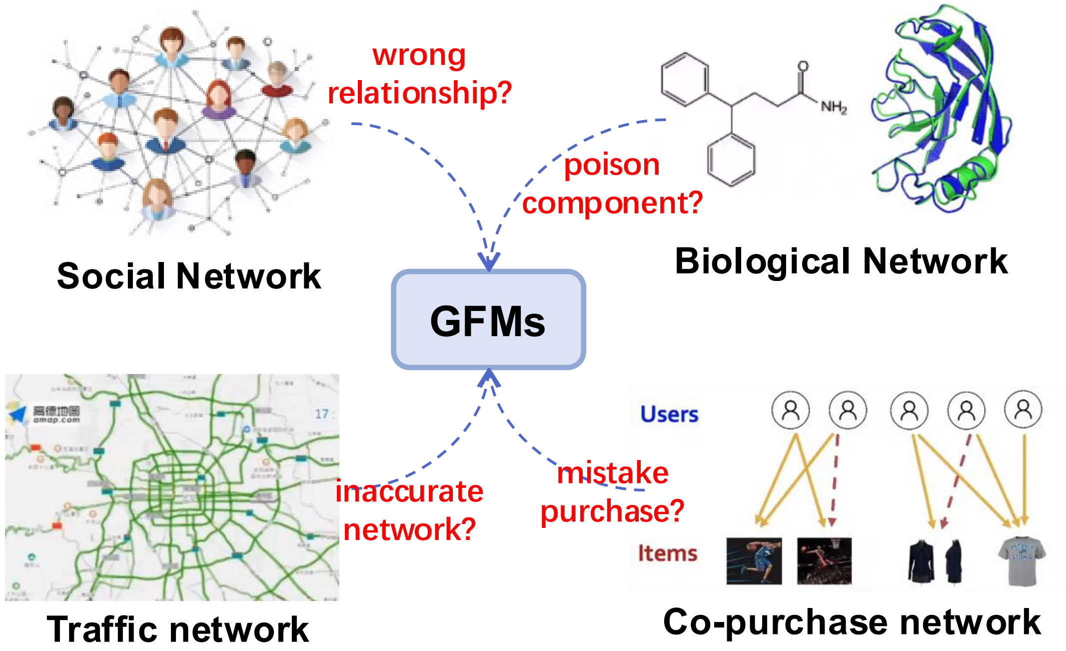
GFM Security

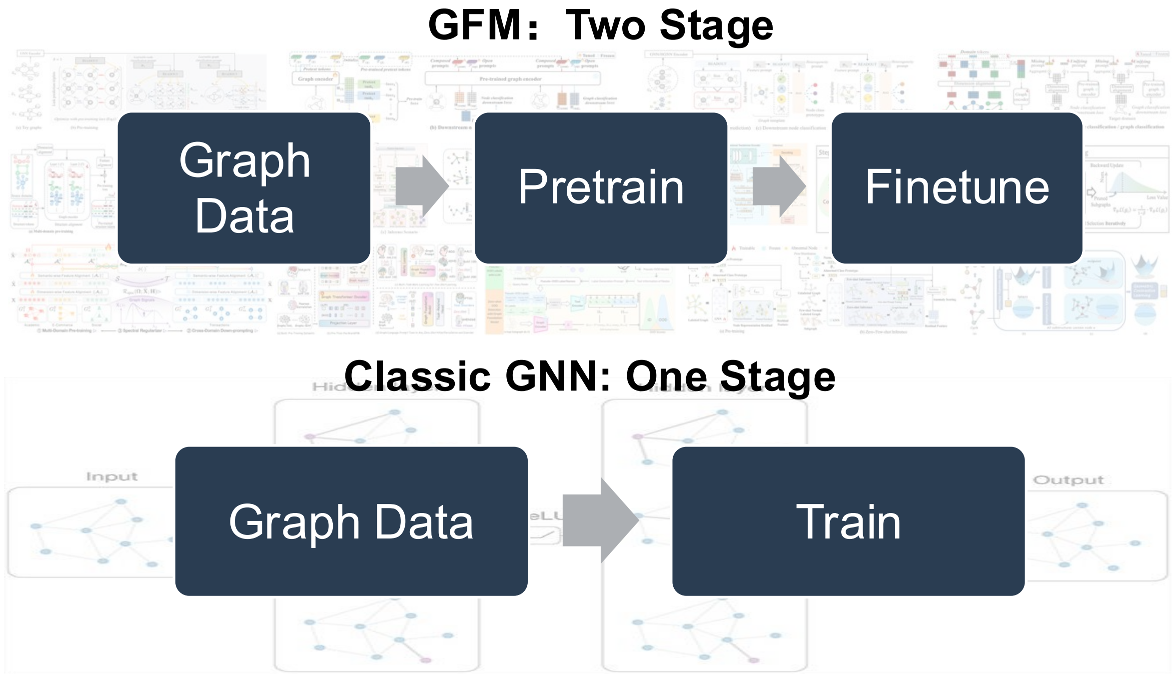
The Security problem in GFMs

■ Defense in GFMs



Data reliability in various open-domain environments cannot be reliably ensured.

■ Attack in GFMs



New training paradigm poses new challenges to adversaries to rethink **how attacks should be formulated**.

