Rethinking Label Poisoning for GNNs: Pitfalls and Attacks

UNIVERSITY OF COLOGNE

CISPA

Vijay Lingam [1] Mohammad Sadegh Akhondzadeh [2] Aleksandar Bojchevski [2]

[1]CISPA Helmholtz Center for Information Security, [2]University of Cologne

TL;DR

- Label poisoning for GNNs is plagued by serious evaluation pitfalls.
- Existing attacks render ineffective post fixing these fallacies.
- We introduce two new simple yet effective family of attacks that are significantly stronger (up to 8%) than previous strongest attacks.

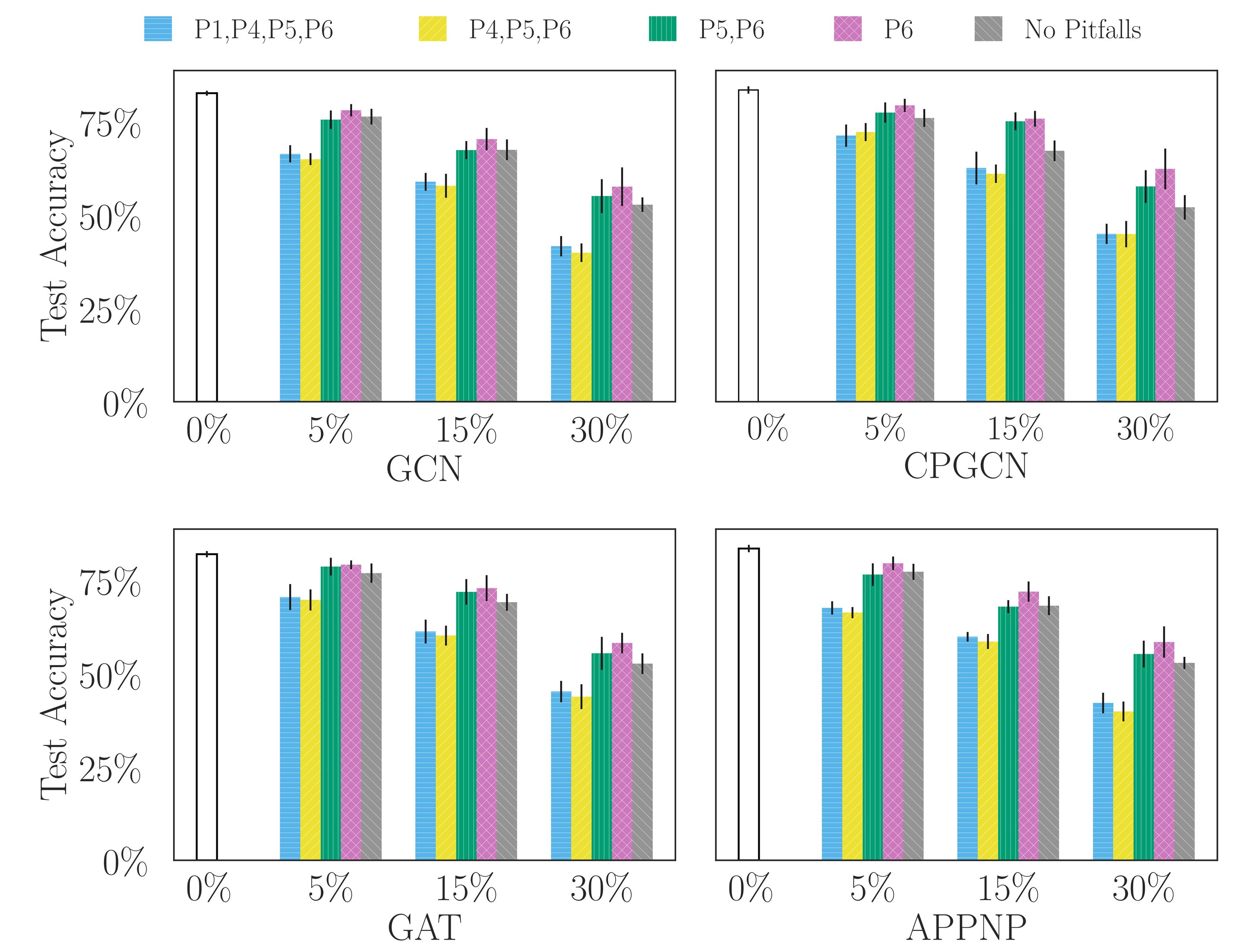
Motivation

GNNs have wide range of applications including critical ones.

