BEYOND OUTCOME FAIRNESS: MECHANISM-QUALIFIED INTERVENTIONS FOR GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Lower group disparity does not tell us which graph edges should change, or when uncertain edits should be withheld. We study these decisions through identify–qualify–expose. MD-SCM parameterizes edge propensity by task similarity, sensitive homophily, and their interaction; a joint IPW-MLE consistently estimates the declared contributions under explicit model and sampling assumptions. MIC (Mechanism Intervention Certificate) transfers compatible-parameter uncertainty into simultaneous soft-adjacency prediction-change bounds for a fixed intervention and withholds unqualified actions. MIFB matches aggregate graph statistics while separating construction-oracle actions, yielding a lower bound for statistics-restricted Lipschitz policies and an action-fidelity audit for general editors. Across four datasets, group disparity falls 15.3–38.5% versus FairSNR, with a 0.11–0.34 percentage-point accuracy cost. At equal editing budgets, disparity decreases from 1.57 to 1.01 versus random abstention, supporting selective intervention under matched evaluation. Under misspecification, target-action error falls 43.4%; MD-SCM-free audits preserve four-dataset ordering. These results support selection quality beyond model matching; certification concerns fixed-mask prediction stability, excluding subsequent edge suppression and fairness improvement. The protocol complements outcome fairness with separately testable claims about which mechanisms are edited, when edits qualify, and whether actions match their targets.

1 INTRODUCTION

Graph neural networks (GNNs) aggregate local neighborhoods and therefore can propagate sensitive-attribute homophily through the adjacency Kipf & Welling (2017); Dong et al. (2022). Yet the same group gap can accompany different editing needs: preserve task-related connections or replace sensitive-driven ones. A useful intervention must therefore specify what changes, how uncertainty limits the action, and whether the selected edits match that purpose.

Recent fair graph learning has moved toward distribution-shift robustness, multi-attribute adaptation, and structure-aware encoding. DANCE and Fair Graph U-Net improve representation fairness under shifted or pooled graphs Wang et al. (2025b;c); FairSNR and homophily-aware counterfactual learning refine causal or contrastive interventions Liu et al. (2026); Kejani et al. (2026); FairGuide and FairGE modify or encode graph structure Lu et al. (2025); Luo et al. (2026). These advances motivate a complementary question: can an edge intervention be justified and audited beyond its outcome gap? We connect declared task, sensitive, and mixed propensities to uncertainty-qualified replacement, then test action fidelity on matched-statistics graphs. Figure 1 illustrates why matched visible statistics can demand different edits, motivating our identify–qualify–expose workflow under explicit modeling assumptions.

Three contributions form this protocol. (1) **Identify:** MD-SCM separates task, sensitive-homophily, and mixed edge-propensity contributions under an explicit, model-conditional identification statement; synthetic parameter recovery and intervention precision test this claim. (2) **Qualify:** MIC converts compatible mechanism parameters into simultaneous soft-adjacency prediction-change bounds and abstains when the bound is ambiguous or non-actionable; calibration, bound magnitude, and failure-regime abstention diagnose its usefulness. (3) **Expose:** MIFB matches pre-

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.

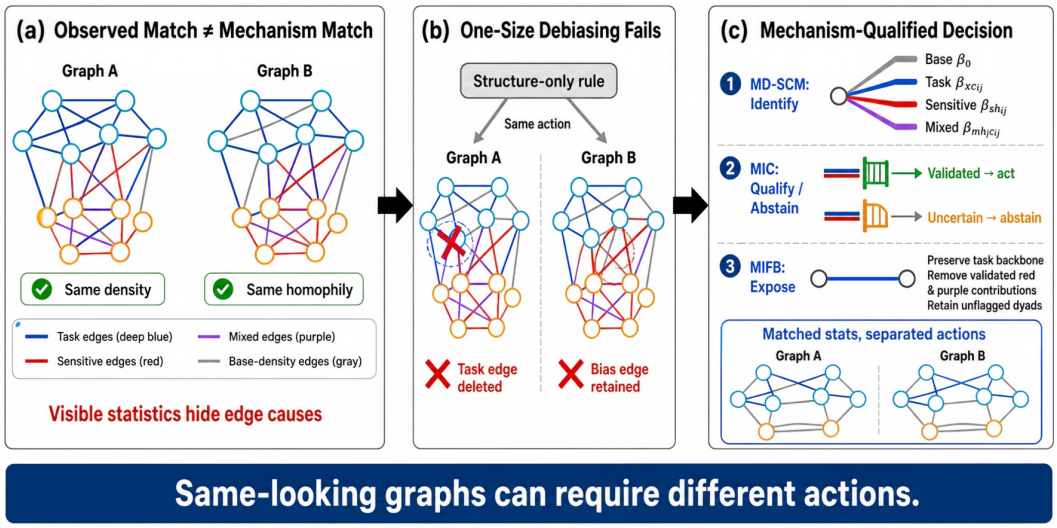


Figure 1: Why mechanism qualification matters. Graphs with matched visible statistics can contain different task, sensitive, and mixed edge contributions, so a shared structure-only action may delete task signal or retain bias. MD-SCM identifies the declared contributions, MIC acts or abstains, and MIFB tests whether matched graphs require separated actions.

defined aggregate graph statistics while separating construction-oracle actions; its theorem covers only $\mathcal{T}$-restricted Lipschitz policies, while full GNNs are evaluated empirically by action fidelity.

The protocol makes three intervention decisions explicit: which declared edge contributions to replace, when to accept the replacement, and how to assess its target error. Four-dataset outcomes test the fairness–utility trade-off; matched-budget and shifted-generator controls test selection and action fidelity. This adds a structural-action audit alongside representation-level fairness objectives.

2 Related Work

Classical counterfactual fairness formalizes protected-attribute interventions Kusner et al. (2017); our intervention instead targets edge propensities.

Fair graph learning uses representation control Dai & Wang (2021); Wang et al. (2022); Liu et al. (2025); Wang et al. (2025b;c); Ye et al. (2026); Luo et al. (2026); Cao et al. (2026), structural editing Dong et al. (2022); Li et al. (2024); Cheng et al. (2024); Zhang et al. (2024a); Lu et al. (2025), disentanglement Lee et al. (2025), invariant learning Zhu et al. (2024), and robustness certification Bojchevski et al. (2020); Zügner & Günnemann (2020); Dong et al. (2026). Counterfactual graph methods Agarwal et al. (2021); Ma et al. (2022); Guo et al. (2023); Wang et al. (2025a); Kejani et al. (2026); Wang et al. (2025d) intervene on attributes, edges, or explanations. These objectives are adjacent to, but not identical to, an edge-generation mechanism certificate.

Closest work and exact boundary. SFG Chen et al. (2025) combines Lipschitz-constrained training with distributionally robust node selection. MIC qualifies a fixed edge intervention after training over compatible mechanism parameters; neither Lipschitz control nor uncertainty-based selection is new by itself. GraphPAR Zhang et al. (2024b) certifies bounded sensitive-semantic perturbations and ELEGANT Dong et al. (2026) adversarial fairness perturbations; our certificate concerns a different intervention domain, not a stronger guarantee. FairSNR Liu et al. (2026) uses probability of necessity and sufficiency to guide representation-level counterfactual interventions. MD-SCM instead models edge-generation propensity; MIC qualifies structural replacement through compatible-set bounds.

CDFG disentangles causal and sensitive representations Zhang et al. (2025); matched comparisons cover German and Credit. Network SCMs estimate interventions for non-IID graphs Dai et al.

(2026), rather than our edge contributions. Relational SCMs identify queries across relational skeletons Ejaz & Bareinboim (2026); our identification is restricted to one declared dyad equation. FairMIB uses IPW for diffusion feature correction Liu et al. (2025); ours weights sampled dyads for propensity estimation. FairWire regularizes dyadic bias; homophily-aware counterfactuals alter augmentation ratios Kose & Shen (2024); Kejani et al. (2026). Marey et al. vary controlled structural biases beyond homophily for link-prediction evaluation; MIFB instead matches declared statistics and tests whether mechanism-separated node-classification pairs require different actions Marey et al. (2026). SF-GNN filters homophilic edges amplified by hubs or triadic closure; FairGDiff constructs sensitive-homophily treatments for fair graph diffusion Lu et al. (2026); Wang et al. (2026). Our evaluation instead links declared edge contributions, compatible-set qualification, and target-action fidelity on matched pairs.

The individual ingredients are classical: the two-order average below is the two-player Shapley value Shapley (1953), dyad logistic interactions are standard propensity parameterizations, and compatible-set maximization is a robust-certification pattern. The contribution is linking a declared intervention to a qualification rule and a target-action audit; the component inequalities retain their established scope. Supplementary Table 9 compares the method scopes.

3 PRELIMINARIES

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{A})$ be an undirected graph. Each node v_i has features X_i, binary sensitive attribute S_i, and binary label Y_i. We use $\Delta_{\mathrm{SP}} = |P(\hat{Y} = 1 \mid S = 1) - P(\hat{Y} = 1 \mid S = 0)|$, equal-opportunity gap, accuracy, and individual counterfactual disparity. For predicted probabilities $p_{\Theta,\phi}$ and factual/counterfactual adjacencies $\mathbf{A}, \mathbf{A}'$, define

$$\mathrm{CF}_i = |p_{\Theta,\phi}(\mathbf{A}, \mathbf{X})_i - p_{\Theta,\phi}(\mathbf{A}', \mathbf{X})_i|, \qquad \text{W5-CD} = \mathrm{mean}_{i \in \mathrm{top}_{5\%}(\mathrm{CF})}\, \mathrm{CF}_i. \tag{1}$$

Definition 1 (Mechanism-Disentangled Causal Parameterization)**.** *For a candidate dyad (i, j), let $X^{(-S,-Y)}$ denote features with sensitive attributes, labels, and direct copies of either removed, let $c_{ij} = \cos(X_i^{(-S,-Y)}, X_j^{(-S,-Y)})$, and let $h_{ij} = \mathbb{1}[S_i = S_j]$. We posit the structural equation*

$$A_{ij} = \mathbb{1}\{U_{ij} \le \beta_0 + \beta_x c_{ij} + \beta_s h_{ij} + \beta_m h_{ij} c_{ij}\}, \tag{2}$$

where $U_{ij} \sim \mathrm{Logistic}(0, 1)$, β_0 is a nuisance base-density offset, and the conditional probability is induced by this equation. Equation (2) *separates the declared edge mechanisms. The task, sensitive, and mixed terms are $\beta_x c_{ij}$, $\beta_s h_{ij}$, and $\beta_m h_{ij} c_{ij}$. We call setting $(\beta_s, \beta_m) = (0, 0)$ on validated dyads a* declared-mechanism replacement intervention*; it preserves the component designated as task-related by the declared model, not necessarily all real-world task information. Proxy information may remain in $X^{(-S,-Y)}$.*

Definition 1 formalizes the decomposition illustrated in Figure 1, making its contrasting editing needs explicit within the declared model.

