GUIC: Certified Graph Unlearning with Individual Fairness Guarantees

Zichong Wang[1], **Tongliang Liu**[2], **Wenbin Zhang**[1*]

[1] Florida International University, Florida, United States
[2] University of Sydney, New South Wales, Australia

Abstract

Graph unlearning, motivated by emerging right to be forgotten regulations, seeks to remove the influence of specific subsets of data (*e.g.*, noisy, poisoned, or privacy-sensitive data) from pre-trained graph learning models. While much attention has focused on the technical feasibility of unlearning, its implications for fairness remain largely unexamined. To address this critical gap, this paper introduces GUIC, the first framework that jointly ensures certified unlearning and individual fairness in graph-based models, introducing a novel perspective on responsible model updates in graph unlearning. Specifically, GUIC employs a principled distance-based rule to pinpoint individual biases arising from node removals and applies a computationally efficient certificate-driven update, preserving the local Lipschitz constraints crucial for individual fairness. Different from computationally expensive retraining or fairness-regularized optimization methods, GUIC provides a lightweight yet verifiable alternative with theoretical fairness guarantees. Experiments on multiple real-world datasets show that our method consistently surpasses existing approaches across key performance metrics.

Introduction

Graph learning models have achieved impressive success across a range of applications, including social media analysis (Liu et al. 2021), financial forecasting (Wang et al. 2025d), and criminal justice (Zhou et al. 2024), by leveraging their capacity to learn intricate structural patterns from large-scale graph-structured data. However, this reliance on extensive and often sensitive personal data raises serious privacy concerns (Wang et al. 2025i). In response, regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) (Politou et al. 2018; Ruan et al. 2021) have introduced the legal notion of the *right to be forgotten*, which allows individuals to request the deletion of their personal information from data repositories and trained models. These privacy mandates have created an urgent need for graph unlearning methods, techniques that can efficiently remove the influence of specific data points from learned models without requiring full retraining (Bourtoule et al. 2021).

Despite the development of numerous graph unlearning approaches to meet this demand (Gupta et al. 2021; Chen et al. 2022; Pan, Chien, and Milenkovic 2023), these methods may inadvertently introduce or amplify biases inherited from training data. For instance, in loan approval scenarios, removing a successful loan applicant's record might unintentionally lower the predicted creditworthiness of similar individuals who would otherwise qualify for loans, unfairly penalizing them solely based on their similarity to the deleted case. While traditional graph fairness literature has not adequately addressed this issue, a recent shift in perspective has emerged, viewing unlearning not only as a privacy-preserving mechanism but also as a potential fairness intervention (Oesterling et al. 2024; Chen et al. 2025). Methods like FMD (Chen et al. 2023) use influence functions to identify and remove training data that contribute to counterfactual bias, suggesting a promising new direction.

However, existing methods leverage unlearning primarily as a tool to address biases explicitly embedded within biased samples, while overlooking a critical source of bias: the unlearning process itself. This bias does not stem from the content of the removed data but rather from structural disruptions introduced by the act of unlearning. Specifically, two key mechanisms: neighbor loss and constraint weakening, contribute to this unintended consequence. Neighbor loss occurs when deleting a node disrupts fairness constraints that previously linked it with similar individuals, leaving the remaining nodes without essential fairness-enforcing neighbors. Constraint weakening arises because fewer node pairs remain available to uphold fairness constraints, thereby diminishing their overall strength. Together, these mechanisms cause the optimization process to prioritize predictive performance at the expense of fairness, leading to divergent outcomes for otherwise similar individuals and violating the fundamental principle of individual fairness.

To account for the bias introduced by the act of unlearning, we propose a novel approach that embeds fairness considerations into the core of the unlearning framework. This ensures that fairness guarantees are preserved even when handling legitimate data deletion requests, an area that remains largely unexplored and poses three distinct challenges: **i) Difficulty in fairly forgetting graph-structured data.** Graph data encodes information in both node attributes and their structural relationships. Fair unlearning therefore requires

*Corresponding author.

jointly removing biased influence from features and connections. This dual dependency complicates the process, as bias may reside in either component or emerge from their interaction, making it challenging to ensure fairness is preserved while honoring deletion requests. **ii) Challenge of certifying fairness-preserving unlearning.** While some certified unlearning methods provide accuracy guarantees, they rarely address fairness constraints, particularly individual fairness, which requires pairwise consistency between similar individuals. Ensuring that these pairwise fairness relationships are upheld after unlearning is nontrivial, especially when the deletion disrupts existing fairness-enforcing pairs. As a result, verifying that fairness constraints still hold post-unlearning presents a distinct and largely unaddressed challenge. **iii) Challenge of quantifying the individual-level bias introduced by unlearning.** After a node is removed, the model's weights typically undergo only minor adjustments; however, even these slight changes can significantly affect predictions for neighboring nodes similar to the deleted one. Consequently, accurately quantifying the bias introduced by each deletion becomes crucial yet challenging, as it requires explicitly measuring the prediction shifts among multiple pairs of similar nodes. This necessitates precisely tracking subtle changes in disparity and clearly linking these small perturbations to measurable changes in individual fairness outcomes.

To address these challenges, this paper proposes **Graph Unlearning with Individual Fairness Certificates (GUIC)**, a novel framework that enables fair AI decision-making in graph-based models while honoring data deletion requests. *To the best of our knowledge, GUIC is the first work that combines certified graph unlearning with provable individual fairness.* Specifically, GUIC introduces a distance-based criterion to identify bias introduced during unlearning and employs a closed-form, certificate-driven update that removes the target node while constraining performance shifts within a subspace that preserves the local Lipschitz bound. This approach provides a lightweight alternative to full retraining, maintains individual fairness across deletion requests, and enables independently verifiable, per-request guarantees. Formal analysis further establishes bounded error in both predictive performance and fairness loss. The main contributions are as follows:

- A distance-based rule is designed to precisely identify individual bias introduced by node removal during the unlearning process, enabling targeted mitigation of fairness loss.
- Building on this insight, GUIC is introduced as a novel, training-free framework that simultaneously ensures individual fairness and certified unlearning.
- Extensive experiments on three real-world datasets validate the effectiveness of GUIC, demonstrating its superiority over existing state-of-the-art unlearning and fairness-aware methods.