Label poisoning poses a distinct threat as training data can be compromised.

Existing attacks are not effective; do better attacks exist?

Existing attacks are not as powerful as claimed



1. **P1**: Large Validation Set
2. **P2**: Missing stdev
3. **P3**: Eval. on undefended models
4. **P4**: Class equalised splits
5. **P5**: Hyper-parameter tuning
6. **P6**: Clean Validation set

Fixing the above pitfalls leads to a massive reduction in LafAK's performance (previous strongest attack).

Threat Model and Baselines

Flip a small fraction of labels to decrease test acc.

Results in a difficult bi-level optimization problem for which we propose different relaxations.

We used two family of attacks as the baselines:

- **Heuristic-based:** Random (**RND**), Degree (**DEG**)
- **Learning-based:** **LP**, LafAK (**LFK**), **MG**

Linear surrogate attacks

Linearize the classifier and compute the optimal weights in closed-form

$$\min_{\boldsymbol{H}\in\{0,1\}^{L\times C}} \mathcal{L}(\boldsymbol{Y}_u, \widehat{\boldsymbol{Y}}_u)$$

$$||\boldsymbol{H} - \boldsymbol{Y}_l||_0 \leq 2\epsilon L$$

$$\widehat{\boldsymbol{Y}}_u = \widehat{\boldsymbol{X}}_u \widetilde{\boldsymbol{X}}_l \boldsymbol{H}$$

$$\boldsymbol{H}\boldsymbol{1}_C = \boldsymbol{1}_L$$

where: $\widetilde{\boldsymbol{X}} = (\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{X}}$ is the closed form solution of LR.

Variant-1: SGC surrogate $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{A}}^2\boldsymbol{X}$

Variant-2: NTK surrogate $\widehat{\boldsymbol{X}}$ = NTK - Kernel

Proposition: LSA closed form solution

Given fixed target labels $\widetilde{\boldsymbol{Y}}_l$, the optimal nodes to poison are the subset of nodes corresponding to the smallest $\lfloor \epsilon L \rfloor$ negative elements of an L-dimensional vector $\boldsymbol{c}$, where the l-th element of $\boldsymbol{c}$ is computed as $c_l = \sum_{ij} Q_{il} P_{lj} R_{ij}$ where $\boldsymbol{Q} = \widehat{\boldsymbol{X}}_u \widetilde{\boldsymbol{X}}_l$, $\boldsymbol{P} = \widetilde{\boldsymbol{Y}}_l - \boldsymbol{Y}_l$, and $\boldsymbol{R} = \boldsymbol{Y}_u$.

Meta attacks

Meta gradients w.r.t. labels by backpropagating through the unrolled inner optimization. The poisoned labels are constructed as follows:

$$\boldsymbol{H} = \mathrm{diag}(\boldsymbol{b})\widetilde{\boldsymbol{Y}} + \mathrm{diag}(\boldsymbol{1}_L - \boldsymbol{b})\boldsymbol{Y}_l$$

where: $\widetilde{\boldsymbol{Y}} = \textit{GumbleSoftmax}(\widetilde{\boldsymbol{Y}}_{\log})$; $\widetilde{\boldsymbol{Y}}_{\log} \in \mathbb{R}^{N\times C}$

