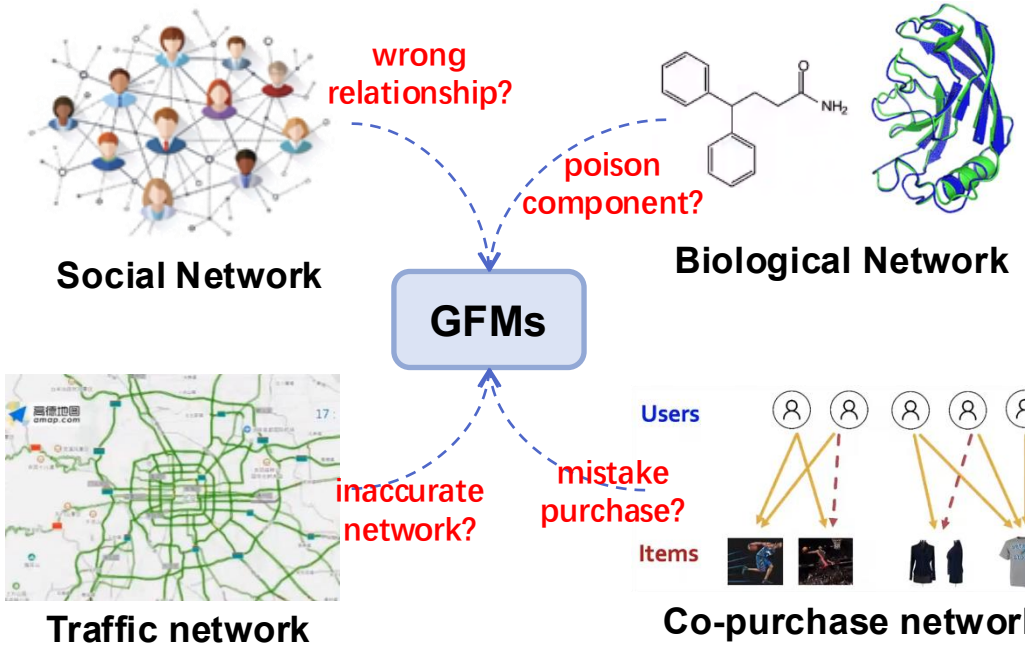
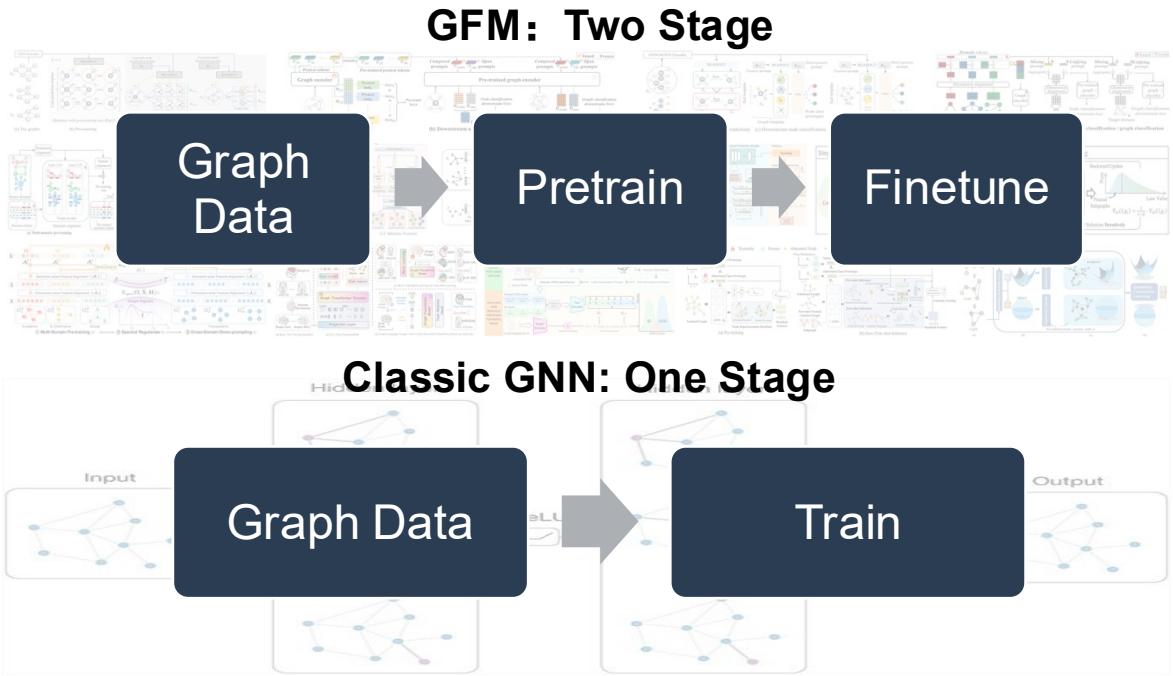
GFM Security

