SLES: NetSafe: Towards a Computational Foundation of Safe Graph Neural Networks

PIs: Hanghang Tong and Jingrui He

University of Illinois at Urbana-Champaign

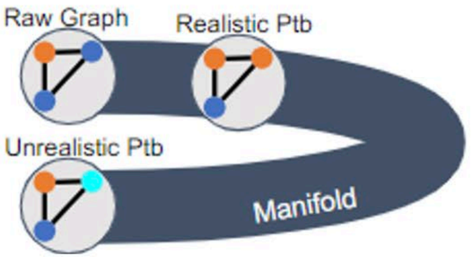
htong@illinois.edu, http://tonghanghang.org

NetSafe Overview: Safe Graph Neural Networks

- **Goal**: Build a Computational Foundation for End-to-end Safe GNNs.
- **Safety Notion**: **Performance Assurance against Hazards**
  - Hazards = external perturbation & dataset shifts
  - Requires the learning model (e.g., GNNs) to generalize well, **ideally with provable guarantee**, in the presence of unintended or unexpected behavior (largely remain same)
  - Three aspects: awareness, robustness, and confidence
    - Collectively identify, endure, and reduce hazards in a machine learning system
- **Research Tasks**: Safe Training (Task 1), Adaptation (Task 2), Testing (Task 3)
- **Evaluation Plan**: Benchmark evaluation, finance and power grid.

Research Tasks & Evaluation

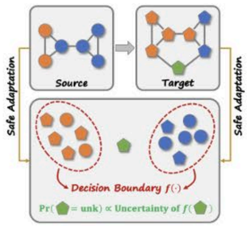
- T1.1: Learning Perturbation
- T1.2: Sensitivity-aware Training

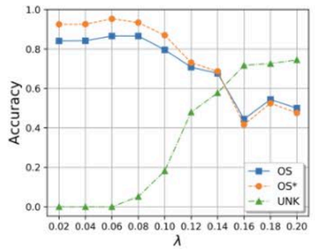


$$\min_{\theta}\left(\mathcal{L}(f_\theta) + \lambda \cdot \mathbb{E}_{G\sim\mathcal{P}}\left[\sup_{\Delta G \in g_\phi(G)} \frac{d_{\text{output}}(f_\theta(G), f_\theta(G+\Delta G))}{d_{\text{graph}}(G, G+\Delta G)}\right]\right)$$

Task 1: Safe Training

- T2.1: GNNs w/ Covariate Shift
- T2.2: GNNs w/ Label Shift



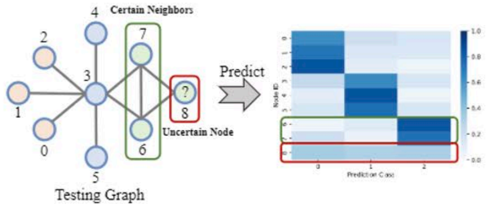


$$d_{\mathcal{G}}(G_s, G_t) = d^{\mathcal{C}}_{\text{GSD}}\left(G_s \otimes Y_s, (G_t \otimes Y_t)_{y\in\mathcal{C}_s}\right) - \rho \cdot d_{\text{GSD}}\left(G_s, G_{t|y\in\mathcal{C}_t\setminus\mathcal{C}_s}\right)$$

Task 2: Safe Adaptation

- T3.1: Closed-set Testing
- T3.2: Open-set Testing

$$\min_{\theta} \quad \mathcal{L}_{\text{test}} = \gamma\mathcal{L}_E + (1-\gamma)\mathcal{L}_D$$
$$s.t. \quad \mathcal{L}_E = \mathbb{E}_{v_i\in\mathcal{V}_{\text{test}}}[\texttt{Entropy}(f(\theta, G, v_i))],$$
$$\mathcal{L}_D = -\texttt{Entropy}(\mathbb{E}_{v_i\in\mathcal{V}_{\text{test}}}[f(\theta, G, v_i)])$$