Related Works

Graph Unlearning

Graph unlearning is a recent concept designed to ensure the right to be forgotten by selectively removing nodes from pre-trained graph-based models without requiring full model retraining (Nguyen et al. 2022). Existing approaches to graph unlearning are broadly categorized into exact unlearning and approximate unlearning (Said et al. 2023). Exact unlearning algorithms aim to update the model after a data removal request such that the updated model exactly matches a model retrained from scratch on the modified dataset (Ullah et al. 2021). For example, GUIDE (Wang, Huai, and Wang 2023) employs balance-guided graph partitioning, enabling efficient retraining of smaller model ensembles while preserving critical graph structural information. In contrast, approximate unlearning methods allow small differences between the updated model and the fully retrained one. The goal is to minimize the influence of deleted data points while accepting a certain level of discrepancy between models. For instance, PUMA (Wu, Hashemi, and Srinivasa 2022) offsets the impact of data deletions through constrained optimization that optimally reweights remaining data points, thus closely approximating the results of retraining without complete model reinitialization. However, most of them overlook fairness considerations, despite the increasing deployment of graph models in high-stakes domains such as finance, healthcare, and criminal justice, where neglecting fairness can lead to unintended bias and ethical concerns. This critical oversight underscores the pressing need for fairness-aware graph unlearning methods.

Fair Graph Learning

Designing fairness-aware graph algorithms has been attracting increasing attention in recent years (Wang and Zhang 2025; Chen et al. 2024; Zhang et al. 2025; Wang et al. 2025c; Wang, Yin, and Zhang 2025). However, most existing fairness-aware approaches rely on pre-processing or training-based in-processing strategies, making them unsuitable for directly achieving fairness in the graph unlearning context without retraining the entire model. On the other hand, a few recent studies (Oesterling et al. 2024; Chen et al. 2023) have begun leveraging unlearning explicitly as a tool to mitigate bias without requiring full retraining. For instance, Kose et al. (Kose, Mateos, and Shen 2025) proposed principled, training-free strategies for detecting and removing bias-inducing node features and structural components from pretrained graph models. Similarly, FROG (Chen et al. 2025) optimizes graph structures to mitigate potential bias amplification during the unlearning process. Nonetheless, existing methods still primarily focus on explicitly removing biased data points, rather than directly addressing the fairness implications stemming from genuine data deletion requests, thus highlighting a critical gap in current fairness-aware unlearning research.

Different from the above works, our work takes a fresh perspective by explicitly addressing fairness implications directly resulting from genuine unlearning requests. It introduces a principled method to precisely identify and mitigate individual-level bias that emerges during graph unlearning, enabling certified unlearning guarantees without compromising individual fairness constraints.

Preliminary

In this section, we first introduce notations, and then formally define the concept of ϵ-δ certified unlearning.

Notations. Consider an undirected attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $|\mathcal{V}| = n$ denotes a set of n nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of edges as unordered pairs of nodes, and each node v_i has an associated D-dimensional feature vector $x_i \in \mathbb{R}^D$, and collectively these features form the node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$. The structure of the graph is captured by an adjacency matrix $\mathbf{A}$, where each element $\mathbf{A}_{i,j}$ equals 1 if nodes v_i and v_j share an edge, and 0 otherwise. Consistent with existing approaches, we concentrate on binary classification tasks at the node level. Additionally, nodes are categorized into labeled and unlabeled sets. The labeled nodes $\mathcal{L} \subseteq \mathcal{V}$ have known labels, forming the set $\mathcal{Y} = \{y_1, y_2, \ldots, y_{N_L}\}$, where N_L denotes the count of labeled nodes. Unlabeled nodes are gathered in the set $\mathcal{U} \subseteq \mathcal{V}$, and their predicted labels are represented as $\hat{y}_i$ for node $v_i \in \mathcal{U}$. It is important to note that the union of the labeled and unlabeled vertex sets equals the entire set of vertices in the graph, *i.e.*, $\mathcal{L} \bigcup \mathcal{U} = \mathcal{V}$.

ϵ-δ Certified Unlearning. Let $\widetilde{\mathcal{G}} \subset \mathcal{G}$ denote the graph dataset obtained after removing unlearned set $\mathcal{G}_U$, from $\mathcal{G}$. Given the original learning process $\mathcal{A}(\mathcal{G})$ parameterized by $\omega \in \mathcal{H}$ that minimizes the empirical risk on $\mathcal{G}$, where $\mathcal{H}$ denotes the hypothesis space. An ϵ-δ certified unlearning algorithm aims to produce an updated model statistically indistinguishable from a model retrained from scratch on $\widetilde{\mathcal{G}}$. Formally, given $\exists\ (\epsilon, \delta) > 0$, an unlearning process $\mathcal{U}$ applied to $(\mathcal{G}, \mathcal{G}_U, \mathcal{A}(\mathcal{G}))$ satisfies ϵ-δ certified unlearning if and only if for all measurable subsets $\mathcal{T} \subseteq \mathcal{H}$:

$$\begin{cases} \mathbb{P}\left(\mathcal{U}(\mathcal{G}, \mathcal{G}_U, \mathcal{A}(\mathcal{G})) \in \mathcal{T}\right) & \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(\widetilde{\mathcal{G}}) \in \mathcal{T}) + \delta \\ \mathbb{P}(\mathcal{A}(\widetilde{\mathcal{G}}) \in \mathcal{T}) & \leq e^{\epsilon}\mathbb{P}\left(\mathcal{U}(\mathcal{G}, \mathcal{G}_U, \mathcal{A}(\mathcal{G})) \in \mathcal{T}\right) + \delta \end{cases} \tag{1}$$

where $\mathbb{P}$ denotes the probability measure over all sources of randomness in the learning and unlearning procedures.