The Security problem in GFMs

Defense in GFMs



Attack in GFMs



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

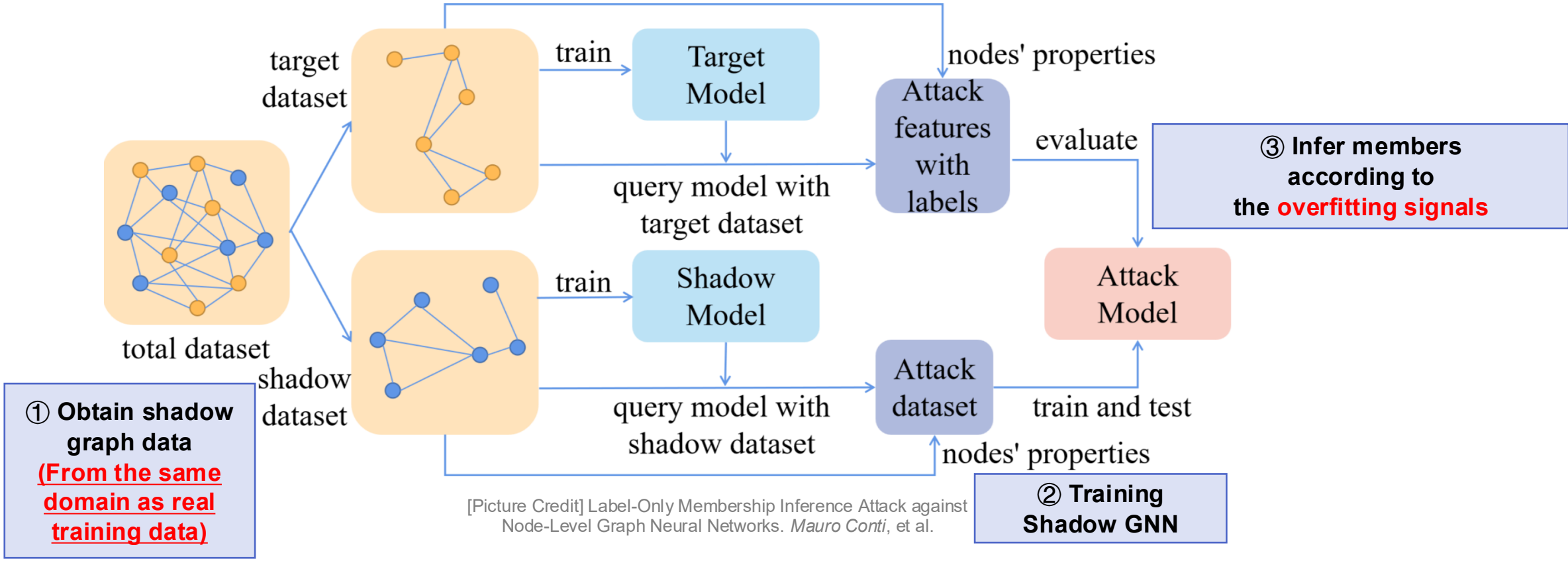
New training paradigm poses new challenges to adversaries to rethink **how attacks should be formulated**.

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 MIA

Graph Membership Inference Attack (MIA)