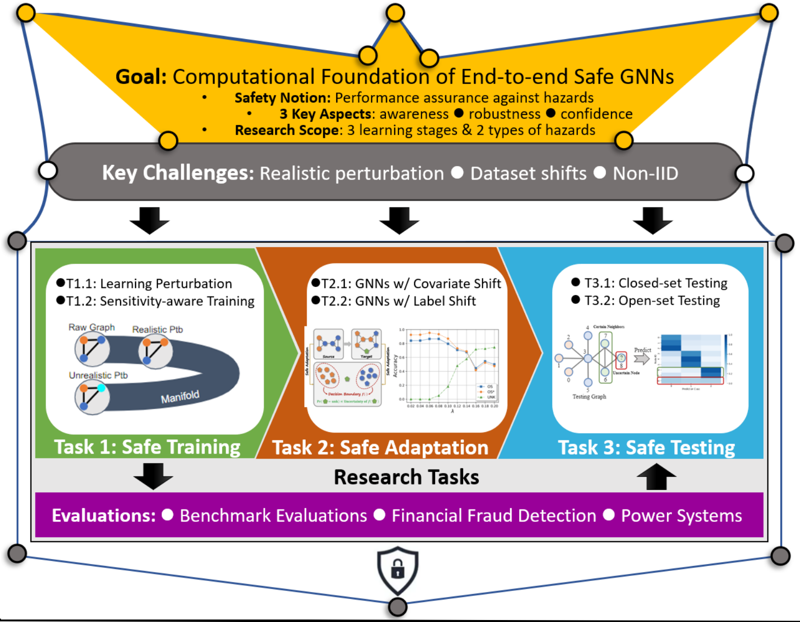
Task 3: Safe Testing

Research Tasks

Evaluations: • Benchmark Evaluations • Financial Fraud Detection • Power Systems

2416070, University of Illinois Urbana-Champaign, Hanghang Tong and Jingrui He

SLES: NetSafe: Towards a Computational Foundation of Safe Graph Neural Networks



Components:
- Safe GNNs Training
- Safe GNNs Adaptation
- Safe GNNs Testing

Rationale:
- Safety Notion: Performance Assurance against Hazards
- 3 Key Aspects: Awareness, Robustness, and Confidence
- Research Scope: 3 Learning Stages & 2 Types of Hazards

Safety Plan:
- Learning Realistic Perturbations for Safe Training
- Safe Adaptation against Covariant and Label Shift
- Closed-set and Open-set Safe Testing

Validation:
- Benchmark and Synthetic Datasets: Efficacy & Robustness
- Two Case Studies: Financial Fraud Detection & Early Event Detection in Power Systems

Intellectual Merit:
This project aims to build a computational foundation for end-to-end safe GNNs. It will establish new theoretical foundations in terms of the sensitivity, NP-hardness, confidence, and generalization error bound of safe GNNs. It will enable learning realistic perturbations and introduce new discrepancy and divergence measures for graphs, which will in turn lead to new algorithms for safe GNNs training, adaptation and testing with better efficacy and robustness.

Broader Impacts Plan:
• Benefit safety critical graph learning based applications, including fraud detection and power systems.
• Curriculum development, with supplemental material for the data mining textbook.
• Engaging minorities through mentoring programs, with an emphasis on bridging activities.
• Disseminating the data, code and manuscripts from this project.