Methodology

Quantifying Individual Graph Unfairness

This subsection examines how individual bias arises during graph unlearning and introduces a quantitative measure to inform mitigation. Figure 1 illustrates a toy loan scenario with three applicants (d_1, d_2, and d_3), each pairwise similar in the input space (distance of 5 units). To ensure similar individuals receive similar outcomes, individual fairness constraints are applied during training to keep interest rate differences within a fixed threshold. Without fairness constraints (orange region), performance-driven objectives would assign different rates, 5% for d_1, 7% for d_2, and 9% for d_3. With fairness enforced (yellow region), all predictions converge to 7%. However, when applicant d_2 requests data deletion, a fast unlearning process removes its gradient contributions, disrupting two of the three original fairness constraints (green region). The lone remaining constraint between d_1 and d_3 is too weak to maintain alignment, leading to a prediction drift: 6% for d_1 and 8% for d_3. This divergence occurs despite no change in input similarity, highlighting the fairness risk introduced by unlearning.

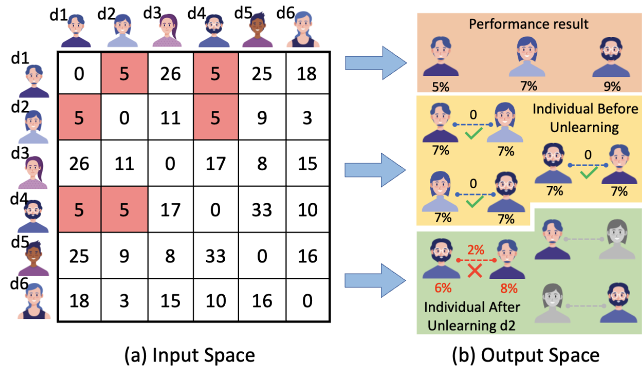

Figure 1: A toy example of the disparity of individual fairness during the unlearning process in a loan approval system.

Motivated by this observation, two key mechanisms are identified through which node removal introduces bias into a trained graph model: i) Neighbor loss. Individual fairness constraints are typically pairwise, when two nodes lie within a certain distance threshold (*i.e.*, similar individuals), a regularization term encourages similar predicted outcomes. Removing a node breaks all associated pairs, causing its neighbors to lose critical fairness, enforcing connections. As a result, subsequent optimization for these remaining nodes becomes increasingly performance-driven, introducing bias as accuracy begins to outweigh fairness. ii) Constraint weakening. Graph unlearning retains the original training objective but applies it to the reduced dataset, effectively optimizing on a reshaped loss surface rather than fully retraining. Although the fairness regularization term is preserved, its relative influence declines due to the removal of nodes and their associated edges that previously encoded fairness constraints. This shift, driven by changes in node representations after deletion, causes optimization to favor data encoding over fairness smoothness, thereby amplifying disparities. To formally quantify the combined effect of these mechanisms, we define the concept of individual graph unlearning bias, which measures how much the total individual bias increases when one or more nodes are deleted. Specifically, it compares the output space differences among all pairs of similar nodes before unlearning to the corresponding sum after unlearning. Formally, this notion is captured in Definition 1 below:

Definition 1 (Individual Graph Unlearning Bias). Let $\mathcal{A}(\mathcal{G})$ be the prediction model trained on the original graph $\mathcal{G}$, and let $\mathcal{U}(\mathcal{G}, \mathcal{G}_U, \mathcal{A}(\mathcal{G}))$ return the updated parameters $\widetilde{\omega}$ after removing the unlearned set $\mathcal{G}_U \subset \mathcal{G}$, and $\widetilde{\mathcal{G}}$ denotes the retained dataset. For every reference node $v_i \in \mathcal{L}$, we denote by $\mathcal{N}(v_i)$ (resp. $\mathcal{N}'(v_i)$) the set of its similar individual (*i.e.*, Dis(v_i, v_j) $\leq \varphi$) before (resp. after) unlearning. The individual bias induced by the unlearning step is defined as:

$$\begin{aligned}\Delta_{bias} &= \sum_{v_i \in \mathcal{L}} \left[\nabla \ell_f(\omega, \mathcal{G}) - \nabla \ell_f(\widetilde{\omega}, \widetilde{\mathcal{G}}) \right] \\ &= \sum_{v_i \in \mathcal{L}} \sum_{v_j \in \mathcal{N}(v_i)} \|\hat{y_i} - \hat{y_j}\|_F^2 - \sum_{v_i \in \mathcal{L}} \sum_{v_j \in \mathcal{N}'(v_i)} \|\hat{y_i} - \hat{y_j}\|_F^2 \\ &= \sum_{v_i \in \mathcal{L}, v_j \in \mathcal{N}(v_i)} L^S_{i,j} Y_i^T Y_j - \sum_{v_i \in \mathcal{L}, v_j \in \mathcal{N}'(v_i)} L^{S'}_{i,j} Y_i^T Y_j\end{aligned} \tag{2}$$

where the similarity Laplacian L^S and $L^{S'}$ are constructed from the neighborhoods $\mathcal{N}(\cdot)$ and $\mathcal{N}'(\cdot)$, respectively. This formal definition quantitatively measures the total increase in output space difference among all pairs of nodes whose input distances are within the threshold φ. To sum, if similar individuals' outputs diverge after node removals, this metric immediately increases; if their predictions remain similar, the metric remains near zero. Therefore, Definition 1 directly operationalizes the individual fairness principle (Dwork et al. 2012), ensuring that similar individuals receive similar algorithmic outcomes after unlearning.

However, directly optimizing the fairness loss from Definition 1 would require adding it to the objective and retraining the entire model, which is impractical in the graph unlearning setting. To this end, given the predictions in GNNs are derived from linear mappings of node representations, we can precisely estimate the representation drift induced by node removals and perform a small compensatory parameter-space update. This compensation keeps representation shifts within a theoretically derived bound, thereby practically estimating the fairness loss described in Definition 1 without retraining the model from scratch. Specifically, Consider a graph learning model where predictions are obtained via a linear mapping of node representations, *i.e.*, $\hat{y}_i = z_i^\top w$, with z_i denoting the representation of node i and w the model's weight vector, the bias increment defined by Definition 1 can be expressed entirely in terms of the representation shifts $z_i - z_j$ induced by node deletions. In other words, quantifying changes in node representations directly allows us to measure the increase in individual bias defined in Definition 1. When nodes are removed, their absence perturbs the representations of neighboring nodes through the message-passing mechanism, causing embedding shifts that directly influence individual fairness.