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

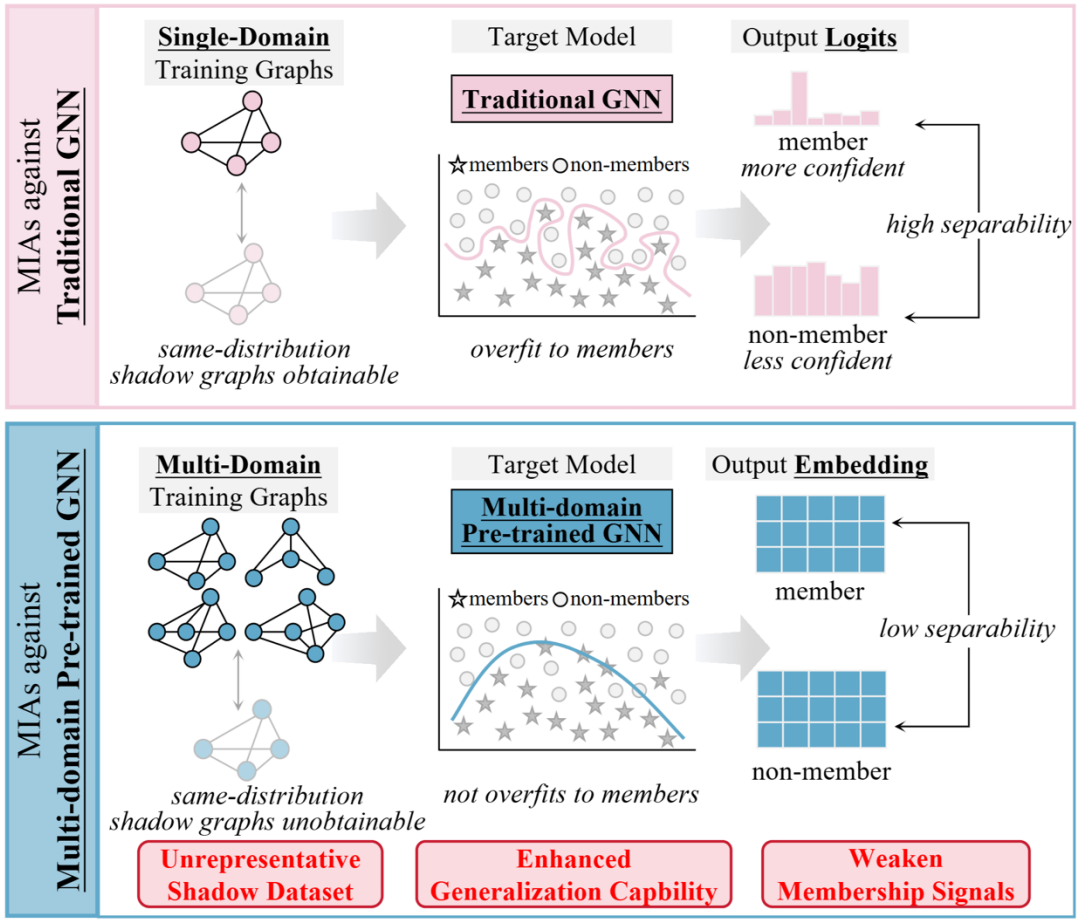


[Picture Credit] Label-Only Membership Inference Attack against Node-Level Graph Neural Networks. *Mauro Conti*, et al.