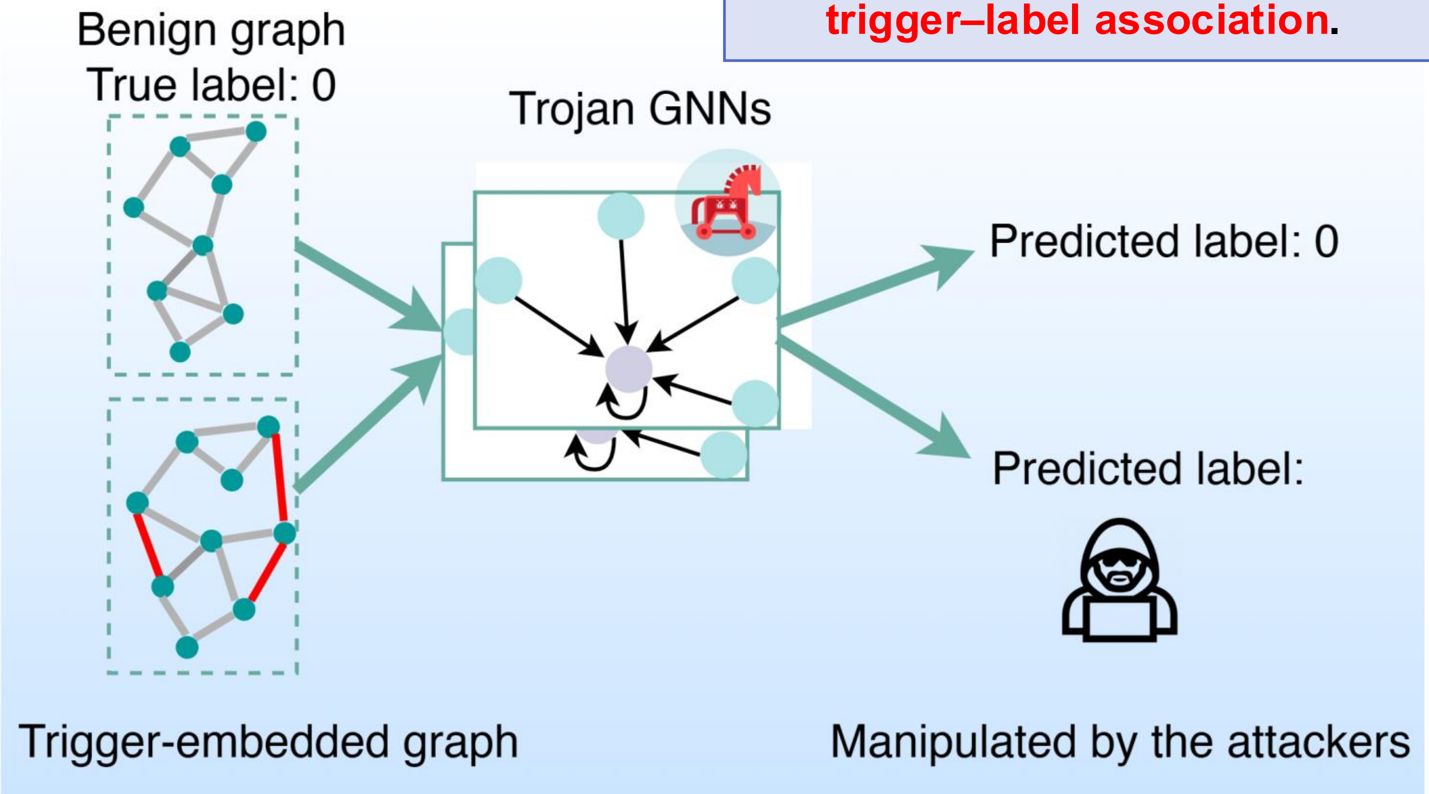
Towards Effective, Stealthy, and Persistent Backdoor Attacks Targeting Graph Foundation Models.
Privacy Auditing of Multi-domain Graph Pre-trained Model under Membership Inference Attacks.

GFM Backdoor

Graph Backdoor

①Poisoning graph training data **(Needs to know the downstream label)**

② Training GNNs on the poisoned graph to inject the **trigger–label association**.



③ Manipulate the backdoored label via trigger

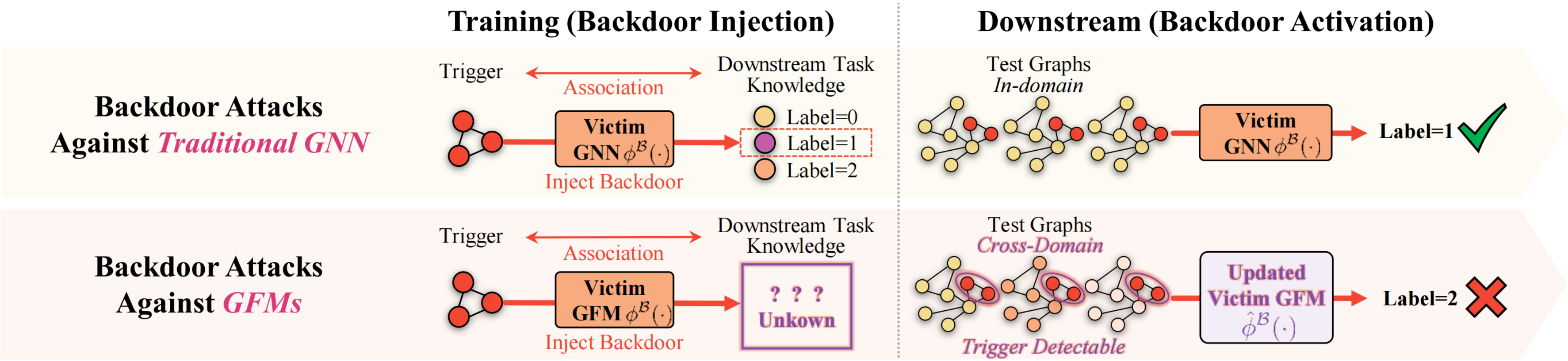
[Picture Credit] Transferable Graph Backdoor Attack. *Shuiqiao Yang*, et al.