To this end, we first analyze precisely how node removals perturb the embeddings of neighboring nodes, quantifying how these perturbations propagate layer-by-layer within the GNN message-passing framework. We take an example with Simple Graph Convolution (SGC) model (Chien, Pan, and Milenkovic 2022), which removes intermediate nonlinearities, thereby simplifying the analysis to repeated propagation through a single linear transformation. Formally, the embedding update rule is given by $\mathbf{H}^{(K)} \triangleq \mathbf{P}^K \mathbf{X}$, where $\mathbf{W}$ denotes the learned weight matrix, and $\mathbf{P} = \mathbf{D}^{-1}\tilde{\mathbf{A}}$ is the left-normalized adjacency matrix with self-loops $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, and $\mathbf{D}$ represents the degree matrix corresponding to $\tilde{\mathbf{A}}$. Building upon this formulation, we explicitly link representation perturbations due to node removals to the resulting bias increments. Specifically, when removing node v_r with feature vector $\mathbf{x}_r$ and diffusion profile $\mathbf{P}_r^K$, assuming uniformly bounded loss gradient derivatives, the bias increment $\Delta\text{bias}^{(w)}$ induced by this node removal is bounded as follows:

$$\left|\Delta^{(w)}_{\text{bias}}\right| \le \frac{d'_{\max}\sqrt{N}}{2\lambda(m-1)} \left[2c\lambda + (c\gamma_1 + \lambda c_1)\|\mathbf{x}_m\|\|\mathbf{P}^K_{:m}\|\right] + \frac{N d'_{\max}}{2\lambda^2(m-1)^2} \left[2c\lambda + (c\gamma_1 + \lambda c_1)\|\mathbf{x}_m\|\|\mathbf{P}^K_{:m}\|\right]^2 \tag{3}$$

where N denotes the number of training nodes, λ the regularization weight, and $d'_{\max}$ represents the maximum degree in the similarity graph after unlearning.

Fair Graph Unlearning Process

In this subsection, we present a practical method to explicitly estimate and mitigate the individual bias introduced during graph unlearning. Central to this theoretical relationship is the concept of the local Lipschitz constant, denoted as $L_{\text{loc}}(v_i)$, which measures in a worst-case scenario, the maximum ratio of prediction changes to input distances among all nodes similar to node v_i (within distance ρ). Intuitively, a larger $L_{\text{loc}}(v_i)$ indicates higher model sensitivity around node v_i, implying that even minor perturbations in the input can significantly widen prediction gaps between similar individuals, thereby undermining individual fairness. Leveraging the explicit bound, GUIC employs a lightweight, certificate-backed update procedure to address this concern in two phases. First, GUIC efficiently estimates how much deleting a node v_r would increase its local Lipschitz constant and consequently, individual bias, without requiring model retraining. Next, GUIC neutralizes this bias increment via a carefully calibrated, single-step weight adjustment. This step simultaneously removes the predictive influence of the deleted node, restores the smoothness constraints essential for maintaining individual fairness, and injects precisely calibrated Gaussian noise $\mathbf{b} \sim \mathcal{N}(0, \sigma^2 I)$ to facilitate certified unlearning. This calibrated noise masks residual node-specific information, thus achieving rigorous (ε, δ)-certified unlearning guarantees. Each component of the procedure is described below, illustrating how GUIC preserves individual fairness throughout the node unlearning process.

Beginning with the bias estimation phase, consider the scenario where a node v_r requests to be forgotten. Building on Definition 1, deleting node v_r will affect neighbor node representations through the message-passing chain. Specifically, the feature vector $\mathbf{x}_r$ is zeroed out, and the r-th row and column of the adjacency matrix are removed. Such perturbations alter all entries of $\mathbf{Z}$, leading to increased local Lipschitz ratios and thus elevated individual bias. GUIC quantifies this bias increment using influence functions, assessing the first-order score changes for each neighbor v_j without retraining. This procedure computes influence values $\delta r \to j$ via a single conjugate-gradient solve:

$$\delta_{r \to j} = -\nabla_\omega f_\theta(\mathbf{x}_j)^\top H_\theta^{-1} \nabla_\omega \ell(f_\theta(\mathbf{x}_r), y_r) \tag{4}$$

Following bias identification, GUIC then creates a proxy node v_p to fill the structural gap left by v_r. The proxy has two goals: i) avoiding reintroducing the predictive influence of v_r (thus fully forgetting it); ii) preserving the local similarity structure so that the fairness-enforcing regularization remains effective. Formally, let $g_r = \nabla_{\mathbf{x}^r}\ell(f\theta(\mathbf{x}_r), y_r)$ and $u_r = g_r/||g_r||_2$ denote the unit direction of strongest influence. GUIC removes precisely this predictive direction and retains only its orthogonal component, constructing the proxy representation as:

$$\mathbf{x}_p = (I_F - u_r u_r^\top)\mathbf{x}_r + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 I_F) \tag{5}$$

The projection matrix $I_F - u_r u_r^\top$ effectively erases the original performance direction, and the added Gaussian perturbation $\boldsymbol{\xi}$ prevents adversaries from reconstructing the original data. As the noise variance σ_g^2 is small relative to $\mathbf{x}_r$, the proxy remains close to the deleted node's embedding in non-performance dimensions. Subsequently, GUIC updates the similarity matrix S' by replacing the r-th row and column with entries computed from the proxy node, preserving local graph connectivity:

$$S'_{rj} = \kappa(\mathbf{x}_p, \mathbf{x}_j) = S'_{jr}, \quad \forall j = 1, \dots, N \tag{6}$$

where $\kappa(\cdot)$ denotes the kernel function.

Building on this, we establish adjusted Laplacian is $\hat{L}_S = D' - S'$ from the proxy-updated similarity matrix, where $D' = \text{diag}(d')$ and $d'_j = \sum_k S'_{jk}$. This adjustment maintains the structural integrity necessary for effective fairness regularization, with the maximum degree $d'_{\max}$ changing only slightly, proportionally to $O(\sigma_g)$.