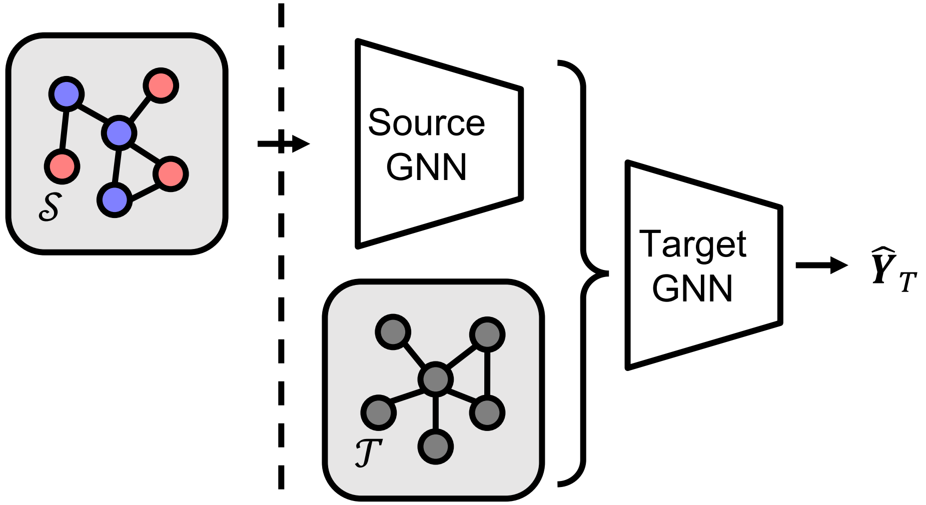
Prior Results:
- Adversarial Graph Training
- Open-set Domain Adaptation for IID Data
- Graph Anomaly and Event Detection

Expected Results:
- Thrust 1: Introducing a formal definition of realistic perturbations with quantified confidence, establishing a new sensitivity measure, and developing new algorithms for robust GNNs training.
- Thrust 2: Introducing a new graph discrepancy measure based on fused Gromov-Wasserstein distance, establishing the label-informed divergence measure for graphs, and unifying covariate shift and label shift for safe GNNs adaptation.
- Thrust 3: Designing an unsupervised objective for closed-set safe GNNs testing, and developing the open-set safe GNNs testing method without the access to the training graph.

Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation (ICLR 2025)

- **Distribution shifts in graphs**:
  - Attribute shift: Node feature distribution is different
    - E.g., LinkedIn and Instagram users have different profile
  - Structure shift: Node connectivity patterns vary
    - E.g., professional connections vs. family & friends

- **Graph test-time adaptation** (graph TTA):
  - Given: source GNN model, unlabeled target graph $\mathcal{T}$
  - Find: Target GNN model
  - Goal: Maximize the node classification accuracy

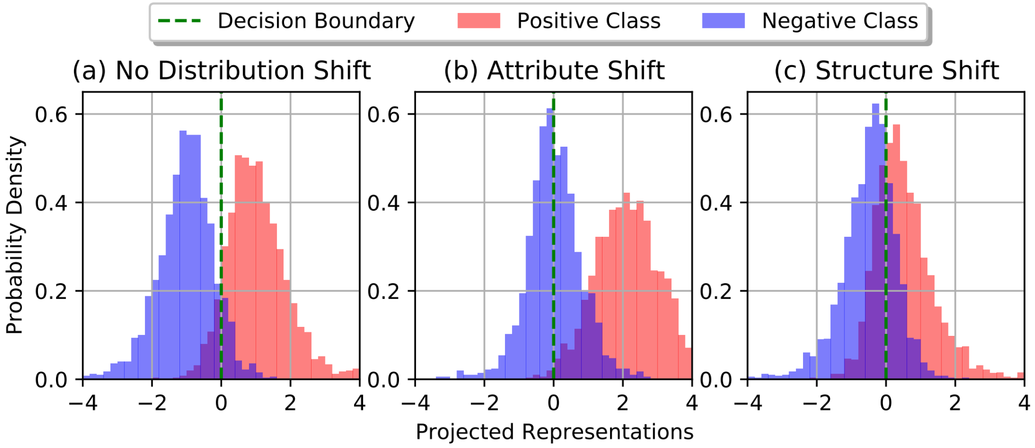


• Wenxuan Bao, Zhichen Zeng, Zhining Liu, Hanghang Tong, Jingrui He. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation. ICLR 2025.

Theoretical Findings

- **Challenge**: Most of the existing generic TTA algorithms, designed for other data (e.g., images), fail on graphs with structure shift.