4 METHODOLOGY

Figure 2 gives a compact procedure: (i) fit contributions, (ii) fix edits, (iii) train and freeze, and (iv) certify or abstain. MIFB separately audits editor actions.

Each stage can fail separately: inaccurate parameters, unresolved bounds, or target-inaccurate edits. The diagnostics follow this order, making the full pipeline open to falsification.

4.1 MD-SCM: MECHANISM IDENTIFICATION

We use a declared candidate-dyad frame $\mathcal{R}$ and a probability sample $\mathcal{D}$ with inclusion probabilities π_{ij} and weights $w_{ij} = 1/\pi_{ij}$. Before stating the result, we make the assumptions explicit: (A1) the dyad structural equation in Definition 1 is correctly specified; (A2) within each edge-status stratum, the declared dyad sample follows simple random sampling without replacement with known positive first-order inclusion probabilities and satisfies a design LLN/CLT; (A3) dyads are conditionally independent; and (A4) $\boldsymbol{\beta}^* \in \mathrm{int}(\mathcal{B}_\beta)$ for a compact parameter space $\mathcal{B}_\beta$, the population weighted

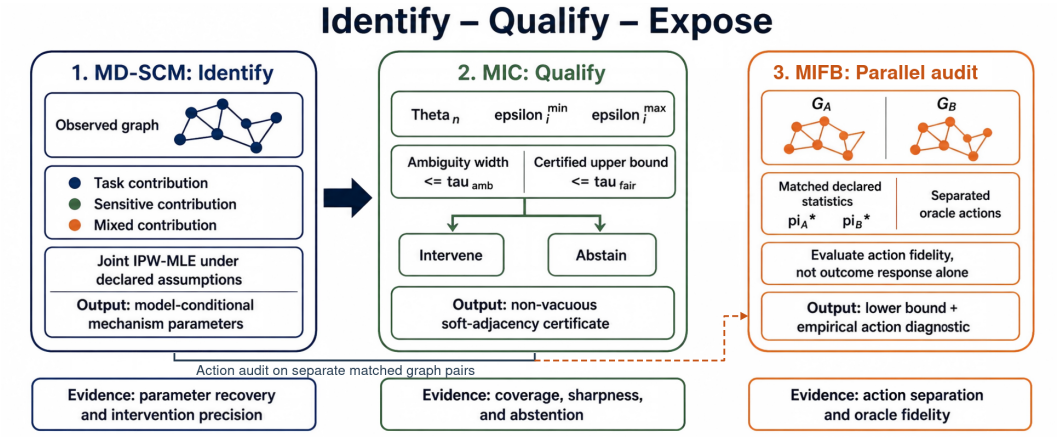


Figure 2: Identify–Qualify–Expose. MD-SCM and MIC define the intervention pipeline; MIFB is a parallel evaluation benchmark, not a deployment stage. Solid: intervention pipeline. Dashed: editor-action audit on separately constructed matched graph pairs.

objective is uniquely maximized at $\boldsymbol{\beta}^*$, the normalized weighted Fisher information has eigenvalues bounded away from zero, and $\max_e w_e / \sqrt{\sum_e w_e^2} \to 0$. These assumptions are about the declared model, not a claim that the model recovers unique real-world event causes. A cluster-dependent extension is outside this theorem and requires an increasing-cluster asymptotic analysis. The observable diagnostics are weighted distance correlation between c and h, conditional mixed-term information, cross-group edge count, and weight concentration.

Theorem 1 (Model-Conditional Identification of MD-SCM Contributions)**.** *Under (A1)–(A4), define the joint IPW-MLE by*

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathcal{B}_\beta} \sum_{(i,j) \in \mathcal{D}} w_{ij} \ell(A_{ij}, \beta_0 + \beta_x c_{ij} + \beta_s h_{ij} + \beta_m h_{ij} c_{ij}) \tag{3}$$

Then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{\sum_e w_e^2} / \sum_e w_e)$. Thus, the coefficients and induced propensity contributions are consistently estimated within the declared dyad model; no event-level causal attribution is implied.

Identification requires both homophily strata and within-stratum similarity variation. Low $\mathrm{dCor}(c, h)$ cannot exclude proxies or latent confounding; passing diagnostics retains model-conditional MIC gating. Failure regimes are CE1 ($\mathrm{dCor}(c, h) > 0.8$), CE2 ($|\mathcal{E}_{\mathrm{cross}}| < 100$), and CE3 ($|\beta_m / \beta_x| > 1.5$).

For $t = \eta_t = \hat{\beta}_0 + \hat{\beta}_x c_{ij}$, $s = \eta_s = \hat{\beta}_s h_{ij}$, and $m = \eta_m = \hat{\beta}_m h_{ij} c_{ij}$, Eqs. (4a)–(4b) average both inclusion orders:

$$\psi_{ij}^{\mathrm{sens}} = \tfrac{1}{2}[\sigma(t+s) - \sigma(t) + \sigma(t+s+m) - \sigma(t+m)], \tag{4a}$$

$$\psi_{ij}^{\mathrm{mix}} = \tfrac{1}{2}[\sigma(t+m) - \sigma(t) + \sigma(t+s+m) - \sigma(t+s)]. \tag{4b}$$

These signed attributions are order-invariant and satisfy $\psi_{ij}^{\mathrm{sens}} + \psi_{ij}^{\mathrm{mix}} = \sigma(t+s+m) - \sigma(t)$. We use $|\psi_{ij}^{\mathrm{sens}}| + |\psi_{ij}^{\mathrm{mix}}|$ only for ranking candidate interventions; the causal interpretation remains conditional on the declared SCM. For dyad (i, j), attribution ranks the edit; its fixed mask determines both endpoint bounds. Failed endpoints suppress incident edits, but the resulting mask does not inherit the original certificate.

4.2 MIC: Mechanism Intervention Certificate

Qualification asks whether a declared edit changes predictions by an acceptably small amount; it does not certify that the edit improves statistical parity.

MIC uses a pre-declared compact parameter box $\mathcal{B} \subset \mathbb{R}^4$ with $\theta^* \in \mathrm{int}(\mathcal{B})$ and defines

$$\Theta_n = \left\{ \theta \in \mathcal{B} : \widehat{D}_{\mathcal{D}}(\widehat{P} \| P_\theta) \le \epsilon_{\mathrm{KL}} \right\}, \tag{5}$$

where ϵ_{KL} is selected by a parametric bootstrap. This calibration is an empirical procedure; the proposition below only transfers a compatible-set event to a simultaneous bound. Freeze the candidate intervention mask before optimizing over Θ_n. For expected soft adjacency $\bar{A}(\theta)$ with self-loops, define

$$\mathcal{A}_i^{\pm}(\Theta_n) = \{\bar{A}^{\mathrm{obs}}(\theta), \bar{A}^{\mathrm{cf}}(\theta) : \theta \in \Theta_n\}, \qquad R_i(\theta) = \|[\bar{A}^{\mathrm{obs}}(\theta) - \bar{A}^{\mathrm{cf}}(\theta)]_{\mathcal{N}_K(i)}\|_F.$$

The fixed support $\mathcal{N}_K(i)$ includes every adjacency entry affecting prediction or degree normalization across compatible graphs. Define $\mathcal{C}_i = \mathrm{conv}\, \mathcal{A}_i^{\pm}(\Theta_n)$ and require $d_v(A) \ge d_0 > 0$ throughout it. The intervention mask and predictor are fixed before certification. Assume g_i is locally Lipschitz on an open neighborhood of $\mathcal{C}_i$, and define

$$\begin{aligned} L_i &= \sup_{A \in \mathcal{C}_i} \sup_{G \in \partial_C g_i(A)} \|G_{\mathcal{N}_K(i)}\|_F, \quad B_i(\theta) = L_i R_i(\theta), \\ \varepsilon_i^{\min} &= \min_{\theta \in \Theta_n} B_i(\theta), \qquad\qquad \varepsilon_i^{\max} = \max_{\theta \in \Theta_n} B_i(\theta). \end{aligned} \tag{6}$$

Here $\partial_C g_i(A)$ is the Clarke generalized gradient; for smooth predictors this reduces to the ordinary gradient supremum. Equation (6) maps parameter ambiguity to nodewise endpoints. Our implementation uses the full-graph bound in Eq. (C.2): spectral norms, the full feature Frobenius norm, and degree bounds over the declared domain. This support includes all normalization dependencies; weight norms alone are insufficient.

Definition 2 (MIC Actionability and Abstention Rule)**.** *A resolved node is actionable iff*

$$\varepsilon_i^{\max} - \varepsilon_i^{\min} \le \tau_{\mathrm{amb}} \quad \textit{and} \quad \min\{1, \varepsilon_i^{\max}\} \le \tau_{\mathrm{fair}}. \tag{7}$$

MIC abstains otherwise. MIC acts only when both gates in Eq. (7) *pass. Since* $0 \le \varepsilon_i^{\min} \le \varepsilon_i^{\max}$*, width never exceeds the upper bound. Thus, when* $\tau_{\mathrm{amb}} \ge \tau_{\mathrm{fair}} < 1$*, the ambiguity gate is redundant; independent screening requires* $\tau_{\mathrm{amb}} < \tau_{\mathrm{fair}}$*.*

Proposition 1 (Conditional Simultaneous Certificate)**.** *Assume* $P(\theta^* \in \Theta_n) \ge 1 - \delta_c$ *and* $\mathrm{CF}_i(\theta) \le B_i(\theta)$ *for all* i *and* $\theta \in \Theta_n$*. Then, with probability at least* $1 - \delta_c$*,* $\mathrm{CF}_i(\theta^*) \le \varepsilon_i^{\max}$ *simultaneously for every node. The proposition transfers compatible-set coverage to all-node prediction bounds; it does not prove bootstrap validity. Abstention selects which certified nodes receive an automatic action, but does not change the bound for retained nodes.*

Theorem 2 (Soft-Adjacency Stability Certificate)**.** *If* g_i *is locally Lipschitz near* $\mathcal{C}_i$ *and* $L_i < \infty$*, then for every compatible* θ *and the fixed declared intervention,*

$$|g_i(\bar{A}^{\mathrm{obs}}) - g_i(\bar{A}^{\mathrm{cf}})| \le L_i \|(\bar{A}^{\mathrm{obs}} - \bar{A}^{\mathrm{cf}})_{\mathcal{N}_K(i)}\|_F. \tag{8}$$

Equation (8) *is a soft-adjacency, not hard-edge, guarantee.*

Theorem 2 certifies prediction change under the expected soft adjacency. For a realized Bernoulli intervention, an additional concentration term is required; the resulting simultaneous statement is summarized in Supplementary Corollary 1. Accordingly, all main-paper certificate numbers are labeled "soft-adjacency" unless this concentration term is included. Numerical acceptance uses $\mathrm{UB}_{\max} - \mathrm{LB}_{\min} \le \tau_{\mathrm{amb}}$ and $\min(1, \mathrm{UB}_{\max}) \le \tau_{\mathrm{fair}}$. Selecting reported nodes preserves simultaneous coverage for this same intervention. Changing the predictor or edge mask requires recertification. Appendix A gives a separate submask extension; its larger-domain constant is not evaluated here.