After repairing the local graph structure, GUIC updates the model weights through a single-step procedure without retraining. Let ω denote all trainable parameters of the model. We decompose the one-step parameter update into a prediction-removal component and a fairness-restoration component:

$$\Delta\omega = \Delta\omega_{\text{pred}} + \Delta\omega_{\text{fair}} \tag{7}$$

where both $\Delta\omega_{\text{pred}}$ and $\Delta\omega_{\text{fair}}$ are update directions in the same parameter space of ω.

The prediction-removal vector removes the predictive influence of the deleted node v_r:

$$\Delta\omega_{\text{pred}} = -H_\theta^{-1}\nabla_\omega \ell(f_\theta(\mathbf{x}_r), y_r) \tag{8}$$

To restore the fairness constraints, GUIC first takes the gradient of the fairness regularizer with respect to ω at the proxy node $\mathbf{x}_p$:

$$\tilde{g}_{\text{fair}}(\omega) = \nabla_\omega \hat{R}_{\text{fair}}(\omega; \mathbf{x}_p) \tag{9}$$

which provides the raw fairness-restoration direction in parameter space. We then remove the component of this gradient that is aligned with $\Delta\omega_{\text{pred}}$, yielding a fairness update direction that is approximately orthogonal to the prediction-removal step:

$$\Delta\omega_{\text{fair}} = \tilde{g}_{\text{fair}} - \frac{\langle \Delta\omega_{\text{pred}}, \tilde{g}_{\text{fair}} \rangle}{\|\Delta\omega_{\text{pred}}\|_2^2 + \varepsilon}\Delta\omega_{\text{pred}} \tag{10}$$

Using backtracking line search, GUIC selects an appropriate step length $\eta \in [0, 1]$, ensuring the local Lipschitz constant returns to pre-deletion levels. This gives the intermediate update:

$$\omega' = \omega + \Delta\omega_{\text{pred}} + \eta\Delta\omega_{\text{fair}} \tag{11}$$

The above-described steps maintain individual fairness but by themselves do not confer an (ε, δ)-certified unlearning guarantee. To achieve rigorous certification, GUIC additionally controls the residual training-loss gradient remaining after a node is deleted. GUIC applies a single Newton-based influence update to compress this residual gradient to a computable upper bound ε', then injects Gaussian noise explicitly calibrated according to this bound. Specifically, given nodal representations extracted by the K–layer Simplified GCN, the residual training-loss gradient after node deletion satisfies:

$$\begin{aligned}\left\|\nabla_\omega L(\omega'; \mathcal{G}')\right\| \ \le\ & 2c\lambda + \left(c\gamma_1 + \lambda c_1\right)\big(\|\mathbf{x}_r\|\sin\theta \\ & + \sigma_g\sqrt{D}\big)\,\|\mathbf{P}_{:r}^K\|\end{aligned} \tag{12}$$

where θ is the angle between the deleted node's feature vector $\mathbf{x}_r$ and its associated loss gradient prior to deletion. Choosing Gaussian release noise with variance $\sigma^2 = c_0, \varepsilon'/\varepsilon$, as derived from equation (12), ensures that the resulting mechanism achieves rigorous (ε, δ)-certified unlearning, where $\delta = 1.5\, e^{-c_0^2/2}$. In practice, after computing the bound ε', GUIC draws Gaussian noise and releases the final updated model as:

$$\tilde{\omega} = \omega' + \mathbf{b} \tag{13}$$

where $\mathbf{b}$ is the Gaussian noise vector.

This explicit bounding and noise calibration strategy allows GUIC to inject precisely the minimal amount of noise required. If the noise calibration step were omitted, a critical trade-off would arise: insufficient noise would leak information about the deleted node, violating certification requirements. By explicitly computing and bounding ε' in advance, GUIC ensures minimal yet sufficient noise injection, thereby preserving model utility while rigorously satisfying the formal certified unlearning guarantee. In conclusion, by seamlessly integrating bias estimation, structural repair, lightweight model updates, and precise noise calibration, GUIC efficiently maintains both individual fairness and certified unlearning guarantees without the need for expensive retraining.

Method	Bail			Pokec-z			Pokec-n		
	Accuracy	F1-Score	FNDCG	Accuracy	F1-Score	FNDCG	Accuracy	F1-Score	FNDCG
Original	81.30 ± 3.51	79.31 ± 2.16	74.50 ± 2.08	68.97 ± 3.21	66.71 ± 2.11	68.04 ± 0.20	66.53 ± 3.74	64.10 ± 3.47	64.26 ± 5.34
Retrain	80.30 ± 2.29	78.07 ± 3.19	72.79 ± 1.15	67.97 ± 0.57	66.10 ± 1.98	67.39 ± 1.08	66.13 ± 0.80	63.73 ± 4.60	63.34 ± 4.15
L-CODEC	77.27 ± 1.61	77.35 ± 3.17	56.46 ± 1.09	67.20 ± 2.70	64.98 ± 3.21	43.33 ± 4.23	63.12 ± 1.36	62.51 ± 3.35	41.53 ± 3.08
SGC	77.83 ± 2.59	78.09 ± 1.28	57.22 ± 0.46	67.63 ± 2.92	65.08 ± 2.86	45.66 ± 1.34	63.63 ± 4.77	62.70 ± 2.41	40.10 ± 2.18
Certified	79.80 ± 0.96	78.28 ± 1.05	57.37 ± 0.41	68.73 ± 3.62	63.68 ± 1.32	49.04 ± 1.38	64.37 ± 3.76	63.01 ± 2.57	42.94 ± 1.43
Fair Unlearning	79.47 ± 1.33	78.55 ± 1.15	57.22 ± 1.09	68.09 ± 1.99	64.25 ± 0.26	48.31 ± 1.54	63.73 ± 1.33	63.53 ± 0.47	43.75 ± 3.41
GUIC	78.70 ± 1.90	77.76 ± 0.14	71.69 ± 0.05	67.87 ± 1.61	64.71 ± 1.56	65.81 ± 0.24	63.23 ± 0.91	62.74 ± 1.49	61.06 ± 2.29

Table 2: Comparison of unlearning methods on Bail, Pokec-z, and Pokec-n datasets using Accuracy, F1-Score and FNDCG.