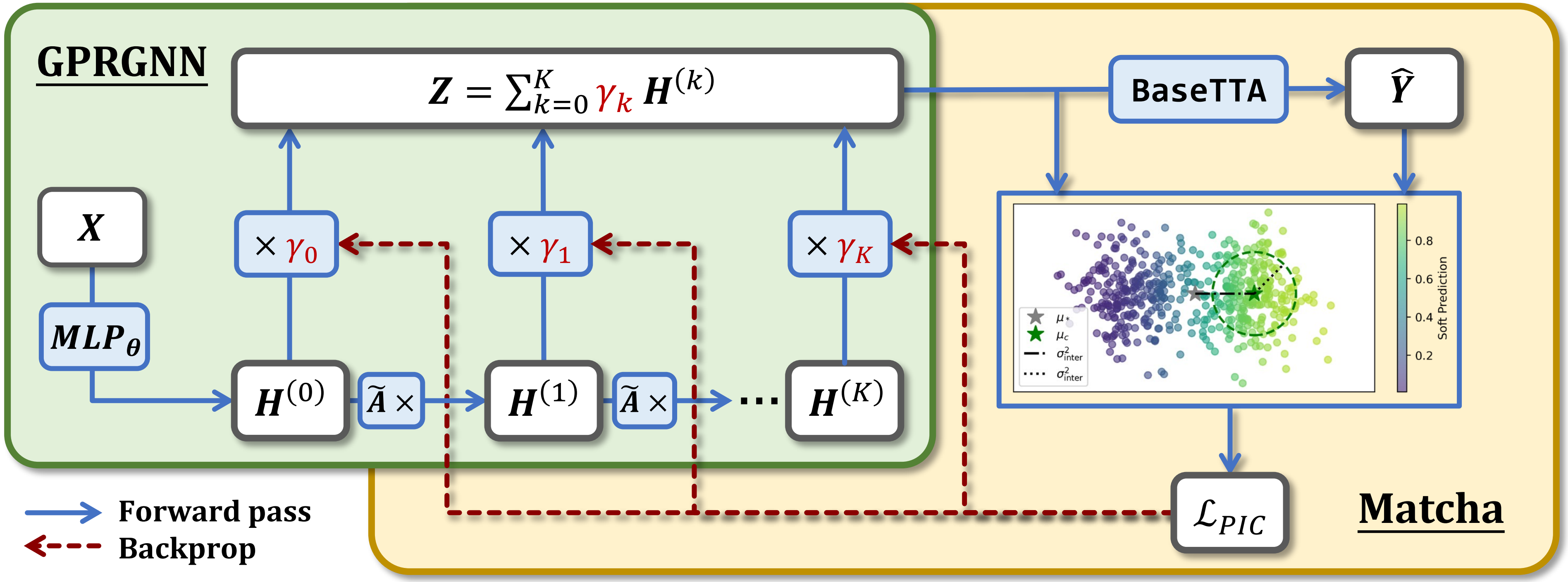
- **Our finding**: Attribute shifts and structure shifts have different impact patterns
  - Compared to attribute shifts (b), structure shifts (c) mix the distributions of node representations from different classes.
  - (See proofs in the paper)



Histograms of Projected Representations Under Different Distribution Shifts

• Wenxuan Bao, Zhichen Zeng, Zhining Liu, Hanghang Tong, Jingrui He. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation. ICLR 2025.