By poisoning the training data, an attacker can manipulate the labels predicted by the backdoored GNN.

Core principle: Enforcing the [trigger]-[target label] association.

GFM Backdoor

Traditional Graph Backdoor *vs.* GFM Backdoor



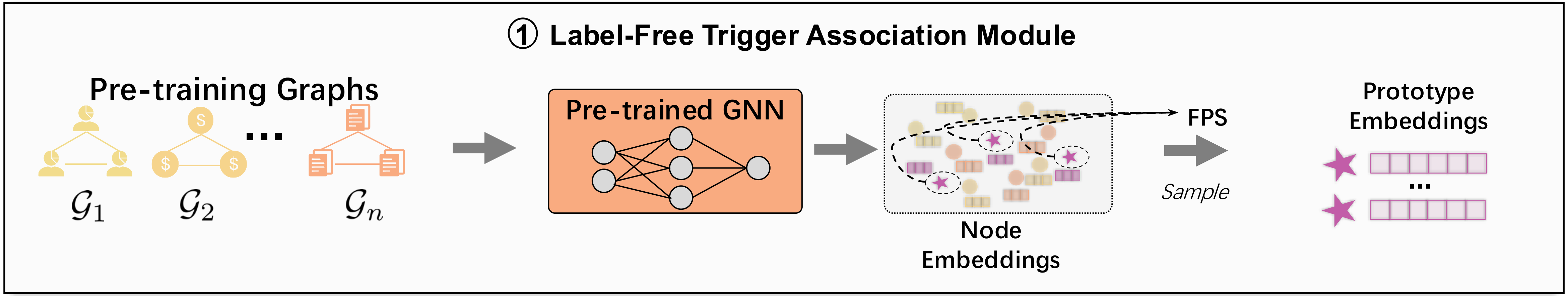
- Three key Challenge in backdooring GFM:
  - (1) ***Effectiveness challenge***: Downstream labels and tasks are unknown during GFM training, making it difficult to inject a backdoor that enforces a trigger–label correlation.
  - (2) ***Stealthiness challenge***: Strong cross-domain distribution shift, so a fixed, universal trigger is more likely to be anomalous and get detected.
  - (3) ***Persistence challenge***: Downstream fine-tuning can overwrite the backdoor, causing the attack to fade

The GFM setting breaks three core assumptions of traditional Graph backdoors:
(1) unknown labels, (2) cross-domain data, and (3) unfixed parameters.

GFM Backdoor

Proposed GFM-BA: (1) Label-free Trigger Association Module

- **Target: Improve Effectiveness**:
  - Sample k prototype embeddings using Farthest Point Sampling while associate trigger to these prototype embeddings. $\mathcal{L}_{\text{eff}} = -\mathbb{E}_{u\sim[n],j\sim[k],i\sim[|\mathcal{G}_u|]}\text{sim}\left(\phi(\text{In}(\mathcal{G}_u,\tilde{\mathcal{G}}_{ij},i)),e_j\right),$
  - At downstream activation time, the attacker uses a small number of trial queries to identify which prototype aligns with the target label, and then uses the corresponding trigger.



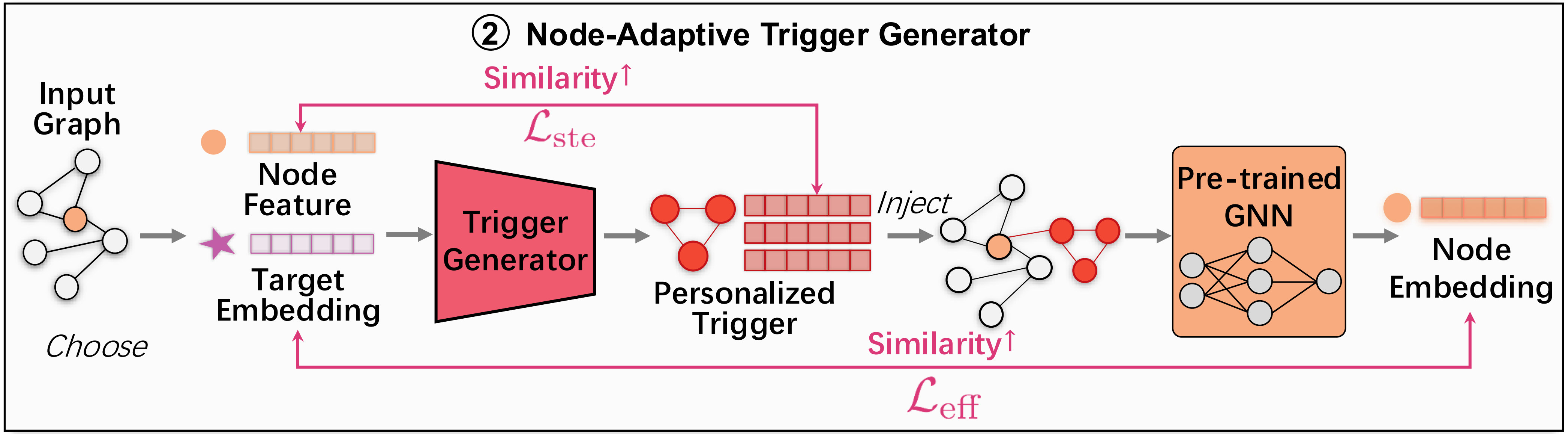
Core Idea: Binding the trigger to a set of prototype embeddings instead of the label (because the label is unknown during pre-training).