4.3 MIFB: Matched-Statistics, Separated-Actions Benchmark

Qualification controls uncertainty for one graph; exposure asks whether a method reacts differently when matched graphs require different actions. MIFB uses the normalized statistic map $\mathcal{T}(G) = [D(G), H_S(G), G_{\deg}(G)]$, comprising edge density, sensitive homophily, and degree inequality. All outputs use the same dyad frame $\mathcal{R}$ with $d(a,b) = |\mathcal{R}|^{-1}\|a - b\|_1$.

$$\begin{aligned} m_{q,e}^{*,\mathrm{do}} &= \mathbf{1}\{|\psi_{q,e}^{\mathrm{sens}}| + |\psi_{q,e}^{\mathrm{mix}}| > \tau_{\mathrm{val}}\}, \\ a_{q,e}^{*,\mathrm{do}} &= m_{q,e}^{*,\mathrm{do}} \sigma(\beta_{0,q} + \beta_{x,q} c_e) + (1 - m_{q,e}^{*,\mathrm{do}}) \bar{A}_{q,e}^{\mathrm{obs}}. \end{aligned} \tag{9}$$

We predeclare $\tau_{\text{val}} = 0.25$ on validation pairs and fix it before test-pair construction; these objects separate construction from evaluation. The theorem covers both mask targets $m_q^{*,\text{do}}$ and soft targets $a_q^{*,\text{do}}$; Table 5 evaluates final soft adjacencies against the latter. The oracle shares our declared SCM and is not independent evidence of real-world causal correctness. The original MIFB cohort uses $\|\mathcal{T}(G_A) - \mathcal{T}(G_B)\|_\infty \leq 0.02$ while separating binary masks; additional audit tolerances are reported with their results.

Theorem 3 (Lower Bound for $\mathcal{T}$-Restricted Policies)**.** *For either target $t_q = m_q^{*,\text{do}}$ or $t_q = a_q^{*,\text{do}}$, put $\Delta_t = d(t_A, t_B)$. Any L_g-Lipschitz policy $r(G) \in [0,1]^{|\mathcal{R}|}$ using only $\mathcal{T}(G)$ satisfies*

$$\max_{q \in \{A,B\}} d(r(G_q), t_q) \geq \frac{[\Delta_t - L_g \varepsilon]_+}{2}, \tag{10}$$

where $\varepsilon = \|\mathcal{T}(G_A) - \mathcal{T}(G_B)\|_\infty$ and $[z]_+ = \max(z, 0)$. This theorem applies to explicitly defined $\mathcal{T}$-only policies; broader GNNs are empirical.

The inequality is elementary; MIFB contributes matched construction and common-frame auditing. If $\Delta_t > L_g \varepsilon$, matched statistics cannot support both targets. Mask separation alone does not ensure this for soft targets. We therefore report action separation $d(f_A, f_B)$ together with target error $d(f_q, a_q^{*,\text{do}})$: different edits are useful only when they track the declared target. This audit can evaluate any editor on the same frame; similar outcome gaps alone cannot establish mechanism blindness.

4.4 Training and Evidence Scope

The released implementation optimizes Eq. (11) over five mechanism environments:

$$\mathcal{L} = \tfrac{1}{5} \sum_{k=1}^{5} \mathcal{L}_{\text{cls}}^{(E_k)} + \lambda \mathcal{L}_{\text{CF}} + \gamma \mathcal{L}_{\text{MI}}, \tag{11}$$

where environments vary $(\beta_x, \beta_s, \beta_m)$, $\mathcal{L}_{\text{CF}}$ uses the soft MD-SCM intervention, and $\mathcal{L}_{\text{MI}}$ is a gradient-reversal sensitive-prediction surrogate. MIC combines spectral norms with feature and degree-normalization factors. Supplementary Algorithm G.1 specifies the pipeline.

Guarantee scope. MD-SCM identifies model-conditional contributions, MIC bounds soft-adjacency prediction change assuming compatible-set coverage, and MIFB bounds mask and soft-target error only for $\mathcal{T}$-only Lipschitz policies. Full-GNN action fidelity and all real-graph causal interpretations remain empirical/model-conditional.

5 Experiments

We ask three questions: what fairness–utility trade-off is achieved (Table 1)? Do mechanism modeling and qualified selection improve editing (Tables 2–4)? Do actions approach their declared targets beyond model matching (Table 5)? These questions connect outcome performance to intervention quality.

5.1 Settings

We use Pokec-z, Pokec-n Dong et al. (2022), German Dai & Wang (2021), and Credit Agarwal et al. (2021). All methods use stratified 60/20/20 train/validation/test splits and five seeds (42, 123, 456, 789, 2025). Hyperparameters are selected on validation data only; the default is Adam, learning rate 5×10^{-4}, for 500 epochs. We report accuracy, Δ_{SP}, W5-CD, mechanism-identification error, controlled certificate coverage/sharpness, abstention reasons, and action fidelity. Table 1 reports five-run means. For Table 5, each graph pair is averaged over five seeds before computing the mean and SD across pairs. Edges exceed validation score percentile 90; Appendix H details controls.

5.2 Four-Dataset Main Results

Table 1 compares structural editing (EDITS), invariance (FairINV), stable representation learning (SFG), and causal representation learning (FairSNR). Cells are ACC/Δ_{SP}/W5-CD (percentage/percentage/proportion).

Method	Pokec-z	Pokec-n	German	Credit
EDITS	67.53/3.91/.149	66.23/4.02/.179	70.92/6.92/.252	74.97/4.12/.179
FairINV	67.29/2.60/.114	67.90/2.80/.148	72.31/4.05/.173	79.80/3.41/.142
CDFG	–	–	73.15/3.53/.118	81.09/2.68/.096
SFG	68.11/1.83/0.094	67.72/1.67/0.089	73.02/3.67/0.126	80.97/2.73/0.098
FairSNR	68.51/1.52/0.083	67.87/1.48/0.078	73.27/3.39/0.109	81.46/2.63/0.093
Ours	68.34/.98/.052	67.53/.91/.049	73.16/2.87/.078	81.32/2.18/.061

Table 1: Matched-protocol outcomes: W5-CD is MD-SCM-relative. Controls: Appendix H. Higher ACC and lower disparities are better. Entries are five-run means; – denotes not evaluated.

Relative to FairINV, paired 95% CIs for Δ_{SP} are $[-1.99, -1.30]$, $[-2.30, -1.55]$, $[-1.54, -0.50]$, and $[-1.47, -0.81]$ percentage points in dataset order. The Pokec-n accuracy CI crosses zero ($[-0.78, 0.59]$); the accuracy difference remains inconclusive.

Against FairSNR, group disparity falls 15.3–38.5% with a 0.11–0.34 percentage-point accuracy cost. MD-SCM-relative W5-CD falls 28.4–37.3%; this declared-intervention diagnostic is not an independent fairness measure. Both disparities also fall against SFG across all datasets. Against CDFG, Δ_{SP} falls from 3.53/2.68 to 2.87/2.18 on German/Credit, with accuracy gains of .01/.23 points. MD-SCM-free W5-CD for Ours/FairSNR/FairINV/EDITS is PZ .052/.061/.073/.084; PN .049/.057/.093/.100; GE .078/.080/.110/.142; CR .061/.068/.092/.101. All values are five-seed means; shared KNN graphs score all nodes, including abstentions.

5.3 Identification, Qualification, and Exposure

Table 2 tests model-relative identification on Syn-ESCM: it separates coefficient RMSE, edge-ranking precision, true-parameter inclusion, and abstention. Standard-regime ranking precision is 0.89; these errors concern the declared generator, not event-level real-world causes. Stress regimes assess degradation rather than establish robustness outside A1–A4.

Regime	RMSE β_x	RMSE β_s	RMSE β_m	Rank precision	True-θ inclusion	Abstention
Standard	.041	.054	.061	.890	.970	.120
Collinear (CE1)	.072	.091	.118	.840	.955	.421
Cross-group-starved (CE2)	.089	.106	.154	.791	.941	.518
Dominant-mix (CE3)	.081	.098	.131	.820	.932	.635

Table 2: Syn-ESCM identification and qualification audit. Stress increases estimation error and abstention; true-parameter inclusion declines from .970 to .932.

Cross-group starvation raises mixed-coefficient RMSE from .061 to .154 and reduces precision from .890 to .791, consistent with losing the cross-group anchor that separates task and mixed effects.

Coverage alone does not show that a certificate is useful. Figure 3 shows the feasible gate geometry, and Table 3 reports the available calibration and magnitude summaries.

At these thresholds, only the magnitude gate can add rejections. Parameter uncertainty still enters $\varepsilon^{\max}$ through maximization over Θ_n. This distinguishes uncertainty-aware qualification from a point-estimate check, while the results provide no evidence for an additional width-gate benefit.