Matcha: Overview



- Matcha adapts the hop-aggregation parameters in GNNs (e.g., $\gamma_0, \cdots, \gamma_K$ for GPRGNN)

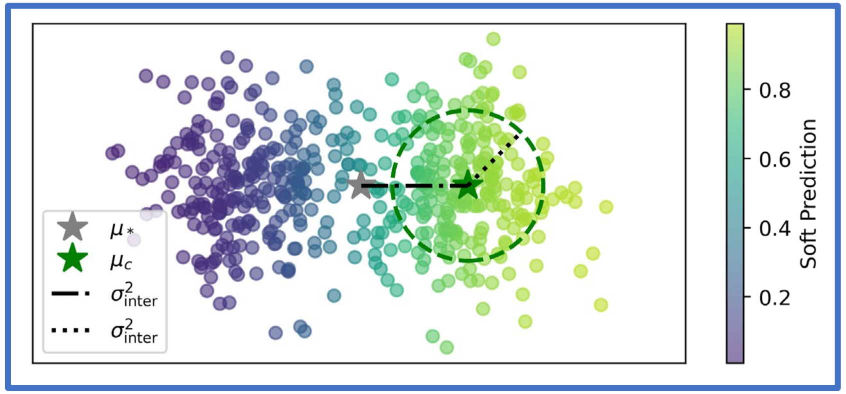
• Wenxuan Bao, Zhichen Zeng, Zhining Liu, Hanghang Tong, Jingrui He. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation. ICLR 2025.

Prediction-Informed Clustering Loss

- We proposed a new loss function:

$$\mathcal{L}_{\text{PIC}} = \frac{\sigma^2_{\text{intra}}}{\sigma^2_{\text{intra}} + \sigma^2_{\text{inter}}}, \text{ where}$$

  - Intra-class variance: $\sum_{i=1}^{M} \sum_{c=1}^{C} \hat{\boldsymbol{Y}}_{i,c} \|\boldsymbol{z}_i - \boldsymbol{\mu}_c\|_2^2$
  - Inter-class variance: $\sum_{c=1}^{C} \left(\sum_{i=1}^{M} \hat{\boldsymbol{Y}}_{i,c} \right) \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_*\|_2^2$
  - Centroid for class c: $\boldsymbol{\mu}_c = \frac{\sum_{i=1}^{M} \hat{\boldsymbol{Y}}_{i,c} \boldsymbol{z}_i}{\sum_{i=1}^{M} \hat{\boldsymbol{Y}}_{i,c}}$
  - Centroid for all nodes: $\boldsymbol{\mu}_* = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{z}_i$

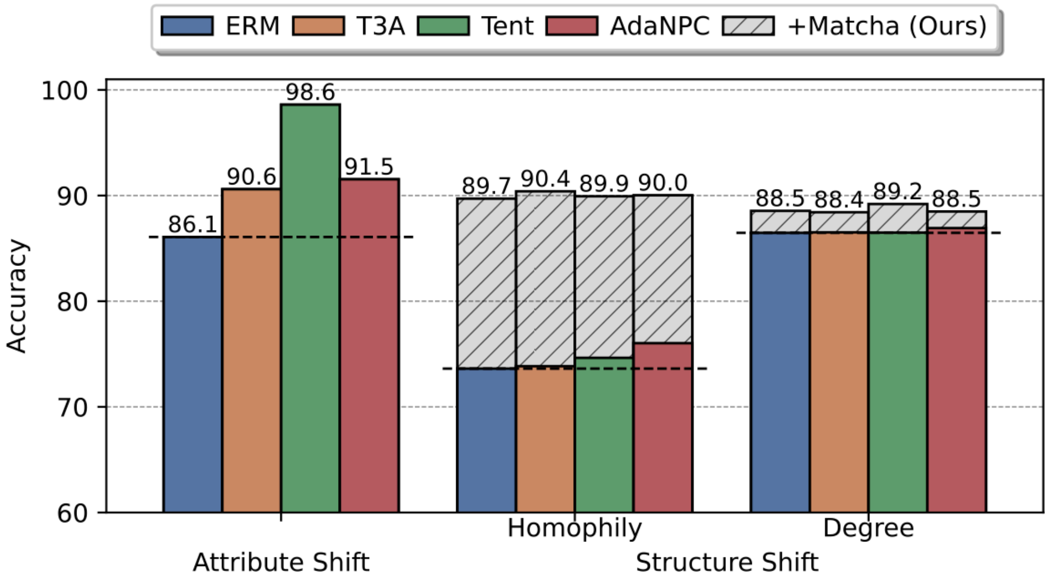


- Intuition
  - Small intra-class variance σ^2_{intra}, large inter-class variance σ^2_{inter}

• Wenxuan Bao, Zhichen Zeng, Zhining Liu, Hanghang Tong, Jingrui He. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation. ICLR 2025.