Experiments

Experimental Settings

Datasets. We conducted experiments on three real-world graphs: Bail (Jordan and Freiburger 2015), Pokec-z, and Pokec-n (Takac and Zabovsky 2012). Specifically, the **Bail dataset**(Agarwal, Lakkaraju, and Zitnik 2021) contains information about defendants who were granted bail in U.S. state courts. Each node represents a defendant, and an edge connecting two nodes indicates similarity in their criminal records and demographic details. Race is considered the demographic information in this dataset. The prediction task focuses on classifying the defendants' outcomes. The **Pokec-z** and **Pokec-n** datasets are extracted from a widely used Slovak social network, representing two distinct provincial sub-networks. Nodes correspond to users characterized by attributes such as gender, age, and interests, while edges represent friendships. In both Pokec datasets, the demographic information is region. The prediction task in these datasets involves classifying users' occupational fields. Detailed statistics are provided in Table **??**.

Evaluation Metrics. We use accuracy and F1-scores to evaluate utility performance, where higher scores indicate better prediction results. In addition, we use FNDCG (Zhang et al. 2023) to assess individual fairness in ranking tasks, where higher values indicate smaller ranking changes after unlearning, thus better preserving fairness and utility after unlearning.

Baselines. To benchmark the performance, we compare our approach with five state-of-the-art baselines: Retrain from Scratch, an ideal baseline achieving exact unlearning by completely retraining the model; L-CODEC (Mehta et al. 2022), which efficiently computes the Hessian using selected parameters guided by the Fisher information matrix; SGC (Chien, Pan, and Milenkovic 2022), providing analytical and certified removal of nodes, edges or features from graph neural networks without expensive retraining; Certified (Zhang et al. 2024), efficiently realizing certified unlearning in deep neural networks through approximate parameter updates and calibrated noise injection; and Fair Unlearning (Kose, Mateos, and Shen 2025), a training-free method removing algorithmic biases from pretrained graph models via certified feature removal and a single-step Newton update.

Dataset	Bail	Pokec-z	Pokec-n
# Vertices	18,876	67,797	66,569
# Edges	311,870	882,765	729129
Feature Dimension	18	59	59
Demographic information	Race	Region	Region

Table 1: Summary of the datasets used in the experiments.

Experiment Results

Unlearning Performance. We compare the utility and fairness of GUIC with five baseline methods on the node classification task. Based on the meaning of graph unlearning, we expect the results of a desirable unlearning method to be close to the results of the retrained model. Hence, we use the retrain-from-scratch baseline as a standard for evaluating the unlearning baselines. Table **??** summarizes the comparison results of GUIC with all baselines across three datasets. We observe that GUIC outperforms all baseline methods across all evaluation metrics in most cases. Specifically, in terms of accuracy and F1-score, GUIC shows comparable performance that is reasonably close to the retrained baseline, achieving accuracy values within 1-2 percentage points across all datasets. In terms of fairness, GUIC significantly outperforms all other unlearning methods in individual fairness results, which are substantially closer to the original model performance compared to other methods that achieve much lower values. The substantial fairness degradation in baseline methods occurs because L-CODEC, SGC, and Certified Unlearning optimize purely performance-driven or privacy-driven objectives, while Fair Unlearning treats unlearning as a tool but fails to address bias arising from real unlearning requests. Once the fairness edges incident to the deleted node disappear, their weight updates push the remaining predictions apart. GUIC avoids this limitation through three key mechanisms: the ghost anchor maintains connectivity in the similarity Laplacian so the regularizer continues pulling neighboring nodes together, the dual compensation step aligns performance-removal and fairness-restoration vectors orthogonally to prevent objective cancellation, and the calibrated Gaussian noise meets the (ϵ-δ) guarantee without compromising restored smoothness. In general, GUIC successfully reconciles the competing goals of certified unlearning and individual fairness preservation, delivering utility close to retraining while maintaining fairness substantially

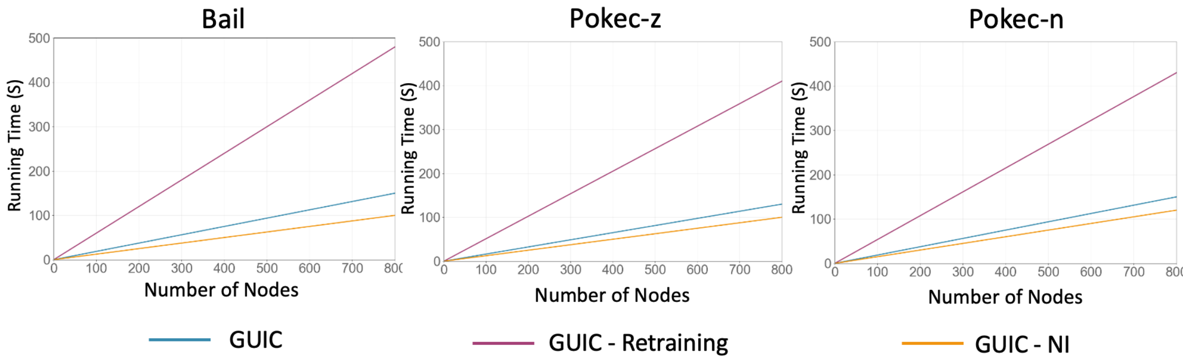

Figure 2: Efficiency comparison of GUIC, GUIC-Retraining and GUIC-NI methods on Bail, Pokec-z and Pokec-n datasets.