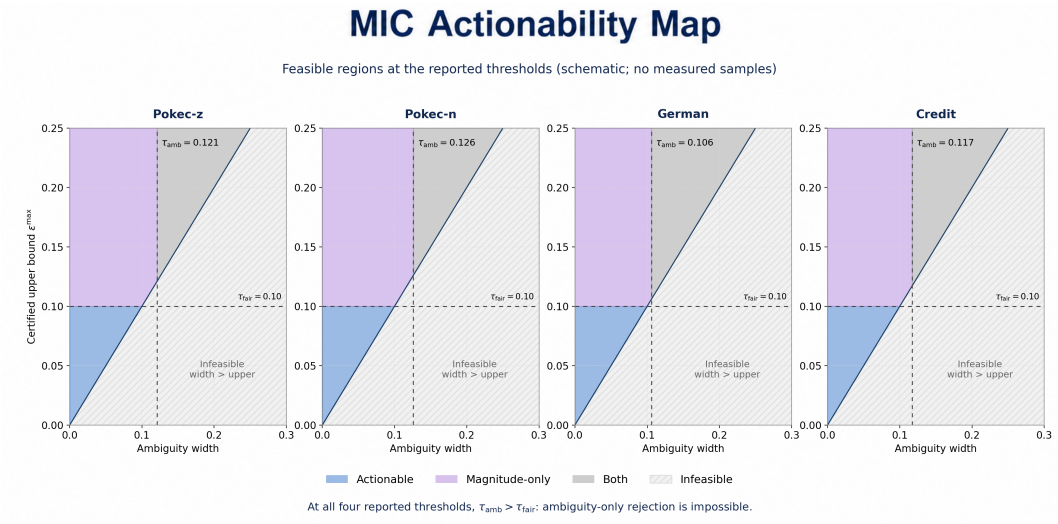


Figure 3: MIC gate geometry, not measured samples. Feasible endpoints satisfy width $\leq \varepsilon^{\max}$; hatching marks impossible pairs. All four reported ambiguity thresholds exceed $\tau_{\text{fair}} = 0.10$, so ambiguity-only rejection is impossible. Shading denotes decision regions, not empirical frequencies.

Dataset	Bootstrap inclusion (95% interval)	Med. $\varepsilon^{\max}$	P90 $\varepsilon^{\max}$	Unresolved
Pokec-z	.965 [.949,.977]	.086	.138	.021
Pokec-n	.961 [.945,.974]	.092	.147	.030
German	.969 [.954,.980]	.074	.121	.022
Credit	.964 [.949,.976]	.081	.131	.020

Table 3: Fitted-model bootstrap inclusion and five-seed fixed-mask soft endpoints using the full-graph bound in Eq. (C.2). Inclusion is model-relative; independent width-gate and post-suppression guarantees are not established by these values.

Bootstrap inclusion does not establish real-world parameter coverage. At nominal 95%, pooled inclusion is 96.8% across all tested configurations, but the minimum configuration achieves 94.0%; pooling hides local undercoverage. At nominal 99%, inclusion is 98.1%. Unresolved cases abstain. The certificate therefore retains its explicit coverage assumption.

Median endpoints lie below 0.10; ninetieth percentiles exceed it. On Pokec-z, .701 of 13,559 test nodes qualify; .299 abstain, including .021 unresolved cases. Qualification excludes subsequent incident-edge suppression; matched controls edit .71 of candidate dyads. Table 4 compares training variants and tests selection at equal edit counts.

E3 tests one environment; without main-path editing, classification/adversarial losses use original graphs, while five environments supply CF.

Variant	Dataset	ACC	Δ_{SP}	W5-CD
Full	Pokec-z	68.34	.98	.052
− interaction	Pokec-z	67.91	1.46	.081
− multi-environment	Pokec-z	67.07	2.21	.093
− adversarial	Pokec-z	67.05	2.23	.094
− main-path editing	Pokec-z	67.11	2.23	.096
− counterfactual loss	Pokec-z	67.10	2.25	.097
Matched Full	Pokec-z	68.32	1.01	0.055
Random abstention	Pokec-z	68.19	1.57	0.089

Table 4: Pokec-z ablation. Matched Full and random abstention each edit .71 of candidate dyads. Matched Full is a separate rate-matched run from the original Full row; the two controls use the same actual budget. Independent width-gate benefits are not claimed.

Component variants reduce accuracy by 1.23–1.29 points and raise Δ_{SP} from .98 to 2.21–2.25; W5-CD rises to .093–.097. Removing counterfactual loss gives the largest disparity increase, but exceeds others by only .02–.04 points, insufficient to establish dominance. Removing interaction raises W5-CD to .081. At equal .71 edit fractions, qualified selection reduces W5-CD from .089 to .055 versus random abstention. Post-suppression masks remain outside certification.

Outcome disparities cannot reveal whether matched graphs trigger different structural responses. Table 5 therefore compares final soft adjacencies on one frame. Here f_q is the final soft adjacency on $\mathcal{R}$. All non-editing predictors share $f_q = \bar{A}_q^{\mathrm{obs}}$ and therefore one reference row; editors report their soft outputs under the stated budget.

Method	$d(f_A, f_B)$	$d(f_A, a_A^{*,\mathrm{do}})$	$d(f_B, a_B^{*,\mathrm{do}})$
No-edit reference	.011±.004	.540±.060	.560±.071
EDITS	.019±.006	.380±.050	.410±.050
Ours	.168±.020	.052±.010	.048±.009
IPW-SCM only	0.142±0.019	0.093±0.016	0.089±0.016
IPW-SCM, shift	0.136±0.021	0.128±0.022	0.123±0.022
Ours, shift	0.155±0.022	0.072±0.014	0.070±0.014

Table 5: MIFB mean ± SD across graph pairs; each pair is first averaged over five seeds. Additional audits use 200 pairs each, $\Delta_{\mathrm{do}} \geq .151$, and statistic mismatch $\leq .022$; original rows retain their original cohort. Shift adds $.31c^2$ to the generating logit, omitted from the fitted SCM; its task-preserving oracle retains this term.

On the original cohort, mean error across both targets is .050 versus .395 for EDITS; the Ours row implies target separation $\geq .0665$, allowing rounding. On the shared shifted cohort, error is .071 versus .1255 for IPW-SCM, a 43.4% reduction. These results support full-pipeline action fidelity; A1 fails under shift, so this evidence is empirical, not a MIC-only guarantee.

Action separation must accompany low target error. The shifted control tests fidelity beyond exact model matching; matched budgets isolate selection. The lower bound covers $\mathcal{T}$-only policies; full-GNN comparisons remain empirical.

5.4 Robustness and Failure Modes

Under identification stress, MIC W5-CD values are 0.095, 0.102, and 0.148 for CE1–CE3, versus 0.142, 0.158, and 0.195 for always intervening. Abstention rises from 42.1% to 63.5% as the failure regime worsens. These comparisons connect failure detection to downstream performance: abstention increases under stress while tail disparity remains below always intervening, supporting conservative action selection within the tested failure regimes. Under 20% label and edge perturbations, the reported Δ_{SP} degradations are 0.41 and 0.54 for our method versus 0.71 and 0.88 for FairINV; paired 95% CIs are $[-.38, -.12]$ and $[-.44, -.15]$. These are soft-adjacency diagnostics for the declared settings, not robustness to arbitrary latent confounding.

6 Limitations

Identification assumes a specified independent-dyad model; no diagnostic detects latent confounding despite checks. MIC certifies interventions; changed masks need recertification; hard graphs need corrections. The shifted audit covers one misspecification; independent auditing is measured on all four datasets under one generator. Single points prove no frontier dominance; MIFB theory remains $\mathcal{T}$-restricted.

7 Conclusion

Identify–qualify–expose improves reported fairness outcomes while making structural decisions testable. Matched-budget gains support selection; four-dataset independent audits support consistent structural-sensitivity gains. Its broader contribution is a common audit for future editors: specify the mechanism to replace, qualify the fixed intervention, and measure error against matched-pair targets. This connects outcome improvements to action quality within explicit model and certificate limits.

Reproducibility Statement

The main paper states the estimands, assumptions, intervention object, certificate scope, datasets, splits, metrics, and five-seed protocol. Appendix A gives complete proofs; Appendices C–E document compatible-set optimization, realized-adjacency scope, environments, preprocessing, and hyperparameters. The supplement specifies the MD-SCM objective, MIC endpoint brackets, MIFB constructor, and evaluation protocol needed to reconstruct the reported pipeline.

Ethical Statement

This work studies algorithmic fairness on graphs containing sensitive attributes. Its certificates are model-conditional and must not be interpreted as guarantees of societal fairness or unique causes of individual outcomes. Dataset preprocessing follows established benchmark protocols; no new human-subject data are collected. In high-stakes use, abstention requires human review, and deployments should audit privacy, subgroup harms, distribution shift, and failures of the declared SCM.

AI Use Statement

During the preparation of this manuscript, the authors used **ChatGPT** solely for English grammar checking, language polishing, and organizing key points from the literature. ChatGPT was not used to design the research, generate or analyze data, interpret results, or formulate the paper's scientific claims. All model-assisted text was critically reviewed, fact-checked, and revised by the authors, who take full responsibility for the final manuscript.

References

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pp. 2114–2124. PMLR, 2021.

Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *International Conference on Machine Learning*, pp. 1003–1013. PMLR, 2020.

Meng Cao, Mingcai Chen, Shuangjie Li, Hualei Yu, Shuai Feng, and Demin Gao. MultiFPT: Towards multi-attribute fairness in pre-trained graph neural networks via prompt tuning. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Qiang Chen, Zhongze Wu, Xiu Su, Xi Lin, Zhe Qu, Shan You, Shuo Yang, and Chang Xu. Stable fair graph representation learning with lipschitz constraint. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 9354–9376, 2025.

Yifan Cheng, Guixian Zhang, Debo Cheng, and Shichao Zhang. FairSIN: Achieving fairness in graph neural networks through sensitive information neutralization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 11661–11669, 2024.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 680–688, 2021.

Yucong Dai, Lu Zhang, Yaowei Hu, Susan Gauch, and Yongkai Wu. Causally fair node classification on non-iid graph data. *Transactions on Machine Learning Research*, 2026.

Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. EDITS: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pp. 1259–1269, 2022.

Yushun Dong, Binchi Zhang, Hanghang Tong, and Jundong Li. Certified defense on the fairness of graph neural networks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 211–220, 2026.

Adiba Ejaz and Elias Bareinboim. Relational structural causal models. In *International Conference on Machine Learning (ICML)*, 2026.

Zhimeng Guo, Jialiang Li, Teng Xiao, Yao Ma, and Suhang Wang. Towards fair graph neural networks via graph counterfactual. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 669–678, 2023.

