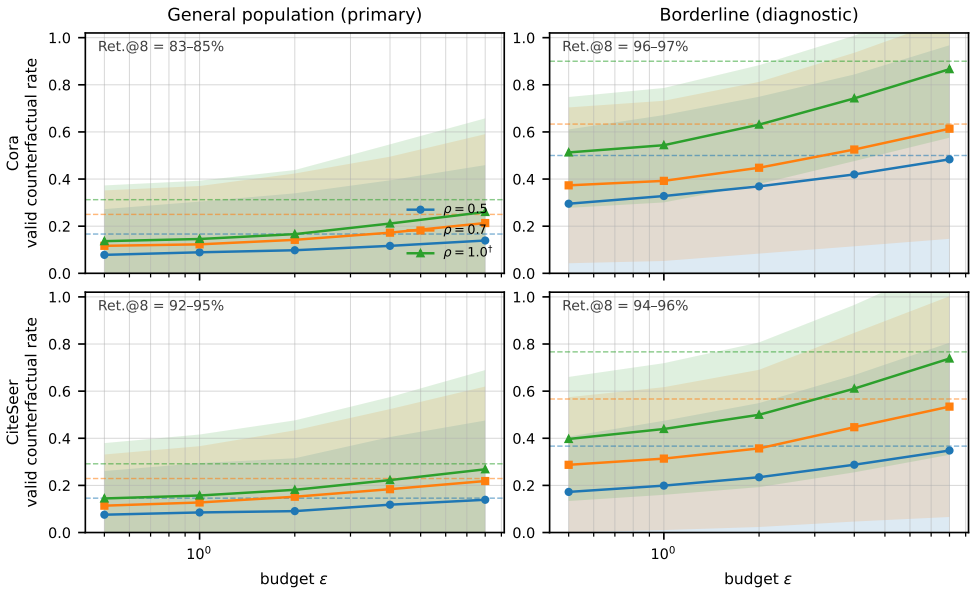

Figure 4: Privacy–utility frontier (3 seeds). Left: *general* population (primary, 16 publicly drawn targets per seed); right: *borderline* targets (diagnostic of the recourse population, 10 lowest-margin nodes per seed). Solid curves: valid counterfactual rate $P(\text{flip})$ of the sampled release versus budget ε (log scale), per public-prior strength ρ; shaded bands: ± 1 std; dashed lines: starCF, the support-restricted non-private ceiling; †: data-dependent upper reference (Sec. 4.1). Ret.@8: retention of the ceiling at ε=8. Privacy noise costs little at $\varepsilon \geq 4$, while the ceiling itself is set by the public prior—the binding constraint (Tables 1 and 2).

Table 7 is the full baseline suite of Sec. 4.4, of which Table 3 in the main text is the condensed recourse-population version.

Table 7: Same-population mechanism comparison: every row shares the same target nodes, the same public prior (ρ=0.5), and the same clean-graph validation. *General* panel: the primary table, 16 fixed publicly drawn targets × 3 seeds. *Borderline* panel: diagnostic, the 10 lowest-margin targets × 3 seeds. CSR: valid counterfactual rate on the clean graph (for Ours-EM, the sampled rate at ε=8; non-empty invalid releases count as failures, and the empty-output rate is 0.1–0.4%). starCF: the support-restricted non-private ceiling—the same-operation-space reference. Spur./Poll: spurious-flip and prediction-pollution rates (diagnostic). Certified: does the quoted ε certify the complete released object? ε=7.2 rows: graph-level Gaussian/Laplace/RR releases of the adjacency; ε=7.16: Opacus RDP, T=100, σ=7. PrivCF-out: heuristic output perturbation—full-release budget 573–753/256 (App. A), listed as non-certified. Ours-EM: pure ε-DP (certified arm, Sec. 4.1). Operation spaces differ (continuous masks vs. ≤ 2 edges + ≤ 3 features); starCF is the same-operation-space non-private anchor.