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

GFM MIA

Traditional Graph MIA *vs.* GFM MIA



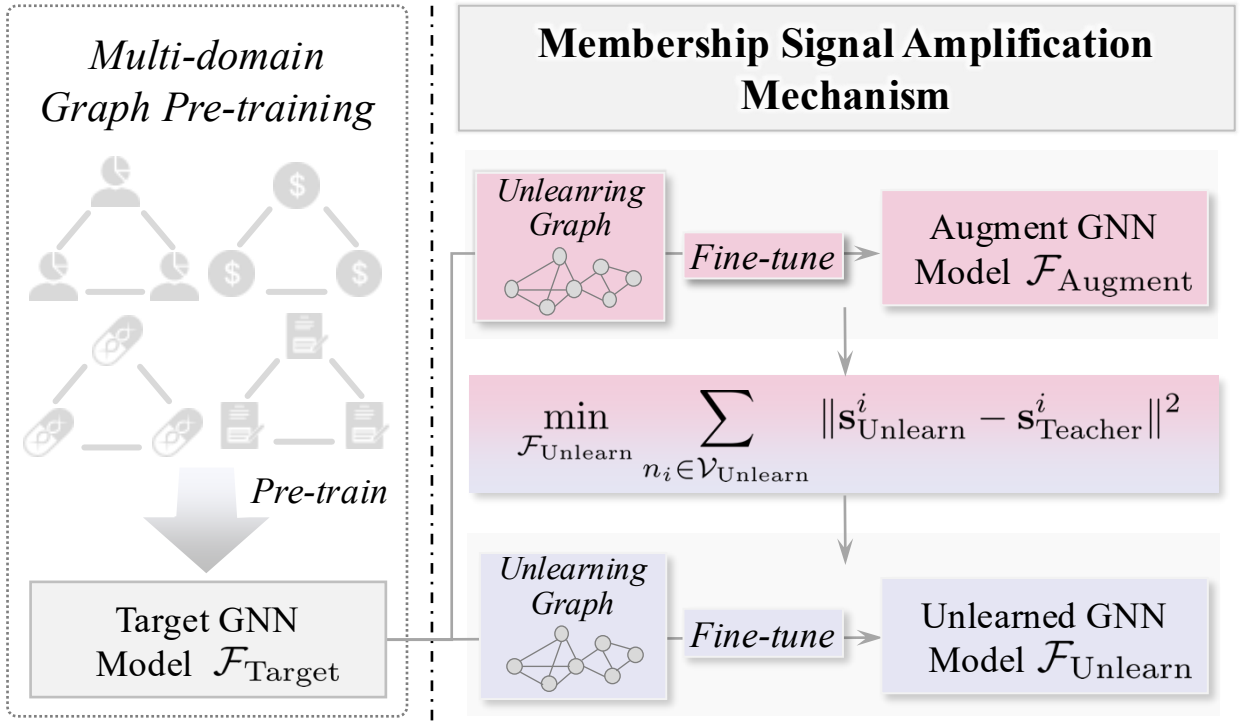
- Three key Challenge in performing MIA against GFM:
  - (1) ***Enhanced Generalization***: reduces overfitting signals of members used by MIAs.
  - (2) ***Unreliable Shadow Graphs***: Attackers rarely obtain shadow data that matches the entire domain; they typically only obtain small graph data from 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.

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

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

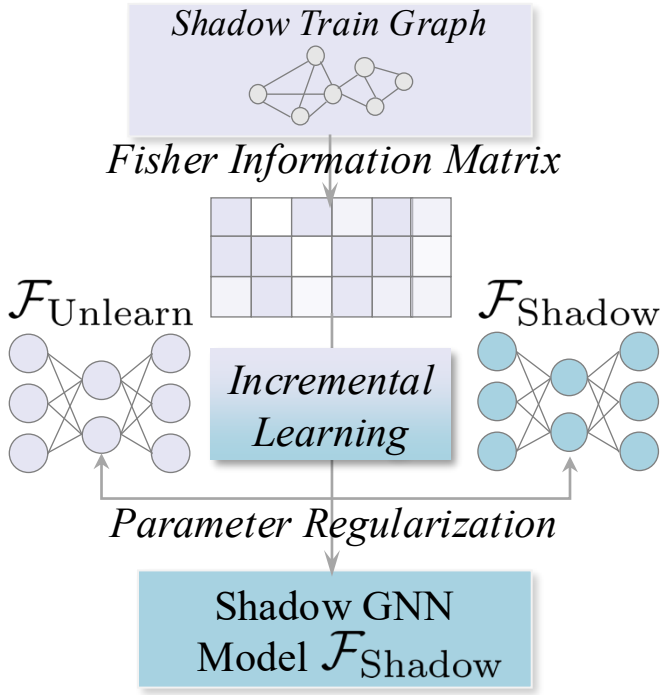
Core Idea: Free memorization capacity to amplify member overfitting signals.

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.