GFM Backdoor

Proposed GFM-BA: (2) Node-Adaptive Trigger Generator

- **Target: Improve Stealthiness**:
  - Introduce a node-adaptive trigger generator with the trigger is a small 3-node fully connected subgraph
  - Trigger node features are generated by an MLP for each target node

$$\mathcal{L}_{\text{ste}} = -\mathbb{E}_{u\sim[n],j\sim[k],i\sim[|\mathcal{G}_u|]}\text{sim}\left(x_i, x_{ij}^{\text{tri}}\right).$$



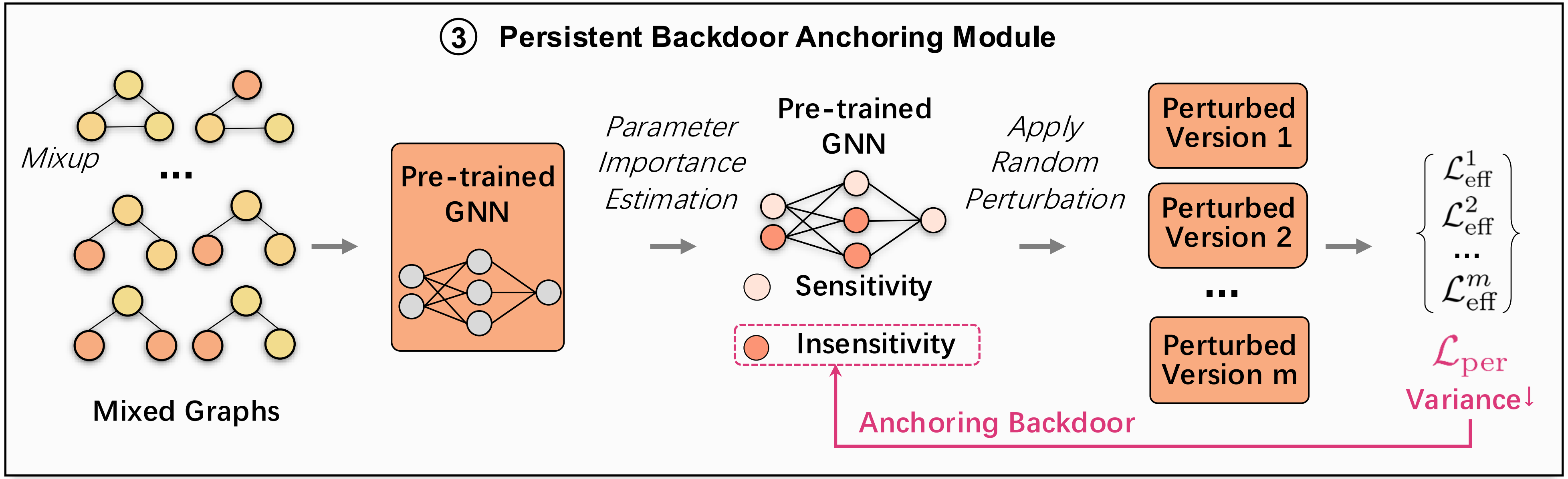
Core Idea: Generating **customized trigger for each target node**, making it harder to detect.

GFM Backdoor

Proposed GFM-BA: (3) Persistent Backdoor Anchoring Module

- **Improve Persistence:**
  - Use graph mixup to generate mixed graphs and approximate potential downstream distributions.
  - Estimate parameter sensitivity and select the most sensitive top s% parameters.
  - Using adversarial perturbation to anchor the backdoor to the insensitive parameters.

$$\mathcal{L}_{\text{per}} = \text{Var}(\{\mathcal{L}_{\text{eff}}^{j}\}_{j=1}^{m}) + \text{Mean}(\{\mathcal{L}_{\text{eff}}^{j}\}_{j=1}^{m}).$$



Core Idea: "Anchor" the backdoor to parameter regions that are less likely to be altered by fine-tuning.

GFM Backdoor

- **Experiments:** Our method’s backdoor activation achieves a near 100% success rate across five datasets .

Victim	Threaten	Cora ASR (*Scen.1*)	Cora ASR (*Scen.2*)	CiteSeer ASR (*Scen.1*)	CiteSeer ASR (*Scen.2*)	PubMed ASR (*Scen.1*)	PubMed ASR (*Scen.2*)	Photo ASR (*Scen.1*)	Photo ASR (*Scen.2*)	Computers ASR (*Scen.1*)	Computers ASR (*Scen.2*)
GCOPE	GCBA_R	26.78±8.89	3.83±1.27	29.64±8.67	4.94±1.45	62.10±12.81	20.70±4.27	26.80±8.79	3.35±1.10	35.40±19.13	3.54±1.91
	GCBA_M	33.42±1.89	4.77±0.27	35.87±6.50	5.98±1.08	64.94±15.18	21.65±5.06	27.80±9.26	3.48±1.16	46.20±12.34	4.62±1.23
	CrossBA	**100.00±0.00**	14.29±0.00	**100.00±0.00**	16.67±0.00	**100.00±0.00**	33.33±0.00	74.00±13.44	9.25±1.68	79.80±7.26	7.98±0.73
	GFM-BA	**100.00±0.00**	**90.40±12.10**	**100.00±0.00**	**89.06±8.02**	**100.00±0.00**	**100.00±0.00**	**93.20±7.05**	**84.53±1.79**	**94.40±7.70**	**78.54±7.28**
SAMGPT	GCBA_R	61.66±8.82	8.81±1.26	51.60±13.57	8.60±2.26	89.02±18.39	29.67±6.13	52.00±9.95	6.50±1.24	66.60±6.66	6.66±0.67
	GCBA_M	61.44±13.41	8.78±1.92	64.42±15.49	10.74±2.58	94.82±4.64	31.61±1.55	57.40±9.58	7.18±1.20	68.00±12.83	6.80±1.28
	CrossBA	95.24±10.64	13.61±1.52	**100.00±0.00**	16.67±0.00	**100.00±0.00**	33.33±0.00	96.80±4.55	12.10±0.57	92.00±11.75	9.20±1.17
	GFM-BA	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**99.80±0.23**	**100.00±0.00**	**100.00±0.00**
MDGPT	GCBA_R	-	-	-	-	-	-	-	-	-	-
	GCBA_M	-	-	-	-	-	-	-	-	-	-
	CrossBA	95.14±5.26	13.59±0.75	92.36±15.87	15.39±2.65	**100.00±0.00**	33.33±0.00	82.60±19.86	10.32±2.48	82.00±19.21	8.20±1.92
	GFM-BA	**100.00±0.00**	**96.61±6.17**	**100.00±0.00**	**99.43±0.95**	**100.00±0.00**	**100.00±0.00**	**98.40±2.30**	**97.68±2.57**	**99.20±1.30**	**93.19±3.66**