Mahdi Tavassoli Kejani, Fadi Dornaika, Charlotte Laclau, and Jean-Michel Loubes. Homophily-aware supervised contrastive counterfactual augmented fair graph neural network. *arXiv preprint arXiv:2604.02342*, 2026.

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

O. Deniz Kose and Yanning Shen. FairWire: Fair graph generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 62246–62268, 2024.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4066–4076, 2017.

Yeon-Chang Lee, Hojung Shin, and Sang-Wook Kim. DAB-GNN: Disentangling, amplifying, and debiasing for fair graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 12013–12021, 2025.

Hourun Li, Yifan Wang, Qinghua Ran, Junyu Luo, Jia Yang, Changling Zhou, Tianzi Zang, and Yu Wang. Fair message passing with edge dropout for graph neural networks. *arXiv preprint arXiv:2412.09507*, 2024.

Chuxun Liu, Debo Cheng, Qingfeng Chen, Jiangzhang Gan, Jiuyong Li, and Lin Liu. Learning fair graph representations with multi-view information bottleneck. *arXiv preprint arXiv:2510.25096*, 2025.

Chuxun Liu, Qingfeng Chen, Debo Cheng, Jiangzhang Gan, Jiuyong Li, and Lin Liu. Learning fair graph representations via probability of necessity and sufficiency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23667–23675, 2026. doi: 10.1609/aaai.v40i28.39540.

Haohui Lu, Jiyuan Tian, Fangyu Zhou, and Shahadat Uddin. Subgraph filtering for fair graph neural networks. *arXiv preprint arXiv:2608.26437*, 2026. URL https://arxiv.org/abs/2608.26437.

Jiahua Lu, Huaxiao Liu, Shuotong Bai, Junjie Xu, Renqiang Luo, and Enyan Dai. Let's grow an unbiased community: Guiding the fairness of graphs via new links, 2025.

Renqiang Luo, Huafei Huang, Tao Tang, Jing Ren, Ziqi Xu, Mingliang Hou, Enyan Dai, and Feng Xia. FairGE: Fairness-aware graph encoding in incomplete social networks. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 695–703, 2022.

Lilian Marey, Mathilde Perez, Tiphaine Viard, and Charlotte Laclau. Structural bias beyond homophily: A study of fairness in link prediction. *arXiv preprint arXiv:2602.11802*, 2026. URL https://arxiv.org/abs/2602.11802.

Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

Wendi Wang, Jiaxi Yang, Yongkang Du, and Lu Lin. Mitigating topology biases in graph diffusion via counterfactual intervention. *arXiv preprint arXiv:2603.02005*, 2026. URL https://arxiv.org/abs/2603.02005.

Xuemin Wang, Tianlong Gu, Xuguang Bao, Qinghua Ran, and Ping Li. Towards fair graph neural networks via graph counterfactual without sensitive attributes. In *International Conference on Data Engineering (ICDE)*, pp. 265–277, 2025a.

Yifan Wang, Hourun Li, Ling Yue, Zhiping Xiao, Jia Yang, Changling Zhou, Wei Ju, Ming Zhang, and Xiao Luo. DANCE: Dual unbiased expansion with group-acquired alignment for out-of-distribution graph fairness learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63516–63529, 2025b. URL https://proceedings.mlr.press/v267/wang25bf.html.

Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1938–1948, 2022.

Zichong Wang, Zhibo Chu, Thang Viet Doan, Shaowei Wang, Yongkai Wu, Vasile Palade, and Wenbin Zhang. Fair graph u-net: A fair graph learning framework integrating group and individual awareness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28485–28493, 2025c. doi: 10.1609/aaai.v39i27.35071.

Zichong Wang, Zhipeng Yin, Fang Liu, Zhen Liu, Christine Lisetti, Rui Yu, Shaowei Wang, Jun Liu, Sukumar Ganapati, Shuigeng Zhou, and Wenbin Zhang. Graph fairness via authentic counterfactuals: Tackling structural and causal challenges. *ACM SIGKDD Explorations Newsletter*, 26(2): 89–98, 2025d. doi: 10.1145/3715073.3715081.

Zhenqiang Ye, Jinjie Lu, Tianlong Gu, Fengrui Hao, and Xuemin Wang. FairGSE: Fairness-aware graph neural network without high false positive rates. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 16163–16171, 2026.

Guixian Zhang, Debo Cheng, Guan Yuan, and Shichao Zhang. Learning fair representations via rebalancing graph structure. *Information Processing & Management*, 61(1):103570, 2024a.

Guixian Zhang, Debo Cheng, Guan Yuan, Shang Liu, and Yanmei Zhang. Causality-inspired disentanglement for fair graph neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 637–645, 2025.

Zhongjian Zhang, Mengmei Zhang, Yue Yu, Cheng Yang, Jiawei Liu, and Chuan Shi. Endowing pre-trained graph models with provable fairness. In *Proceedings of the ACM Web Conference 2024*, pp. 1045–1056, 2024b. doi: 10.1145/3589334.3645703.

Yuchang Zhu, Jintang Li, Yatao Bian, Zibin Zheng, and Liang Chen. One fits all: Learning fair graph neural networks for various sensitive attributes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4688–4699, 2024. doi: 10.1145/3637528.3672029.

Daniel Zügner and Stephan Günnemann. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1656–1665, 2020.

A Proofs and Assumptions

A.1 Proof of Theorem 1

Motivation. The theorem is a statement about the declared candidate-dyad frame, not an unrestricted claim about why a social edge exists. The proof therefore has three separate contracts: the frame and its sampling design, the weighted score, and the nonsingular information needed to turn a zero score into a unique parameter estimate.

Step 0: candidate frame and sampling design. Let $\mathcal{R}$ contain the dyads to which the declared edge equation is applied, and let $\mathcal{D} \subseteq \mathcal{R}$ contain every observed edge together with a without-replacement probability sample of non-edges. Within each edge-status stratum, inclusion probabilities $\pi_{ij} > 0$ are logged and $w_{ij} = 1/\pi_{ij}$. This is precisely the A2 design condition; it allows sparse-graph non-edge probabilities to decrease with graph size. We do not replace it by a fixed lower bound. The weighted information condition in A4 controls the resulting concentration through

$$n_{\text{eff}} = \frac{(\sum_{e\in\mathcal{D}} w_e)^2}{\sum_{e\in\mathcal{D}} w_e^2}, \qquad r_w = \frac{\sqrt{\sum_{e\in\mathcal{D}} w_e^2}}{\sum_{e\in\mathcal{D}} w_e}. \tag{A.1}$$

Step 1: weighted score. Put $z_{ij} = (1, c_{ij}, h_{ij}, h_{ij}c_{ij})^\top$ and $\eta_{ij}(\beta) = z_{ij}^\top\beta$. The weighted log-likelihood and score are

$$\ell_w(\beta) = \sum_{(i,j)\in\mathcal{D}} w_{ij}\{A_{ij}\eta_{ij}(\beta) - \log(1 + e^{\eta_{ij}(\beta)})\}, \tag{A.2a}$$

$$U_w(\beta) = \sum_{(i,j)\in\mathcal{D}} w_{ij} z_{ij}\{A_{ij} - \sigma(\eta_{ij}(\beta))\}. \tag{A.2b}$$

Normalize the score as $\bar{U}_w(\beta) := (\sum_{e\in\mathcal{D}} w_e)^{-1} U_w(\beta)$. Under A1, $\mathbb{E}[\bar{U}_w(\beta^*)] = 0$ for the declared edge equation. The without-replacement design LLN preserves this expectation after inverse-probability weighting within each stratum; A3 supplies the conditional independence needed for the corresponding variance calculation.

Step 2: information and local expansion. Define $\bar{H}_w(\beta) = -\nabla \bar{U}_w(\beta)$. The normalized information is

$$I_w(\beta^*) := \operatorname{plim} \bar{H}_w(\beta^*) = \frac{\mathbb{E}_s\left[w_{ij} z_{ij} z_{ij}^\top \sigma(\eta_{ij}(\beta^*))[1-\sigma(\eta_{ij}(\beta^*))]\right]}{\mathbb{E}_s[w_{ij}]}. \tag{A.3}$$

By A4 this normalized information has eigenvalues bounded away from zero, while $\max_e w_e / \sqrt{\sum_e w_e^2} \to 0$ prevents a single sampled dyad from dominating. The design LLN/CLT together with A3–A4 makes the IPW score asymptotically linear with effective sample size n_{eff}. The design LLN gives $\bar{H}_w(\tilde{\beta}) \xrightarrow{p} I_w(\beta^*)$, while the design CLT and weight-concentration condition give $\bar{U}_w(\beta^*) = O_p(r_w)$. Therefore,

$$0 = \bar{U}_w(\widehat{\beta}) = \bar{U}_w(\beta^*) - \bar{H}_w(\tilde{\beta})(\widehat{\beta} - \beta^*), \tag{A.4}$$

implies $\widehat{\beta} - \beta^* = O_p(r_w)$. This is model-conditional estimation, not event-level causal discovery. For $h = 0$, the design identifies (β_0, β_x); for $h = 1$, it identifies $(\beta_0 + \beta_s, \beta_x + \beta_m)$. Positive probability for both strata and nonzero within-stratum variance of c give full rank. Identification is thus a standard regression consequence, not a new causal-discovery result.

Step 3: initialization is not identification. Cross-group fitting supplies a stable initializer for (β_0, β_x) because $h_{ij} = h_{ij} c_{ij} = 0$ on cross-group dyads. Same-group reduced fits initialize β_s and β_m. The final claim, however, follows from the joint maximizer of ℓ_w and the positive-definite information above, not from either reduced fit. Uniform convergence on $\mathcal{B}_\beta$, together with A4's unique population maximizer, then makes the joint IPW estimator consistent.

Step 4: order-invariant contribution identity. Let $p_t = \sigma(\beta_0 + \beta_x c)$, $p_s = \sigma(\beta_0 + \beta_x c + \beta_s h)$, $p_m = \sigma(\beta_0 + \beta_x c + \beta_m hc)$, and $p_f = \sigma(\beta_0 + \beta_x c + \beta_s h + \beta_m hc)$. The two-order contributions are

$$\phi_s = \tfrac{1}{2}[(p_s - p_t) + (p_f - p_m)], \tag{A.5a}$$

$$\phi_m = \tfrac{1}{2}[(p_m - p_t) + (p_f - p_s)]. \tag{A.5b}$$

Summing Eqs. (A.5a)–(A.5b) gives $\phi_s + \phi_m = p_f - p_t$. Averaging the two orders removes dependence on whether the sensitive or mixed term is removed first. Therefore, the theorem identifies model-conditional propensity contributions; it does not identify a unique event-level social cause.