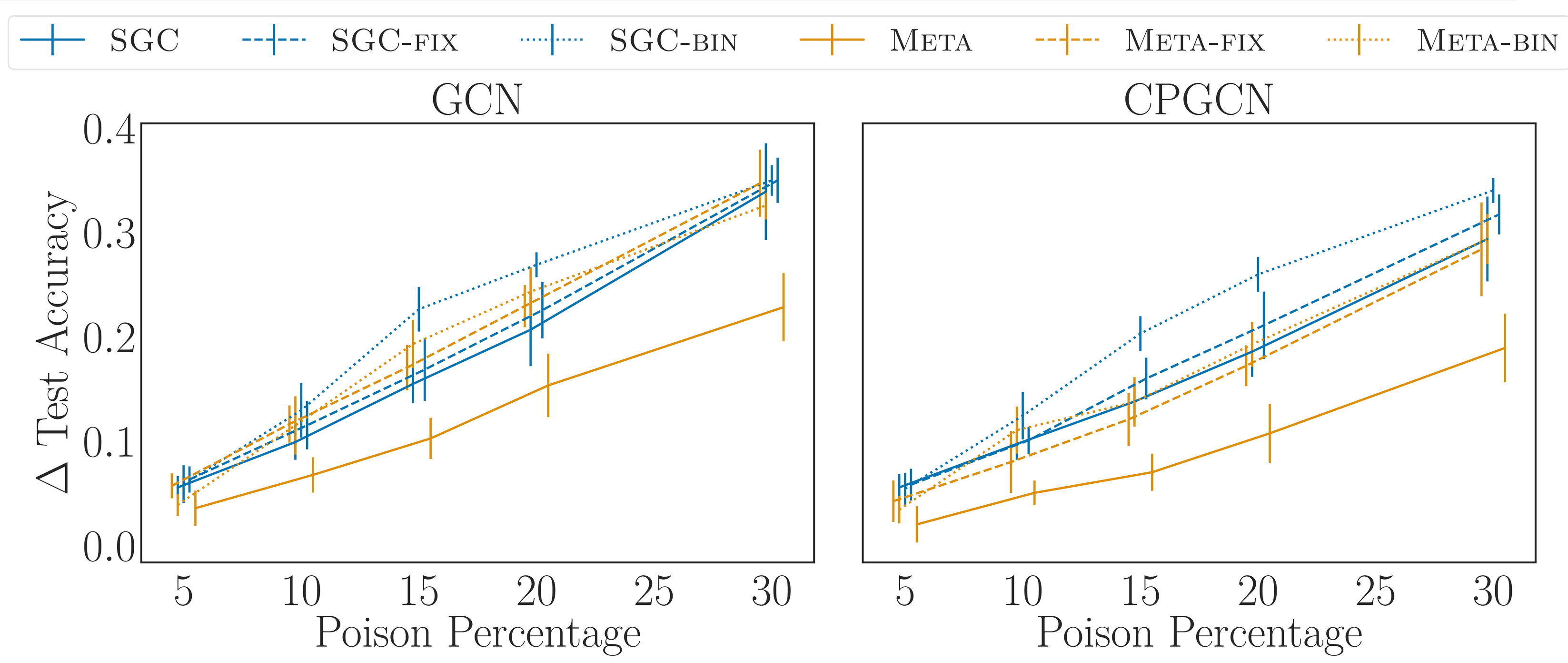
$\boldsymbol{b} = top_k(\tilde{\boldsymbol{b}})$; $\boldsymbol{b} \in \mathbb{R}^N$

Note: since topk is not differentiable, we apply soft-top-k followed by k-subset selection.