Table 1: The results of attack effectiveness. *Scen.1* refers to *target-uncontrolled attack*, and *Scen.2* refers to *target-controlled attack*. “-” indicates that the method is not applicable to perform the backdoor attack (GCBA (Zhang et al. 2023) is tailored for the graph contrastive learning method). The best results are shown in **bold** and the runner-ups are underlined.

Victim	Threaten	Cora ACC (*Clean*)	Cora ASR (*Purified*)	CiteSeer ACC (*Clean*)	CiteSeer ASR (*Purified*)	PubMed ACC (*Clean*)	PubMed ASR (*Purified*)	Photo ACC (*Clean*)	Photo ASR (*Purified*)	Computers ACC (*Clean*)	Computers ASR (*Purified*)
GCOPE	GCBA_R	59.18±3.93	20.87±9.57	56.30±2.86	27.28±12.83	55.00±4.39	42.18±13.59	57.00±3.24	26.60±9.56	44.60±3.71	35.00±20.04
	GCBA_M	59.78±7.20	23.27±8.89	56.12±2.13	22.62±6.51	51.10±5.79	48.71±22.34	58.80±7.33	26.80±8.90	47.20±7.26	43.20±13.55
	CrossBA	60.52±1.68	52.04±2.42	59.06±4.45	57.58±2.80	51.36±5.37	90.65±2.52	65.60±3.51	67.20±9.73	50.40±4.83	48.80±15.85
	GFM-BA	**61.46±3.25**	**100.00±0.00**	**60.10±5.72**	**100.00±0.00**	**54.44±4.43**	**100.00±0.00**	**65.80±4.97**	**100.00±0.00**	**54.80±7.56**	**100.00±0.00**
SAMGPT	GCBA_R	60.36±1.89	38.24±6.24	44.10±4.47	43.92±10.22	58.92±4.35	77.80±17.21	79.20±4.32	17.80±3.11	69.00±6.00	26.20±4.15
	GCBA_M	61.94±3.56	29.32±5.45	48.76±4.09	48.38±12.30	63.02±2.60	86.26±9.05	76.20±5.67	18.20±3.11	**71.20±2.59**	23.20±4.82
	CrossBA	62.82±2.85	70.28±11.22	59.04±2.86	75.90±1.95	65.64±4.40	74.08±10.75	80.40±4.34	67.80±2.39	67.20±7.12	50.60±6.95
	GFM-BA	**63.54±3.73**	**88.28±3.08**	**61.72±3.61**	**86.04±1.56**	**65.92±5.65**	**93.62±2.20**	**80.60±3.58**	**86.00±2.12**	69.20±4.09	**84.60±2.19**
MDGPT	GCBA_R	-	-	-	-	-	-	-	-	-	-
	GCBA_M	-	-	-	-	-	-	-	-	-	-
	CrossBA	42.36±7.08	47.76±13.84	37.82±5.00	61.12±20.64	50.20±6.87	69.10±20.81	68.20±9.09	35.40±5.73	50.20±6.10	37.00±8.69
	GFM-BA	**60.88±4.83**	**81.30±3.60**	**60.58±2.82**	**85.78±2.96**	**62.48±4.71**	**92.36±1.97**	**79.20±6.14**	**89.60±5.68**	**71.80±3.49**	**85.20±4.82**

Table 2: Results of attack stealthiness. ACC reports the accuracy of the backdoored model on clean, non-triggered input graphs. *Purified* refers to applying edge-based purification to the triggered graphs under Scenario 1.