Experiments: Matcha Enhances the Performance of Existing TTA Methods

- Synthetic CSBM dataset with different types of structure shifts



- Real-world datasets

Method	Syn-Cora	Syn-Products	Twitch-E	OGB-Arxiv
ERM	65.67 ± 0.35	37.80 ± 2.61	56.20 ± 0.63	41.06 ± 0.33
+ Matcha	78.96 ± 1.08	69.75 ± 0.93	56.76 ± 0.22	41.74 ± 0.34
T3A	68.25 ± 1.10	47.59 ± 1.46	56.83 ± 0.22	38.17 ± 0.31
+ Matcha	78.40 ± 1.04	69.81 ± 0.36	56.97 ± 0.28	38.56 ± 0.27
Tent	66.26 ± 0.38	29.14 ± 4.50	58.46 ± 0.37	34.48 ± 0.28
+ Matcha	78.87 ± 1.07	68.45 ± 1.04	$\mathbf{58.57 \pm 0.42}$	35.20 ± 0.27
AdaNPC	67.34 ± 0.76	44.67 ± 1.53	55.43 ± 0.50	40.20 ± 0.35
+ Matcha	77.45 ± 0.62	71.66 ± 0.81	56.35 ± 0.27	40.58 ± 0.35
GTrans	68.60 ± 0.32	43.89 ± 1.75	56.24 ± 0.41	41.28 ± 0.31
+ Matcha	$\mathbf{83.49 \pm 0.78}$	$\mathbf{71.75 \pm 0.65}$	56.75 ± 0.40	41.81 ± 0.31
SOGA	67.16 ± 0.72	40.96 ± 2.87	56.12 ± 0.30	41.23 ± 0.34
+ Matcha	79.03 ± 1.10	70.13 ± 0.86	56.62 ± 0.17	41.78 ± 0.34
GraphPatcher	63.01 ± 2.29	36.94 ± 1.50	57.05 ± 0.59	41.27 ± 0.87
+ Matcha	80.99 ± 0.50	69.39 ± 1.29	57.41 ± 0.53	$\mathbf{41.83 \pm 0.90}$

- Wenxuan Bao, Zhichen Zeng, Zhining Liu, Hanghang Tong, Jingrui He. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation. ICLR 2025.

Experiments: Matcha Restores the Representation Quality

- While structure shifts blur the boundary of node classes (b), Matcha can restore the representation quality (f), better than other loss functions.

Loss	Homophily shift		Degree shift	
	homo $\rightarrow$ hetero	hetero $\rightarrow$ homo	high $\rightarrow$ low	low $\rightarrow$ high
(None)	73.62 $\pm$ 0.44	76.72 $\pm$ 0.89	86.47 $\pm$ 0.38	92.92 $\pm$ 0.43
Entropy	75.89 $\pm$ 0.68	89.98 $\pm$ 0.23	86.81 $\pm$ 0.34	93.75 $\pm$ 0.72
PseudoLabel	77.29 $\pm$ 3.04	89.44 $\pm$ 0.22	86.72 $\pm$ 0.31	93.68 $\pm$ 0.69
$\sigma^2_{\text{intra}} - \sigma^2_{\text{inter}}$	76.10 $\pm$ 0.43	72.43 $\pm$ 0.65	82.56 $\pm$ 0.99	92.92 $\pm$ 0.44
PIC (Ours)	**89.71 $\pm$ 0.27**	**90.68 $\pm$ 0.26**	**88.55 $\pm$ 0.44**	**93.78 $\pm$ 0.74**