A.2 Proof of Proposition 1

Step 1: bootstrap compatible set. Let $\widehat{\beta}^{(0)}$ be the joint IPW estimate. Generate B parametric bootstrap samples from $\widehat{P}_{ij}^{(0)} = \sigma(z_{ij}^\top \widehat{\beta}^{(0)})$, refit the same estimator under the same declared design, and compute the normalized weighted discrepancy $\widehat{D}_{\mathcal{D}}(\widehat{P}^{(0)} \| P_{\widehat{\beta}^{(b)}})$. With the $(1-\delta_c)$ empirical quantile ϵ_{KL}, define

$$\Theta_n = \left\{ \theta \in \mathcal{B} : \widehat{D}_{\mathcal{D}}(\widehat{P}^{(0)} \| P_\theta) \le \epsilon_{\text{KL}} \right\}. \tag{A.6}$$

The proposition is conditional on the calibrated event $\mathcal{C}_n = \{\theta^* \in \Theta_n\}$; bootstrap validity is an empirical calibration question, not a consequence of the algebra below.

Step 2: simultaneous transfer. On $\mathcal{C}_n$, the assumed soft-adjacency Lipschitz inequality holds for every $\theta \in \Theta_n$. Thus, for every node, simultaneously,

$$\mathrm{CF}_i(\theta^*) \le B_i(\theta^*) \le \max_{\theta \in \Theta_n} B_i(\theta) = \varepsilon_i^{\max}. \tag{A.7}$$

Equation (A.7) is simultaneous over nodes on $\mathcal{C}_n$. If $P(\mathcal{C}_n) \ge 1 - \delta_c$, this is an all-node statement with the same probability, not a separate per-node correction. Any data-dependent subset selected after this event inherits the same simultaneous event.

Step 3: abstention is post-selection. MIC retains node i only when $\varepsilon_i^{\max} - \varepsilon_i^{\min} \le \tau_{\text{amb}}$, the clipped upper bound is at most τ_{fair}, and the endpoint solver is resolved. Otherwise, it abstains. This changes deployment eligibility but does not change the certified upper bound for a retained node and does not create a new conditional-coverage claim.

A.3 Proof of Theorem 2

For a fixed compatible parameter θ, write $R_i(\theta) = \|(\bar{A}^{\mathrm{obs}}(\theta) - \bar{A}^{\mathrm{cf}}(\theta))_{\mathcal{N}_K(i)}\|_F$. Let $A(t) = \bar{A}^{\mathrm{obs}} + t(\bar{A}^{\mathrm{cf}} - \bar{A}^{\mathrm{obs}})$, which stays in $\mathcal{C}_i$. Local Lipschitzness makes $g_i \circ A$ absolutely continuous. The Clarke chain rule bounds its derivative for almost every t by $L_i\|(\bar{A}^{\mathrm{obs}} - \bar{A}^{\mathrm{cf}})_{\mathcal{N}_K(i)}\|_F$. Integrating proves Eq. (8) for the declared intervention. Uniformity over $\theta \in \Theta_n$ permits simultaneous node selection, but does not authorize an untested change to the edge mask. The degree restriction $d_v(A) \geq d_0$ keeps the normalization factor finite; ReLU nondifferentiability therefore does not invalidate the certificate. For the two-layer normalized GNN, Eq. (C.2) bounds L_i using classifier and convolution spectral norms, the full feature Frobenius norm, and the compatible-domain degree-normalization factor. Taking the minimum and maximum of $B_i(\theta)$ over Θ_n produces the endpoint brackets in Eq. (6). This proves the soft-adjacency certificate; it does not silently replace an expected adjacency by a Bernoulli realization.

Optional submask extension. For symmetric $M \in [0,1]^{|\mathcal{R}|}$, let $A_M = \bar{A}^{\mathrm{obs}} + M \odot (\bar{A}^{\mathrm{cf}} - \bar{A}^{\mathrm{obs}})$ and $\mathcal{C}_i^{\mathrm{env}} = \mathrm{conv}\{A_M(\theta) : \theta \in \Theta_n, M \in [0,1]^{|\mathcal{R}|}\}$. Define L_i^{env} by the same gradient supremum over this larger domain, assuming local Lipschitzness and positive degrees there. Since $\|A_M - \bar{A}^{\mathrm{obs}}\|_{F,\mathcal{N}_K(i)} \leq R_i(\theta)$, the same integration proves $|g_i(A_M) - g_i(\bar{A}^{\mathrm{obs}})| \leq L_i^{\mathrm{env}} R_i(\theta)$ uniformly over submasks. Generally $L_i^{\mathrm{env}} \geq L_i$; the fixed-mask magnitudes in Table 3 do not validate this optional extension.

For a hard Bernoulli realization A^{hard} with independent candidate entries, concentration controls each receptive-field deviation $\|A^{\mathrm{hard}} - \bar{A}\|_F$. A union bound over nodes gives the following corollary.

Corollary 1 (Realized-adjacency extension)**.** *With probability at least $1 - \delta_h$, the realized intervention satisfies*

$$|g_i(A^{\mathrm{obs}}) - g_i(A^{\mathrm{cf}})| \leq L_i R_i(\theta) + L_i\{\rho_i^{\mathrm{obs}}(\delta_h) + \rho_i^{\mathrm{cf}}(\delta_h)\}, \tag{A.8}$$

where each ρ is the corresponding Bernoulli concentration radius on the receptive field. Here L_i must be recomputed on a domain containing both hard realizations and their connecting paths. Joint parameter and realization coverage is at least $1 - \delta_c - \delta_h$; main-paper values omit the radii.

A.4 Proof of Theorem 3

Step 1: matched statistics and separated actions. MIFB uses the statistic map in Eq. (B.1) and the common-frame action distance in Eq. (D.2). For fixed N and matched E, the density coordinate is redundant, but retaining it makes the theorem match the released constructor. The construction oracle uses only the mechanism labels and declared target replacement. It is therefore independent of the evaluated policy. A matched pair satisfies $\|\mathcal{T}(G_A) - \mathcal{T}(G_B)\|_\infty \leq \varepsilon$ while its binary masks are separated by $\Delta_{\mathrm{do}} = d(m_A^{*,\mathrm{do}}, m_B^{*,\mathrm{do}})$.

Step 2: Lipschitz response and triangle inequality. The Lipschitz condition gives $d(r(G_A), r(G_B)) \leq L_g\varepsilon$. By the triangle inequality,

$$\Delta_{\mathrm{do}} \leq d(m_A^{*,\mathrm{do}}, r_A) + d(r_A, r_B) + d(r_B, m_B^{*,\mathrm{do}}). \tag{A.9}$$

If both endpoint errors were smaller than $(\Delta_{\mathrm{do}} - L_g\varepsilon)/2$, Eq. (A.9) would be violated. Replacing both masks by the soft targets gives the identical argument with $\Delta_t = d(a_A^{*,\mathrm{do}}, a_B^{*,\mathrm{do}})$, proving Eq. (10) for either target. The constructor guarantees binary separation only; a nonvacuous soft bound additionally requires $\Delta_t > L_g\varepsilon$. For the audited construction, the replacement-action gate enforces $\Delta_{\mathrm{do}} \geq \delta_{\min} = 0.15$ and $\varepsilon = 0.02$; the theorem keeps the general positive-part form because the bound depends on the policy-specific L_g. It applies only to policies that use the declared $\mathcal{T}$ and satisfy the stated Lipschitz constant. Full GNN action fidelity remains an empirical diagnostic under the same common-frame protocol.

B Definitions and Applicability

B.1 Statistic map used by MIFB

MIFB uses the three-dimensional aggregate map implemented in `src/mifb.py`. For a graph with N nodes, E undirected edges, sensitive homophily H_S, edge density $D = 2E/[N(N-1)]$, and degree Gini $G_{\deg} = \sum_{u,v\in\mathcal{V}} |d_u - d_v|/[2N\sum_v d_v]$, we use

$$\mathcal{T}(G) = [D(G), H_S(G), G_{\deg}(G)]\,. \tag{B.1}$$

All three coordinates lie in $[0,1]$; when all degrees are zero, we define $G_{\deg} = 0$. Pairs are accepted when $\|\mathcal{T}(G_A) - \mathcal{T}(G_B)\|_\infty \le 0.02$. Actions are vectors on the common candidate frame $\mathcal{R}$ and use $d(a,b) = |\mathcal{R}|^{-1}\|a-b\|_1$.

B.2 Mechanism-level structural counterfactual fairness

Definition 3 (Mechanism-Level Structural Counterfactual Fairness)**.** *Let $p_{\Theta,\phi} = c_\phi \circ f_\Theta$ be the predicted positive-class probability. An encoder satisfies ε-mechanism-level structural counterfactual fairness on node i if, for every soft counterfactual adjacency generated by the declared masked intervention,*

$$\left|p_{\Theta,\phi}(\bar{A}^{\mathrm{obs}}, X)_i - p_{\Theta,\phi}(\bar{A}^{\mathrm{cf}}, X)_i\right| \le \varepsilon_i. \tag{B.2}$$

The intervention sets the sensitive-homophily and mixed mechanisms to zero only on flagged dyads and preserves the task component and the nuisance base-density offset elsewhere. This definition is formalized in Eq. (B.2)*; it is relative to this declared intervention object; it is not a claim of societal or event-level fairness.*

B.3 Identifiability contract

The main theorem uses A1–A4 as a contract rather than as informal caveats. A1 is a correct specification of the declared dyad equation. A2 is the within-stratum without-replacement sampling design with known positive first-order inclusion probabilities and a design LLN/CLT. A3 is conditional dyad independence. A4 specifies the compact parameter space, unique population maximizer, positive-definite normalized weighted information, and weight concentration. The following table records what each condition does and how a failure is detected before interpreting a certificate.