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

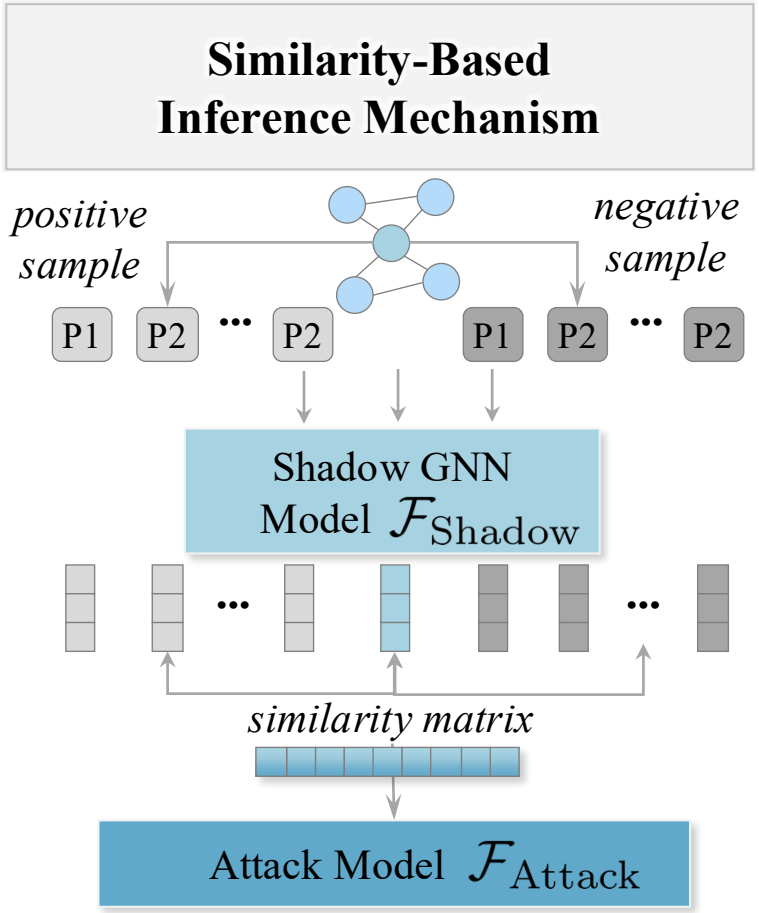
$$\min_{\Theta_{\text{Shadow}}} \sum_{v \in \mathcal{G}^{\text{Train}}_{\text{Shadow}}} \mathcal{L}_{\text{task}}(\mathcal{F}_{\text{Shadow}}; v) + \alpha \sum \mathbf{I}^{(i)}_{\text{Unlearn}} \left(\Theta^{(i)}_{\text{Shadow}} - \Theta^{(i)}_{\text{Unlearn}} \right)^2,$$

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.

GFM MIA

Experiments

- On average, our method achieves a 15% increase in MIA accuracy across five datasets over six state-of-the-art baselines.

(1) link-prediction-based GFM

Dataset		Cora		CiteSeer		PubMed		Photo		Computers	
Victims	Method	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MDGPT	Embed-MIA	68.89±1.83	60.31±4.27	66.53±1.26	53.25±2.90	60.60±0.80	61.93±0.68	60.81±2.52	64.99±1.16	61.54±0.51	64.70±0.73
	Grad-MIA	51.51±1.41	22.03±7.08	50.76±0.49	14.29±1.84	49.21±2.79	35.29±4.61	50.74±5.16	16.65±6.53	55.15±4.12	41.17±7.13
	NLO-MIA	59.39±0.69	50.04±1.76	60.75±0.74	53.42±2.11	54.31±0.52	54.51±0.83	55.46±2.76	54.70±3.49	60.85±3.15	61.55±1.99
	GLO-MIA	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00
	GE-MIA	60.79±1.63	67.63±1.97	54.90±1.60	61.06±1.99	51.69±0.83	55.04±3.70	53.34±2.44	58.83±7.07	53.16±1.58	59.99±3.41
	GPIA	72.20±16.41	76.41±15.07	68.58±17.65	48.45±38.29	65.75±4.34	62.19±15.12	61.95±16.17	73.13±8.84	68.35±4.30	65.84±15.79
	MGP-MIA	**81.79**±0.94	**83.99**±0.87	**77.36**±1.31	**80.06**±2.08	**74.77**±0.47	**77.09**±1.06	**74.05**±0.33	**77.23**±0.55	**80.66**±1.43	**82.05**±1.33
BRIDGE	Embedding	66.62±1.02	58.93±1.74	65.61±1.50	52.31±2.95	55.46±0.48	58.65±1.07	59.43±1.15	64.47±1.01	58.94±1.59	64.20±1.17
	Gradient	49.91±1.75	32.35±2.33	50.00±2.30	40.33±11.27	51.45±3.02	51.39±3.68	49.06±5.35	49.60±3.89	45.44±3.32	47.14±3.13
	NLO-MIA	60.97±1.02	53.83±1.53	61.48±1.63	53.53±2.28	52.17±0.56	52.13±1.05	53.38±4.49	55.48±4.72	54.54±3.83	56.39±3.23
	GLO-MIA	50.92±1.28	66.25±0.93	50.52±0.72	66.06±0.94	49.98±0.07	66.62±0.12	50.02±0.04	65.47±2.67	48.16±4.01	66.63±0.08
	GE-MIA	55.75±3.43	59.34±4.37	52.86±2.36	52.77±8.76	50.66±0.30	52.13±2.74	52.63±1.22	62.41±1.84	53.20±1.85	57.58±7.01
	GPIA	66.76±11.87	66.51±13.16	62.77±15.71	46.22±42.40	59.34±5.59	55.36±10.67	53.28±5.64	47.08±30.14	54.38±4.59	38.15±26.29
	MGP-MIA	**81.20**±1.10	**79.97**±1.26	**79.57**±1.25	**80.94**±1.46	**74.93**±0.69	**79.05**±0.28	**70.36**±1.74	**73.06**±2.75	**73.39**±0.43	**76.13**±1.05