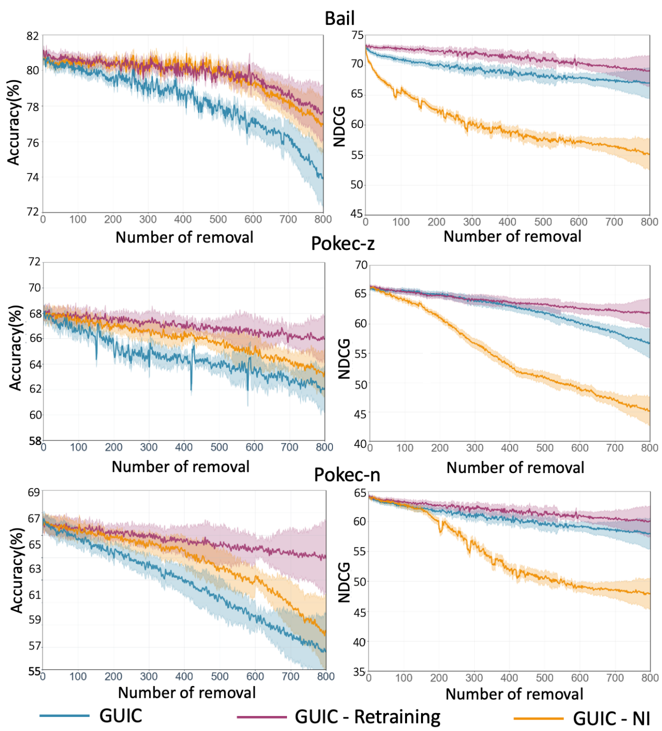

Figure 3: Performance comparison of GUIC, GUIC-Retraining, and GUIC-NI on three datasets. The left panels illustrate accuracy, while the right panels show the fairness metric FNDCG.

above other fast unlearning alternatives.

Ablation Study. We conduct ablation studies to gain insights into the effect of individual fairness component of GUIC on improving fairness. Specifically, we denote GUIC-NI as the variant without the individual fairness component. Figure 3 presents ablation results on Pokec-z, Pokec-n, and Bail datasets, including the accuracy, F1-score, and computational time of unlearning. We observe that GUIC performs better than the ablation variant on both fairness and utility metrics. Specifically, compared with GUIC-NI, GUIC achieves significantly higher individual fairness scores and approaches the performance of the retrained model. This improvement occurs because GUIC explicitly mitigates the individual fairness loss caused by node removal during the unlearning process, thereby preventing the decline in model fairness that typically accompanies node deletion. Additionally, GUIC maintains similar utility performance to both GUIC-NI and the retrained baseline, which demonstrates GUIC's ability to balance model utility and fairness preservation during the unlearning process. These results confirm that the individual fairness component is essential for achieving the dual objectives of effective unlearning and fairness maintenance.

Efficiency study. Following the above experimental setting, we further investigate the computational efficiency of GUIC, GUIC-NI, and retraining methods. Figure 2 presents the runtime comparison across different numbers of nodes on the Bail, Pokec-z, and Pokec-n datasets. We observe that GUIC demonstrates substantial computational advantages over retraining from scratch. Specifically, compared to retraining, GUIC maintains comparable performance while running at least $3\times$ faster across all datasets and scales. While GUIC is slightly slower than GUIC-NI due to the additional individual fairness computations, the runtime difference remains within a reasonable range, demonstrating that the fairness preservation mechanism does not introduce prohibitive computational overhead. These results confirm that GUIC achieves an effective balance between computational efficiency and fairness preservation, making it practically viable for large-scale graph unlearning applications where both speed and fairness are important considerations.

Conclusion

Despite the extensive study of certified unlearning in many graph learning models, individual fairness remains a largely unaddressed challenge. In this paper, we propose GUIC, the first framework designed to achieve certified graph unlearning with explicit guaranties for individual fairness. In addition, we theoretically established formal bounds on the individual fairness loss arising from legitimate deletion requests and introduced a certificate-driven update mechanism that preserves local Lipschitz continuity critical for fairness. Extensive evaluations demonstrate that GUIC consistently maintains fairness and predictive utility throughout the unlearning process, thereby representing an essential advancement toward ethical and responsible graph-based decision-making in real-world applications.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2404039, and the Dissertation Fellowship from Florida International University.

References

Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, 2114–2124. PMLR.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, 141–159. IEEE.

Chen, A.; Rossi, R. A.; Park, N.; Trivedi, P.; Wang, Y.; Yu, T.; Kim, S.; Dernoncourt, F.; and Ahmed, N. K. 2024. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6): 1–23.

Chen, M.; Zhang, Z.; Wang, T.; Backes, M.; Humbert, M.; and Zhang, Y. 2022. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 499–513.

Chen, R.; Yang, J.; Xiong, H.; Bai, J.; Hu, T.; Hao, J.; Feng, Y.; Zhou, J. T.; Wu, J.; and Liu, Z. 2023. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36: 14516–14539.

Chen, Z.; Cheng, J.; Tolomei, G.; Liu, S.; Amiri, H.; Wang, Y.; Nag, K.; and Lin, L. 2025. FROG: Fair removal on graphs. *arXiv preprint arXiv:2503.18197*.

Chien, E.; Pan, C.; and Milenkovic, O. 2022. Certified graph unlearning. *arXiv preprint arXiv:2206.09140*.

Chu, Z.; Wang, Z.; and Zhang, W. 2024. Fairness in Large Language Models: A Taxonomic Survey. *ACM SIGKDD Explorations Newsletter, 2024*, 34–48.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gupta, V.; Jung, C.; Neel, S.; Roth, A.; Sharifi-Malvajerdi, S.; and Waites, C. 2021. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34: 16319–16330.

Jordan, K. L.; and Freiburger, T. L. 2015. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3): 179–196.

Kose, O. D.; Mateos, G.; and Shen, Y. 2025. Unlearning Algorithmic Biases over Graphs. *arXiv preprint arXiv:2505.14945*.

Liu, Y.; Zeng, K.; Wang, H.; Song, X.; and Zhou, B. 2021. Content matters: A GNN-based model combined with text semantics for social network cascade prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 728–740. Springer.

Mehta, R.; Pal, S.; Singh, V.; and Ravi, S. N. 2022. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10422–10431.

Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Oesterling, A.; Ma, J.; Calmon, F.; and Lakkaraju, H. 2024. Fair machine unlearning: Data removal while mitigating disparities. In *International Conference on Artificial Intelligence and Statistics*, 3736–3744. PMLR.

Pan, C.; Chien, E.; and Milenkovic, O. 2023. Unlearning graph classifiers with limited data resources. In *Proceedings of the ACM Web Conference 2023*, 716–726.