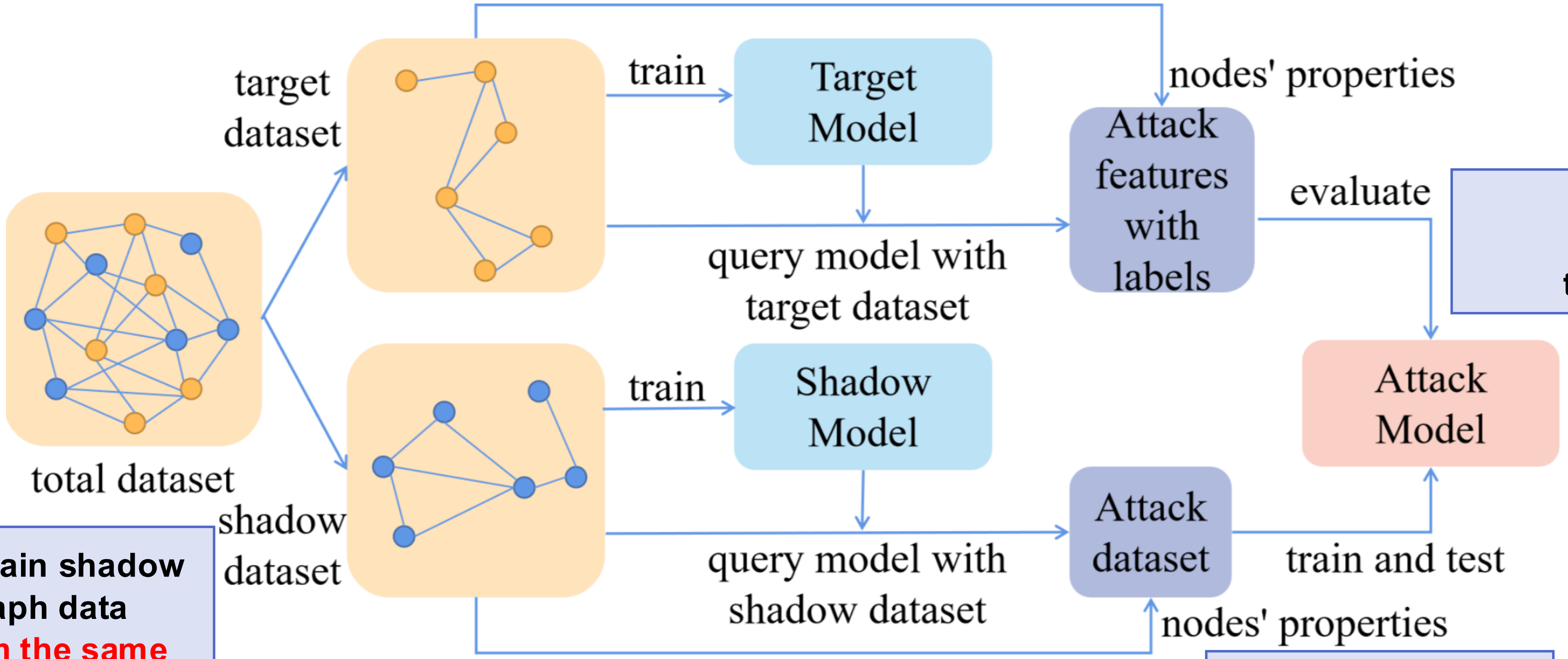
Achieve an almost 100% backdoor attack success rate

Maintain a high attack success rate even under backdoor defenses.

GFM MIA

Graph Membership Inference Attack (MIA)

- **MIA: Tries to determine whether a specific data sample was included in a model's training set.**



③ Infer members according to the overfitting signals

① Obtain shadow graph data
(From the same domain as real training data)

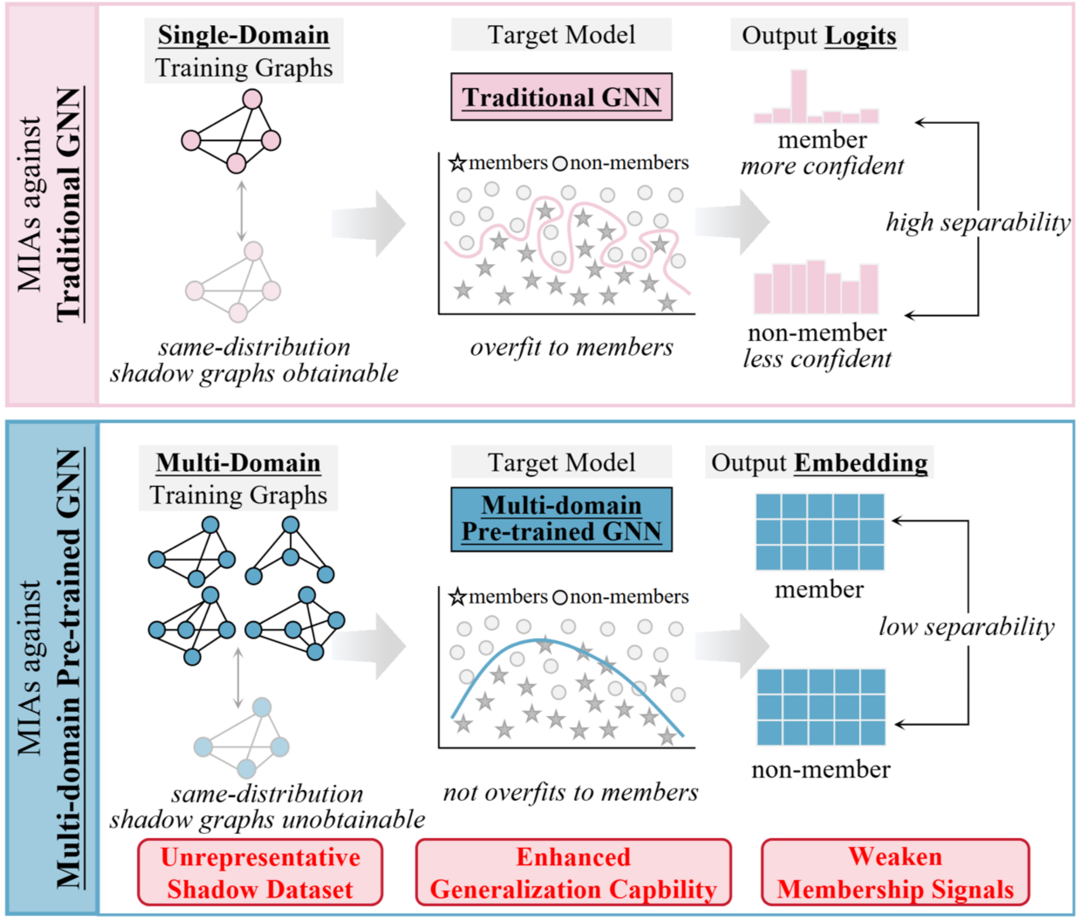
② Training Shadow GNN

[Picture Credit] Label-Only Membership Inference Attack against Node-Level Graph Neural Networks. *Mauro Conti*, et al.