Condition	Role	Failure mode	Observable diagnostic
A1: correct specification	Makes β^* the target of the weighted likelihood	Misspecified edge mechanism or omitted interaction	held-out weighted log-likelihood and synthetic recovery
A2: declared sampling design	Makes IPW score unbiased and supplies the design LLN/CLT	Unlogged inclusion or invalid within-stratum sampling	inclusion-probability log and sampling audit
A3: conditional dyad independence	Supplies the score variance decomposition	Cluster or temporal dependence	cluster sensitivity analysis; no automatic theorem extension
A4: parameter space, identification, information, and weights	Gives global identification, local invertibility, and effective-sample-size control	Non-unique population optimum, collinearity, or one dyad dominates the weighted score	multi-start objective audit, Fisher condition number, and $\max w_e/\sqrt{\sum_e w_e^2}$
CE1 stress check	Flags severe task–sensitive collinearity	$\mathrm{dCor}(c,h) > 0.8$	weighted distance correlation
CE2 stress check	Checks cross-group signal for β_x	$\lvert\mathcal{E}_{\mathrm{cross}}\rvert < 100$	cross-group edge count
CE3 stress check	Checks mixed-term dominance	$\lvert\widehat{\beta}_m/\widehat{\beta}_x\rvert > 1.5$	fitted coefficient ratio

Table 6: Identifiability contract and applicability diagnostics. CE1–CE3 are stress regimes, not additional assumptions. A failed diagnostic triggers qualification or abstention; it does not prove that the declared model is false.

CE1 uses 0.8 for synthetic stress; 0.03 is a heuristic diagnostic threshold, not evidence that German or Pokec passes it. Removing sensitive columns leaves possible proxies; low dependence between c and h does not establish independence between remaining features and S. If diagnostics pass, the declared model and MIC gates still determine action; no latent-confounding safeguard is implied.

C MIC COVERAGE AND SOLVER AUDIT

C.1 REPEATED SYNTHETIC COVERAGE

The coverage audit repeats the complete compatible-set construction 2,400 times at three nominal levels. A success means that the injected synthetic parameter belongs to the fitted compatible set and that every reported node endpoint obeys the simultaneous upper-bound event. The Wilson intervals below are descriptive calibration intervals; they are not a claim that every configuration reaches its nominal target.

Nominal level	Successful repetitions	Empirical rate	Wilson 95% CI
0.90	2237/2400	.932	[.921,.941]
0.95	2323/2400	.968	[.960,.974]
0.99	2354/2400	.981	[.975,.986]

Table 7: Synthetic repeated-coverage audit for the MIC compatible-set event. The pooled 0.95 result is conservative, while a configuration-level rate can be as low as .940; the 0.99 row is .981 and therefore is not reported as attaining 99% nominal coverage.

C.2 FROM EDGE PROBABILITIES TO A NODE CERTIFICATE

For a fixed compatible parameter θ, let $q_e^{\mathrm{obs}}(\theta)$ and $q_e^{\mathrm{cf}}(\theta)$ be the observed and masked-do edge probabilities on the K-hop receptive field of node i. The soft adjacency difference is bounded by

$$R_i(\theta) = \left\|\bar{A}^{\mathrm{obs}}(\theta) - \bar{A}^{\mathrm{cf}}(\theta)\right\|_{F,\mathcal{N}_K(i)} \le \sqrt{m_i} \max_{e\in\mathcal{N}_K(i)} \left|q_e^{\mathrm{obs}}(\theta) - q_e^{\mathrm{cf}}(\theta)\right|, \tag{C.1}$$

where m_i counts the affected matrix entries, including both symmetric entries for each undirected dyad. For a two-layer GNN with nonexpansive activations, an explicit conservative bound on the full N-node support is

$$L_i \le 2\|W_{\mathrm{cls}}\|_2\|W_1\|_2\|W_0\|_2\|X\|_F \left(d_0^{-1} + d_{\max}\sqrt{N}\, d_0^{-2}\right). \tag{C.2}$$

For symmetric nonnegative A, write $S(A) = D^{-1/2}AD^{-1/2}$. Along the declared domain, $\|S(A)\|_2 \le 1$, $\|A\|_2 \le d_{\max}$, and $\|\mathrm{d}D\|_2 \le \sqrt{N}\|\mathrm{d}A\|_F$. Differentiating both degree factors bounds $\|\mathrm{d}S\|_F$ by the parenthesized factor times $\|\mathrm{d}A\|_F$; the two convolution occurrences give the factor two. The reported endpoints use Eq. (C.2) on full-graph support, including all degree-normalization dependencies. Here d_0 and $d_{\max}$ bound degrees throughout the declared compatible domain, rather than only the observed graph. Combining Eqs. (C.1) and (C.2) with the local Lipschitz inequality yields

$$\mathrm{CF}_i(\theta) \le B_i(\theta) = L_i R_i(\theta). \tag{C.3}$$

Optimizing Eq. (C.3) gives the endpoints in Eq. (6); the endpoint quantities are not re-numbered here. This is a soft-adjacency result. A realized Bernoulli intervention adds the concentration radii in Corollary 1.

C.3 OUTER COVER AND ENDPOINT BRACKETS

The solver partitions the entire predeclared box $\mathcal{B}$; a smaller box is valid only after proving that it contains Θ_n. The candidate mask is frozen during this optimization, so threshold crossings do not invalidate cellwise continuity. For cell C_j with center θ_j and radius r_j, the KL lower bound is obtained from the center value and a conservative local Lipschitz remainder:

$$\mathrm{KL}_{\mathrm{LB}}(C_j) = \max\{0, \widehat{D}_{\mathcal{D}}(\widehat{P}\|P_{\theta_j}) - K_{\mathrm{KL}} r_j\}. \tag{C.4}$$

For Euclidean parameter cells, $\|z_e\|_2 \le 2$ gives $K_{\mathrm{KL}} = 2$ for normalized weighted Bernoulli KL. With a fixed mask, $K_i^\theta = L_i\sqrt{m_i}$ is conservative because each probability-difference gradient has norm at most one. Cells with the lower bound in Eq. (C.4) above ϵ_{KL} are discarded; feasible centers

provide lower-bound witnesses. If K_i^θ is the cellwise Lipschitz constant of B_i, the four brackets are

$$\mathrm{LB}_{\max} = \max_{\theta \in F} B_i(\theta), \tag{C.5a}$$

$$\mathrm{UB}_{\max} = \max_{j \in J} \{B_i(\theta_j) + K_i^\theta r_j\}, \tag{C.5b}$$

$$\mathrm{LB}_{\min} = \min_{j \in J} \max\{0, B_i(\theta_j) - K_i^\theta r_j\}, \tag{C.5c}$$

$$\mathrm{UB}_{\min} = \min_{\theta \in F} B_i(\theta). \tag{C.5d}$$

The solver uses Eqs. (C.5a)–(C.5d) to close both endpoint brackets, where J is the surviving-cell set and F the feasible-witness set. The solver reports a node as unresolved when either endpoint gap exceeds η_{opt}; unresolved nodes abstain. The released audit uses 1,600 visited cells, maximum endpoint gap .009, and $q^* = 5$ for each dataset (Table 12). This makes the numerical stopping rule part of the certificate contract rather than an unreported implementation detail.

C.4 Realized Bernoulli adjacency

The main certificate concerns soft adjacency. An optional hard deployment requires an enlarged Lipschitz domain before sampling; a bound valid only on the soft compatible hull is insufficient. Let A^{hard} have independent Bernoulli entries with mean $\bar{A}$ on a receptive field of size m_i. A Hoeffding-type radius $\rho_i(\delta_h)$ is chosen so that

$$\Pr\big(\|A^{\mathrm{hard}} - \bar{A}\|_{F,\mathcal{N}_K(i)} > \rho_i(\delta_h)\big) \le \delta_h/(2N). \tag{C.6}$$

Applying Eq. (C.6) to the factual and counterfactual draws and taking a union bound over both sides and all nodes yields Eq. (A.8) with probability at least $1 - \delta_h$. The main-paper MIC values intentionally report only the first term, because they certify the expected soft adjacency. The hard-adjacency radii change the certified magnitude and must be included when rechecking deployment eligibility.

D MIFB Construction Audit

MIFB fixes one candidate-dyad frame $\mathcal{R}$ and constructs two graphs from a shared undirected base graph. The task-dominated graph uses a large β_x/β_s ratio; the bias-dominated graph uses the reciprocal regime. Degree-preserving 2-switches keep density and degree Gini fixed, and a switch is retained only when sensitive homophily stays within ε. On the same frame, the construction oracle is the replacement-action mask

$$m_{q,e}^{*,\mathrm{do}} = \mathbf{1}\{|\psi_{q,e}^{\mathrm{sens}}| + |\psi_{q,e}^{\mathrm{mix}}| > \tau_{\mathrm{val}}\}, \tag{D.1}$$

where one denotes executing $(\beta_s, \beta_m) \leftarrow (0, 0)$ on dyad e and zero denotes preserve/abstain. The mask uses a pre-declared validation threshold $\tau_{\mathrm{val}} = 0.25$ fixed before test-pair construction, and never reads the evaluated model's predictions.

For each pair, the action metric and construction quantities are

$$d(a,b) = |\mathcal{R}|^{-1}\|a - b\|_1, \qquad \Delta_{\mathrm{do}} = d(m_A^{*,\mathrm{do}}, m_B^{*,\mathrm{do}}), \qquad \|\mathcal{T}(G_A) - \mathcal{T}(G_B)\|_\infty \le \varepsilon. \tag{D.2}$$

The implementation uses $\varepsilon = 0.02$ and accepts only pairs whose binary replacement-action separation satisfies $\Delta_{\mathrm{do}} \ge \delta_{\min} = 0.15$. These gates define the original cohort in Table 5; the additional cohorts use Appendix H's .022 tolerance and .151 mask-separation gate. The theorem then gives the policy-dependent lower bound in Eq. (10). Because the released full-GNN path does not expose a single independent L_g for all models, we retain this formula rather than promoting a fixed numeric lower bound to a theorem claim. The value .061 may be reported only for a separately declared policy and audited L_g. Thus Table 5 is an empirical full-GNN action-fidelity diagnostic, while the lower bound is restricted to $\mathcal{T}$-only Lipschitz policies.

E Experimental Protocol

E.1 Environment construction

The five training environments are the exact defaults in `src/train.py`; β_0 is recalibrated for the target density and is not treated as a mechanism axis. They span task-strong, balanced, and

bias-strong task-to-sensitive ratios, so the audit tests mechanism shift rather than a single density perturbation.

Environment	β_x	β_s	β_m	Regime
E_1	1.5	0.2	0.1	task-strong
E_2	1.0	0.4	0.2	task-moderate
E_3	0.7	0.7	0.3	balanced
E_4	0.4	1.0	0.3	bias-moderate
E_5	0.2	1.5	0.2	bias-strong

Table 8: MD-SCM parameters used to construct the five training environments.