Dataset	Method	CSR↑	Spur.↓	Poll↓	ε	Certified?
General population (primary)						
Cora	starCF (support ceiling)	0.167±.373	0.000	0.000	∞	—
Cora	CF (no privacy)	1.000	0.000	0.000	∞	—
Cora	PNS (causal)	0.958±.029	0.000	0.000	∞	—
Cora	RCExplainer (rob.)	0.854±.029	0.000	0.000	∞	—
Cora	COMBINEX (unif.)	0.646±.059	0.000	0.000	∞	—
Cora	NaiveDP (graph-level)	0.625±.051	0.333±.106	0.812±.051	7.2	✓(graph)
Cora	LapDP (graph-level)	0.729±.078	0.250±.102	0.812±.051	7.2	✓(graph)
Cora	RR-LDP (graph-level)	0.979±.029	0.021±.029	0.062±.051	7.2	✓(graph)
Cora	DP-SGD backbone	0.938±.088	0.750±.000	0.417±.078	7.16	✓(model)
Cora	PrivCF-out (heur.)	1.000	0.000	0.000	7.2	×
Cora	PrivCF-SGD	0.125±.051	0.000	0.000	7.16	✓
Cora	**Ours-EM** (ρ=0.5)	0.139±.320	0.000	0.000	8	**✓(pure)**
Borderline (diagnostic)						
Cora	starCF (support ceiling)	0.500±.500	0.000	0.000	∞	—
Cora	CF (no privacy)	1.000	0.000	0.000	∞	—
Cora	PNS (causal)	1.000	0.000	0.000	∞	—
Cora	RCExplainer (rob.)	1.000	0.000	0.000	∞	—
Cora	COMBINEX (unif.)	0.800±.082	0.000	0.000	∞	—
Cora	NaiveDP (graph-level)	0.667±.205	0.333±.125	0.733±.047	7.2	✓(graph)
Cora	LapDP (graph-level)	0.700±.082	0.300±.141	0.733±.047	7.2	✓(graph)
Cora	RR-LDP (graph-level)	0.633±.125	0.367±.125	0.600±.141	7.2	✓(graph)
Cora	DP-SGD backbone	0.967±.047	0.433±.047	0.767±.047	7.16	✓(model)
Cora	PrivCF-out (heur.)	1.000	0.000	0.000	7.2	×
Cora	PrivCF-SGD	0.500±.141	0.000	0.000	7.16	✓
Cora	**Ours-EM** (ρ=0.5/0.7)	0.484±.484/0.614±.467	0.000	0.000	8	**✓(pure)**
General population (primary)						
CiteSeer	starCF (support ceiling)	0.146±.353	0.000	0.000	∞	—
CiteSeer	CF (no privacy)	0.958±.029	0.000	0.000	∞	—
CiteSeer	PNS (causal)	0.958±.029	0.000	0.000	∞	—
CiteSeer	RCExplainer (rob.)	0.854±.059	0.000	0.000	∞	—
CiteSeer	COMBINEX (unif.)	0.292±.128	0.000	0.000	∞	—
CiteSeer	NaiveDP (graph-level)	0.688±.088	0.312±.051	0.917±.030	7.2	✓(graph)
CiteSeer	LapDP (graph-level)	0.833±.059	0.167±.059	0.917±.030	7.2	✓(graph)
CiteSeer	RR-LDP (graph-level)	0.875±.088	0.104±.030	0.188±.051	7.2	✓(graph)
CiteSeer	DP-SGD backbone	1.000	0.771±.030	0.458±.030	7.16	✓(model)
CiteSeer	PrivCF-out (heur.)	0.958±.029	0.000	0.000	7.2	×
CiteSeer	PrivCF-SGD	0.146±.029	0.000	0.000	7.16	✓
CiteSeer	**Ours-EM** (ρ=0.5)	0.139±.336	0.000	0.000	8	**✓(pure)**
Borderline (diagnostic)						
CiteSeer	starCF (support ceiling)	0.367±.482	0.000	0.000	∞	—
CiteSeer	CF (no privacy)	1.000	0.000	0.000	∞	—
CiteSeer	PNS (causal)	1.000	0.000	0.000	∞	—
CiteSeer	RCExplainer (rob.)	1.000	0.000	0.000	∞	—
CiteSeer	COMBINEX (unif.)	0.433±.125	0.000	0.000	∞	—
CiteSeer	NaiveDP (graph-level)	0.733±.170	0.267±.125	0.700±.082	7.2	✓(graph)
CiteSeer	LapDP (graph-level)	0.767±.205	0.233±.047	0.700±.082	7.2	✓(graph)
CiteSeer	RR-LDP (graph-level)	0.733±.047	0.267±.047	0.500±.082	7.2	✓(graph)
CiteSeer	DP-SGD backbone	1.000	0.400±.141	0.633±.094	7.16	✓(model)
CiteSeer	PrivCF-out (heur.)	1.000	0.000	0.000	7.2	×
CiteSeer	PrivCF-SGD	0.533±.125	0.000	0.000	7.16	✓
CiteSeer	**Ours-EM** (ρ=0.5/0.7)	0.348±.458/0.534±.468	0.000	0.000	8	**✓(pure)**

Table 8 reports the same-operation-space placement control quoted in Sec. 4.4.