Core principle: Members exhibit stronger overfitting signals than non-members.

GFM MIA

Traditional Graph MIA *vs.* GFM MIA



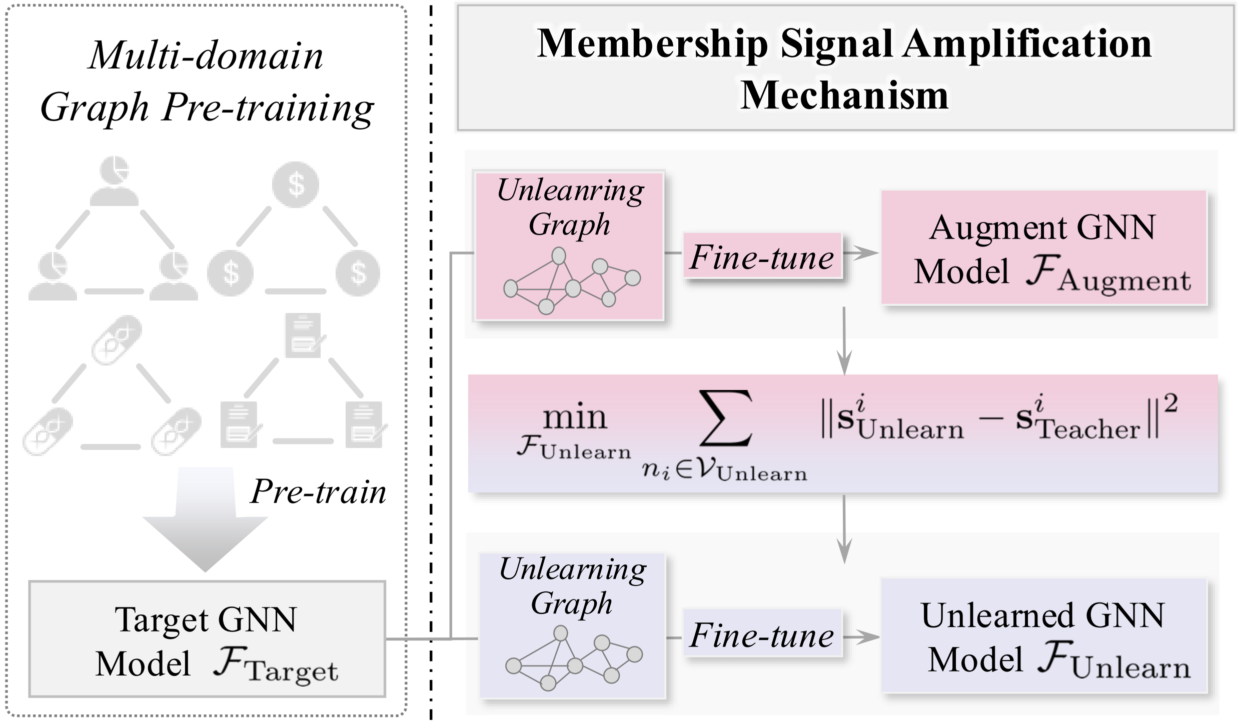
- Three key Challenge in performing MIA against GFM:
  - (1) ***Enhanced Generalization***: reduces overfitting signals of members used by MIAs.
  - (2) ***Unreliable Shadow Graphs***: Attackers rarely obtain shadow data that matches the entire domain; they typically only obtain small graph data from a single domain.
  - (3) ***Weaken Membership Signals***: The model output is an embedding rather than logits/confidence scores, making classic confidence-based MIA more difficult.

The GFM setting breaks three core assumptions of traditional Graph MIA:
(1) Lower member/non-member distinguishability, (2) Unreliable Shadow Data (3) Weaken Overfitting Signal.

GFM Backdoor

Proposed GFM-MIA: (1) Membership Signal Amplification

Enhance Overfitting → *Building Shadow Model* → *Training Attack Model*



- **Enhance Member/non-member distinguishability:**
  - We construct a small unlearning set.
  - Apply unlearning to the target model to free up memorization capacity, thereby strengthening memorization of the remaining members and amplifying membership signals.

$$\mathbf{s}^i_{\text{Teacher}} = \mathbf{s}^i_{\text{Target}} - \lambda \cdot \left(\mathbf{s}^i_{\text{Target}} - \mathbf{s}^i_{\text{Augment}}\right),$$

$$\min_{\mathcal{F}_{\text{Unlearn}}} \sum_{n_i \in \mathcal{V}_{\text{Unlearn}}} \left\| \mathbf{s}^i_{\text{Unlearn}} - \mathbf{s}^i_{\text{Teacher}} \right\|^2,$$