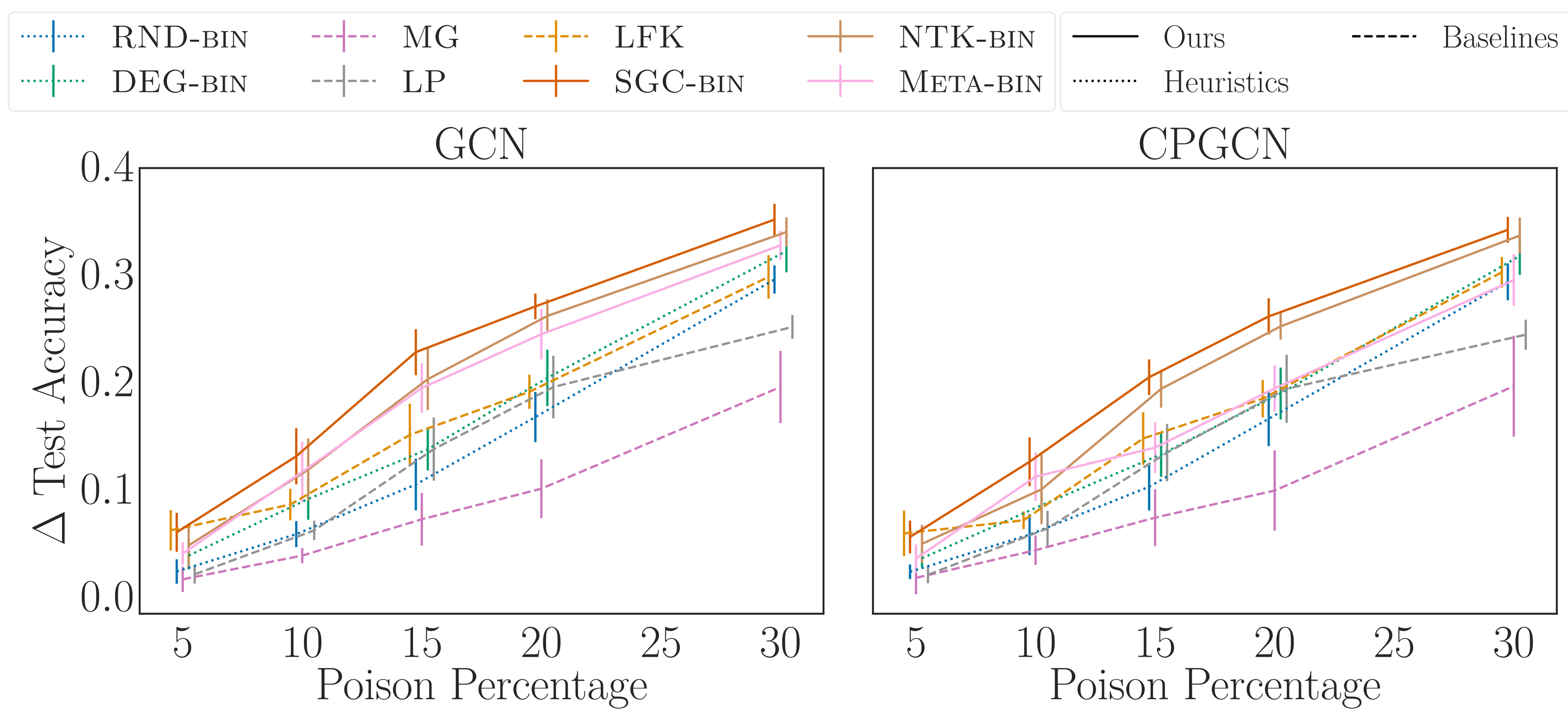
Proposition: Optimality of binary random attack

Let the adversary flip label p to label $q \neq p$ with probability $\frac{\epsilon}{s} \cdot t_{pq}$ and retain label p with probability $1-\epsilon$, where ϵ is the poisoning budget, $t_{pq} \in \{0,1\}$ indicates whether the adversary is allowed to flip p to q, and $s = \sum_{q\neq p} t_{pq}$ is the number of allowed classes. The test accuracy of the Bayes optimal classifier trained on randomly flipped labels is minimized for $s = 1$ (binary flips).

LSA outperforms meta & Binary outperforms multi-label



Our proposed attacks significantly outperform baselines



Key takeaways

- Faithfully simulating the defender is crucial to evaluate the efficacy of an attack.
- Simple label poisoning attacks are surprisingly powerful.
- Our findings highlight the need to further study label poisoning attacks as well as develop defences.