Table 8: Same-operation-space placement control (borderline protocol, ρ=0.5, 3 seeds, n=30 targets). Every row uses the *identical* candidate support; validity columns are clean-graph CSR / P(flip). starCF: support-restricted non-private optimum on the clean graph. restricted-NaiveDP: privatize the graph first (Gaussian, ε=7.2), take the support-restricted argmax on the privatized graph, then verify it on the clean graph (protocol in Ablation (j)). Ours-EM: sampled release at ε=8.

Dataset	starCF	restr.-NaiveDP	Poll.	Ours-EM@8
Cora	0.500±.500	**0.000±0.000**	0.733	0.485±.486
CiteSeer	0.367±.482	**0.000±0.000**	0.700	0.349±.459

Figure 5 gives the deployment-guidance view behind finding (iii) of Sec. 4.3.

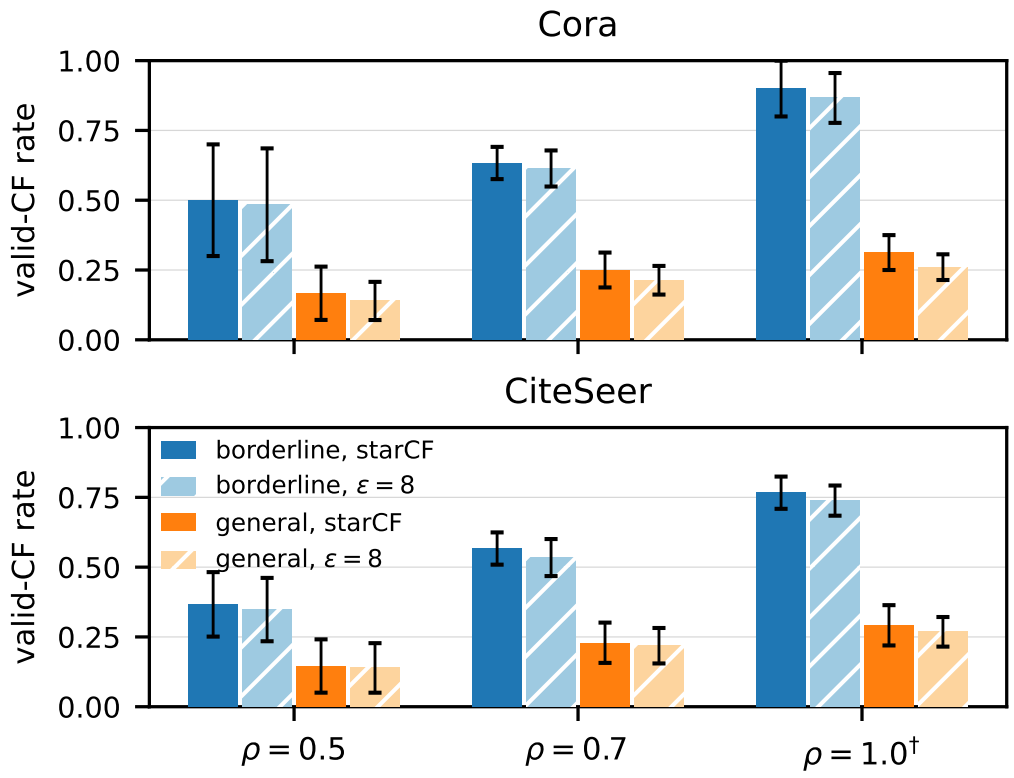

Figure 5: Population gating (deployment guidance). Solid bars: starCF, the support-restricted non-private ceiling; hatched bars: valid-CF rate of the sampled release at ε=8; whiskers: ±1 std across 3 seeds; †: data-dependent upper reference (Sec. 4.1). The ceiling is a one-number function of the target's margin (borderline ≈0.008 vs. general 0.69–0.75): the owner may choose *who* to explain from public information at no privacy cost, while margin-based gating would itself disclose private margin information and is reported only as a diagnostic.