Politou, E.; Michota, A.; Alepis, E.; Pocs, M.; and Patsakis, C. 2018. Backups and the right to be forgotten in the GDPR: An uneasy relationship. *Computer Law & Security Review*, 34(6): 1247–1257.

Ruan, W.; Xu, M.; Jia, H.; Wu, Z.; Song, L.; and Han, W. 2021. Privacy compliance: Can technology come to the rescue? *IEEE Security & Privacy*, 19(4): 37–43.

Said, A.; Zhao, Y.; Derr, T.; Shabbir, M.; Abbas, W.; and Koutsoukos, X. 2023. A survey of graph unlearning. *arXiv preprint arXiv:2310.02164*.

Takac, L.; and Zabovsky, M. 2012. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1.

Ullah, E.; Mai, T.; Rao, A.; Rossi, R. A.; and Arora, R. 2021. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, 4126–4142. PMLR.

Wang, C.-L.; Huai, M.; and Wang, D. 2023. Inductive graph unlearning. In *32nd USENIX Security Symposium (USENIX Security 23)*, 3205–3222.

Wang, Z.; Chu, Z.; Doan, T. V.; Ni, S.; Yang, M.; and Zhang, W. 2025a. History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 5(3): 1955–1971.

Wang, Z.; Chu, Z.; Viet Doan, T.; Wang, S.; Wu, Y.; Palade, V.; and Zhang, W. 2025b. 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, 28485–28493.

Wang, Z.; Hoang, N.; Zhang, X.; Bello, K.; Zhang, X.; Iyengar, S. S.; and Zhang, W. 2025c. Towards Fair Graph Learning without Demographic Information. In *The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, 2107–2115.

Wang, Z.; Liu, F.; Pan, S.; Liu, J.; Saeed, F.; Qiu, M.; and Zhang, W. 2025d. fairGNN-WOD: Fair Graph Learning Without Demographics. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*.

Wang, Z.; Narasimhan, G.; Yao, X.; and Zhang, W. 2023a. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *2023 IEEE International Conference on Data Mining (ICDM)*, 638–647. IEEE.

Wang, Z.; Palikhe, A.; Yin, Z.; and Zhang, W. 2025e. Fairness in language models: A tutorial. In *Proceedings of the 34th*

ACM International Conference on Information and Knowledge Management, 6849–6852.

Wang, Z.; Saxena, N.; Yu, T.; Karki, S.; Zetty, T.; Haque, I.; Zhou, S.; Kc, D.; Stockwell, I.; Bifet, A.; et al. 2023b. Preventing Discriminatory Decision-making in Evolving Data Streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Wang, Z.; Wallace, C.; Bifet, A.; Yao, X.; and Zhang, W. 2023c. FG^{2}AN: Fairness-Aware Graph Generative Adversarial Networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 259–275. Springer Nature Switzerland.

Wang, Z.; Wu, A.; Moniz, N.; Hu, S.; Knijnenburg, B.; Zhu, X.; and Zhang, W. 2025f. Towards fairness with limited demographics via disentangled learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*.

Wang, Z.; Yin, Z.; Liu, Z.; Yap, R. H.; Zhang, X.; Hu, S.; and Zhang, W. 2025g. Redefining fairness: A multi-dimensional perspective and integrated evaluation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 336–353. Springer Nature Switzerland Cham.

Wang, Z.; Yin, Z.; Yang, L.; Zhuang, J.; Yu, R.; Kong, Q.; and Zhang, W. 2025h. Fairness-Aware Graph Representation Learning with Limited Demographic Information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Nature Switzerland*.

Wang, Z.; Yin, Z.; Yap, R.; and Zhang, W. 2025i. Ai fairness beyond complete demographics: Current achievements and future directions. In *28th European Conference on Artificial Intelligence*.

Wang, Z.; Yin, Z.; and Zhang, W. 2025. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wang, Z.; Yin, Z.; Zhang, Y.; Yang, L.; Zhang, T.; Pissinou, N.; Cai, Y.; Hu, S.; Li, Y.; Zhao, L.; et al. 2025j. FG-SMOTE: Towards fair node classification with graph neural network. *ACM SIGKDD Explorations Newsletter*, 26(2): 99–108.

Wang, Z.; and Zhang, W. 2025. FDGen: A Fairness-Aware Graph Generation Model. In *Forty-second International Conference on Machine Learning*.

Wu, G.; Hashemi, M.; and Srinivasa, C. 2022. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8675–8682.

Wu, M.; Lin, L.; Zhang, W.; Wang, X.; Yang, Z.; and Hu, S. 2025. Preserving auc fairness in learning with noisy protected groups.

Yin, Z.; Wang, Z.; Palikhe, A.; and Zhang, W. 2025. Uncertain boundaries: A tutorial on copyright challenges and cross-disciplinary solutions for generative ai. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 6853–6856.

Zhang, B.; Dong, Y.; Wang, T.; and Li, J. 2024. Towards certified unlearning for deep neural networks. *arXiv preprint arXiv:2408.00920*.

Zhang, W. 2024a. AI fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3): 386–395.

Zhang, W. 2024b. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22685–22685.

Zhang, W. 2025. Online and Customizable Fairness-aware Learning. *Knowledge and Information Systems*, 1–28.

Zhang, W.; Hernandez-Boussard, T.; and Weiss, J. 2023. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 14611–14619.

Zhang, W.; and Ntoutsi, E. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1480–1486.

Zhang, W.; Wang, Z.; Kim, J.; Cheng, C.; Oommen, T.; Ravikumar, P.; and Weiss, J. 2023. Individual Fairness Under Uncertainty. In *Frontiers in Artificial Intelligence and Applications*, volume 372, 3042–3049. ECAI 2023.

Zhang, W.; and Weiss, J. C. 2022. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, 12235–12243.

Zhang, W.; Zhou, S.; Walsh, T.; and Weiss, J. C. 2025. Fairness amidst non-IID graph data: A literature review. *AI Magazine*, 46(1): e12212.

Zhou, B.; Zhou, H.; Wang, W.; Chen, L.; Ma, J.; and Zheng, Z. 2024. Hdm-gnn: A heterogeneous dynamic multi-view graph neural network for crime prediction. *ACM Transactions on Sensor Networks*.