The environments vary the task-to-sensitive ratio while recalibrating base density, so environment shift does not become a trivial density shift.

E.2 Datasets and preprocessing

The real-data audit uses Pokec-z (67,796 nodes, 617,824 undirected edges), Pokec-n (66,569 nodes, 584,199 edges), German (1,000 nodes, 22,242 edges), and Credit (30,000 nodes, 1,436,858 edges). The loader symmetrizes the edge list, removes self-loops, normalizes node features, and uses the declared binary sensitive attribute. Each seed uses a stratified 60/20/20 train/validation/test split; the five seeds are 42, 123, 456, 789, and 2025. These counts describe the released benchmark preprocessing and are not claims about an unprocessed source dump.

E.3 Selection and reporting protocol

The validation search covers learning rates $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$, $\lambda \in \{0.05, 0.1, 0.2\}$, and $K \in \{3, 5, 10\}$. The MIC ambiguity gate is the validation 85th percentile with $B = 100$ bootstrap repetitions; the reported magnitude sensitivity uses $\tau_{\text{fair}} \in \{0.05, 0.10, 0.15\}$. The final defaults are Adam with learning rate 5×10^{-4}, $K = 5$, $\lambda = 0.1$, $\gamma = 0.01$, 500 epochs, and spectral-normalization coefficient 2.0. All choices are made within each training seed using the validation split; no test seed is used for selection. Paired comparisons are reported over the five held-out test seeds. Hardware and dependency details are intentionally compressed because they do not support a scientific claim.

F Supplementary Scope Audit and Full Results

After checking claim mechanics, we ask whether the reported advantage survives broader comparators. Tables 9 and 10 separate capability scope from outcome performance.

Method	EG	SI	ME	SB	AB	MP	LB
NIFTY	–	○	–	–	–	–	–
EDITS	–	○	–	–	–	–	–
FairGNN	–	–	–	–	–	–	–
DAB-GNN	○	–	–	–	–	–	–
FairINV	–	–	○	–	–	–	–
ELEGANT	–	–	–	○	–	–	–
SFG	–	–	–	○	○	–	–
Ours	✓	✓	✓	✓	✓	✓	✓

Table 9: Descriptive scope audit. EG=edge-generation decomposition; SI=selective intervention; ME=mechanism environments; SB=simultaneous bound; AB=actionability/abstention; MP=matched pairs; LB=lower bound. Columns use our intervention-specific definitions, not general capability rankings. ✓ denotes this scope, ○ adjacent scope, and – no corresponding claim. SFG selects training nodes; MIC qualifies post-training interventions.

Method	Pokec-z Δ_{SP}/W5	Pokec-n Δ_{SP}/W5	German Δ_{SP}/W5	Credit Δ_{SP}/W5
GCN	8.97/.298	9.08/.287	10.18/.299	12.58/.445
FairGNN	3.15/.163	3.96/.182	7.51/.252	14.36/.373
EDITS	3.91/.149	4.02/.179	6.92/.252	4.12/.179
FairINV	2.60/.114	2.80/.148	4.05/.173	3.41/.142
FMED	4.24/.142	4.46/.173	6.76/.206	3.75/.160
DAB-GNN	3.42/.158	3.60/.156	5.49/.213	4.50/.188
ELEGANT	2.67/.125	3.53/.151	5.45/.196	3.17/.165
Ours	**.98/.052**	**.91/.049**	**2.87/.078**	**2.18/.061**

Table 10: Full disparity and tail view of the main comparison. The complete seed-level CSV is in `codey/results/main_seed_metrics.csv`.

Scope breadth is not itself performance evidence. The outcome table supports lower disparity under the reported convention, while Table 1 remains authoritative for same-protocol utility.

Method	Pokec-z	Pokec-n	German	Credit
DAB-GNN	67.38/3.42/.158	66.26/3.60/.156	71.31/5.49/.213	76.67/4.50/.188
ELEGANT	66.63/2.67/.125	66.66/3.53/.151	70.74/5.45/.196	76.99/3.17/.165

Table 11: Previously reported ACC/Δ_{SP}/W5-CD entries moved from Table 1 to retain its size when adding closer comparisons. These values are unchanged.

G Diagnostics and Reproducibility

The implementation reports width, bound magnitude, gate failures, unresolved cases, and qualified nodes. No-witness and Gap-fail cases are included in abstention. The mutually exclusive proportions below distinguish unresolved cases from resolved gate failures; they measure fixed-mask qualification, not execution or nominal coverage.

The main paper reports the stress-regime outcomes; the following table records the calibration and solver defaults without duplicating those results.

Field	Default
Bootstrap repetitions B	100
Counterfactual samples K	5
τ_{amb}	validation 85th percentile
τ_{fair}	0.10 (reported sensitivity: .05/.10/.15)
Optimizer	Adam, lr 5×10^{-4}
Epochs / seeds	500 / 5

Dataset	Visited cells	Max endpoint gap	q^*
Pokec-z	1600	.009	5
Pokec-n	1600	.009	5
German	1600	.009	5
Credit	1600	.009	5

Table 12: Reproducibility defaults and MIC solver audit; endpoint gaps are summarized over resolved nodes, and unresolved nodes abstain.

Among resolved nodes, the maximum endpoint gap is .009, below the .01 tolerance. Across 13,559 Pokec-z test nodes, on the fixed symmetric mask before failure-induced suppression, mutually exclusive proportions are .0105 without a witness, .0105 without bracket closure, .278 resolved but failing the gates, and .701 qualified. Abstention is .299; unresolved nodes always abstain. These are qualification statistics, not realized execution counts. Candidate edges require two qualified endpoints. After one fixed-mask recertification, failure withholds every incident-edge intervention. Such suppression changes the mask: the fixed-mask certificate does not automatically certify the remaining intervention. Qualification is not execution coverage.

G.1 Executable Pipeline

Algorithm 1: Identify–Qualify–Expose.
Input: graph (X, S, Y, A), candidate frame $\mathcal{R}$, seeds, thresholds.
1. Compute c_{ij} and h_{ij} on $\mathcal{R}$; sample dyads and fit the weighted joint MD-SCM. 2. Compute ψ_{ij}^{sens} and ψ_{ij}^{mix} using both inclusion orders; flag validated dyads and construct $\bar{A}^{\text{cf}}$ with $\beta_s = \beta_m = 0$ on them. 3. Train over five mechanism environments, freeze the predictor, then bootstrap ϵ_{KL} and solve conservative compatible-set endpoint brackets. 4. Use conservative endpoint brackets for both gates; unresolved nodes abstain. Node selection preserves coverage only for the already-certified intervention. 5. With the predictor fixed, recertify the fixed symmetric mask once. Failed nodes withhold all incident-edge interventions. This subsequent suppression is outside the reported fixed-mask certificate; no further certification is claimed. Evaluate MIFB on accepted pairs.

H Additional Control Protocols

Closer baselines. SFG/FairSNR cover four datasets; CDFG covers German/Credit with matched splits, preprocessing, seeds, and validation selection. Table 1 supports a utility–fairness trade-off, not uniformly better accuracy.

Matched withholding. Within each seed, the predictor, fitted SCM, and candidate dyads are fixed. Both controls use symmetric masks with equal realized edit counts (.71 fraction), separately from Full. Unresolved cases abstain; post-suppression masks are outside the reported certificate. Table 4 supports selection at matched budgets.

Model-matching control. IPW-SCM only uses the same fitted four-parameter model and plug-in replacement mask, without MIC qualification or multi-environment classifier training. All methods use the same dyad frame, generator information available at training, and soft-adjacency metric. The shifted generator adds $.31c^2$ to its logit, while the fitted four-parameter SCM omits this term. The construction oracle retains $.31c^2$ as task structure while removing the sensitive and mixed terms. Each additional audit uses 200 accepted pairs with statistic mismatch at most .022 and binary oracle-mask separation at least .151. IPW-only and Full share the shifted cohort; original rows retain their own cohort and are not pooled with the additional audits. This deliberate misspecification is an empirical stress test, not an extension of A1 or a real-world causal claim.

Separate counterfactual audit. The CDFG-inspired generator Zhang et al. (2025) fixes non-sensitive features, samples a temporary Bernoulli(.5) sensitive channel, and builds cosine KNN graphs (threshold 0; undirected union) without exposing that channel to predictors. Pokec-z uses $k = 18$; other datasets freeze k from training-graph median degree. All methods share five counterfactual samples per node; no test labels or fitted MD-SCM parameters enter generation. Mean$\pm$SD W5-CD in Ours/FairSNR/FairINV/EDITS order is PZ .052$\pm$.004/.061$\pm$.005/.073$\pm$.006/.084$\pm$.007; PN .049$\pm$.004/.057$\pm$.005/.093$\pm$.008/.100$\pm$.009; GE .078$\pm$.006/.080$\pm$.007/.110$\pm$.010/.142$\pm$.013; and CR .061$\pm$.005/.068$\pm$.006/.092$\pm$.008/.101$\pm$.009. All test nodes, including abstentions, are compared with the original factual graph.

Component controls. In Eq. (11), removing adversarial or counterfactual loss sets $\gamma = 0$ or $\lambda = 0$. Single-environment training uses E3 with classification weight one for 500 epochs. Without main-path editing, classification and adversarial losses use the original graph; five auxiliary environments contribute equally to counterfactual loss ($\lambda = .1$, $\gamma = .01$). Each environment has weight .2; auxiliary graph editing remains active. Broader controls (ACC/Δ_{SP}/W5-CD): no MD-SCM 65.82/3.52/.141; random edges 67.98/2.31/.118; always intervene 67.98/1.87/.142.

Edge-mask percentile sensitivity. Main-table edge selection uses scores strictly above the 90th percentile of validation candidate edges; ties at the threshold are excluded. This is neither 90% edge editing nor a guaranteed final 10% edit fraction. A separate Pokec-z audit reports ACC/Δ_{SP}/W5-CD of 68.33/.98/.054, 68.35/.97/.053, and 68.36/.95/.052 at percentiles 85, 90, and 95. Ranges are .03 accuracy points, .03 disparity points, and .002 W5-CD. The 90th-percentile entry is a separate repeated evaluation, not the original Full run. These points show local stability, not universal optimality; edge percentiles differ from MIC and MIFB thresholds.