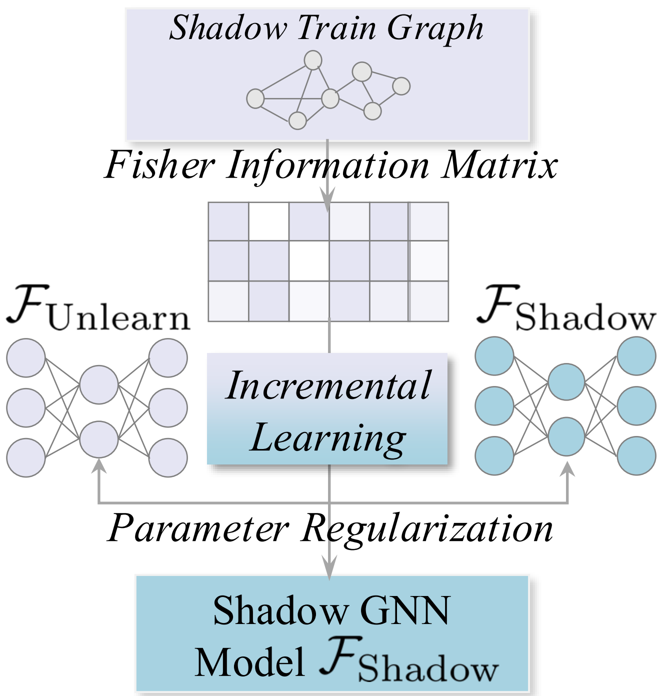
Core Idea: Free memorization capacity to amplify member overfitting signals.

GFM Backdoor

Proposed GFM-MIA: (2) Incremental Shadown Model Construction

Enhance Overfitting → *Building Shadow Model* → *Training Attack Model*

Incremental Shadow Model Construction Mechanism



- **Build Reliable Shadow Model:**
  - Incrementally fine-tune the target model with a Fisher-based regularizer to build the shadow model.
  - This keeps key parameters close to the target model, making the shadow model more faithful and improving transferability for membership inference.

$$\mathbf{I}_{\text{Unlearn}}(\theta) = \mathbb{E}_{v\sim\mathcal{G}^{\text{Train}}_{\text{Shadow}}}\left[\frac{\partial^2\mathcal{L}_{\text{task}}(\mathcal{F}_{\text{Unlearn}};v)}{\partial\theta^2}\Big|\theta\right],$$

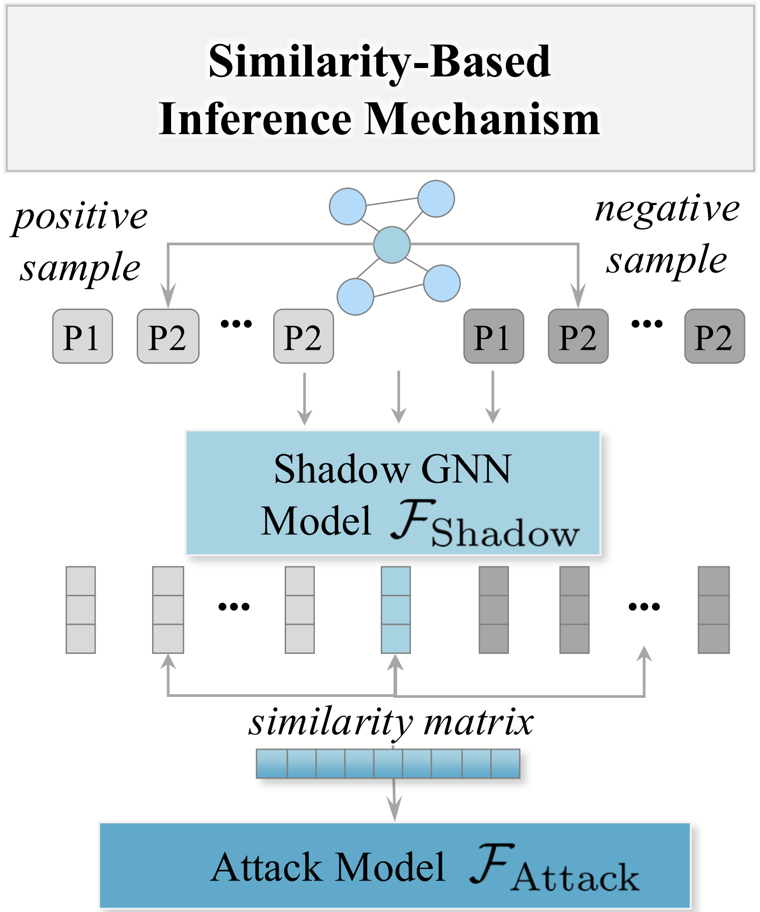
$$\min_{\Theta_{\text{Shadow}}} \sum_{v\in\mathcal{G}^{\text{Train}}_{\text{Shadow}}} \mathcal{L}_{\text{task}}(\mathcal{F}_{\text{Shadow}};v) + \alpha\sum_{i}\mathbf{I}^{(i)}_{\text{Unlearn}}\left(\Theta^{(i)}_{\text{Shadow}}-\Theta^{(i)}_{\text{Unlearn}}\right)^2,$$

Core Idea: Incrementally construct a reliable shadow model from the target model with limited shadow data.

GFM Backdoor

Proposed GFM-MIA: (3) Similarity-Based Inference Mechanism

Enhance Overfitting → *Building Shadow Model* → *Training Attack Model*



- **Membership Inference without Confidence Score**:
  - We discover that members have higher similarity to their positive samples than non-members.
  - Select shadow models's outputs for positives and negatives samples to construct an attack training set.
  - We train a two-layer MLP to predict membership from these positive/negative similarity features, without relying on confidence scores.

Core Idea: Members have higher similarity to their positive samples than non-members.