Table 1: Membership inference attack performance against **link-prediction-based** multi-domain graph pre-trained models. Best results are in **bold**, and runner-ups are underlined.

(2) constractive-learning-based GFM

Dataset		Cora		CiteSeer		PubMed		Photo		Computers	
Victims	Method	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
GCOPE	Embed-MIA	60.00±22.36	73.33±14.91	50.00±0.00	66.67±0.00	70.32±27.10	76.73±22.27	60.00±22.36	20.00±44.72	70.00±27.39	40.00±54.77
	Grad-MIA	48.42±4.94	50.71±7.92	51.02±3.30	54.16±0.47	57.81±3.86	57.11±6.99	52.47±7.20	55.51±7.53	57.97±4.67	60.53±2.22
	NLO-MIA	54.58±2.80	53.93±3.64	53.97±2.09	53.10±1.54	51.94±0.44	48.64±1.83	55.21±4.53	57.11±4.39	55.22±3.28	57.76±6.51
	GLO-MIA	41.89±7.82	33.12±27.55	44.98±0.69	60.81±1.17	48.64±2.34	65.42±2.16	45.46±6.10	62.30±6.06	49.87±0.16	53.21±29.75
	GE-MIA	51.19±1.36	55.32±9.01	50.63±0.55	38.08±16.77	50.85±0.44	32.04±6.76	50.96±0.94	35.96±27.88	50.71±0.30	16.62±8.18
	GPIA	80.00±27.39	60.00±54.77	70.00±27.39	66.67±40.82	60.00±41.83	40.00±54.77	80.00±27.39	60.00±54.77	**90.00**±22.36	**93.33**±14.91
	MGP-MIA	**87.21**±1.04	**88.13**±0.83	**85.86**±0.75	**87.44**±0.60	**80.20**±1.77	**83.37**±1.17	**83.19**±1.05	**85.04**±0.69	84.80±1.43	86.35±0.84
SAMGPT	Embedding	54.39±9.81	14.10±31.53	51.18±10.02	43.49±33.11	48.61±5.00	16.54±29.76	50.17±1.43	40.22±36.62	49.80±1.14	42.90±32.76
	Gradient	61.82±2.33	60.71±2.50	52.19±2.19	47.71±3.34	50.03±1.22	51.76±2.20	48.22±1.82	52.86±7.42	54.70±3.70	61.21±4.21
	NLO-MIA	52.87±1.97	52.66±3.14	49.84±2.58	49.63±4.49	49.51±0.18	51.07±2.62	49.11±4.85	49.68±4.85	50.53±2.80	50.67±0.77
	GLO-MIA	61.02±0.61	63.52±3.11	55.16±5.31	47.16±30.81	53.71±2.09	59.74±11.10	51.18±1.93	63.45±7.48	50.98±1.23	35.32±30.08
	GE-MIA	73.32±3.88	74.99±4.52	73.97±4.44	75.67±5.64	55.21±0.35	51.03±6.16	50.61±0.66	44.41±11.64	55.77±0.68	55.34±3.39
	GPIA	58.55±2.76	59.11±7.01	55.31±2.30	57.25±5.80	54.59±1.23	61.83±3.24	84.54±4.89	83.94±3.92	73.33±16.28	77.20±11.04
	MGP-MIA	**99.91**±0.20	**99.88**±0.27	**98.83**±1.17	**98.86**±1.14	**91.30**±8.61	**92.37**±7.20	**98.11**±3.22	**98.12**±2.93	**91.72**±17.29	**93.92**±12.38

Table 2: Membership inference attack performance against **contrastive-learning-based** multi-domain graph pre-trained models. Best results are in **bold**, and runner-ups are underlined.

MIA accuracy increases 15% target
both (1) link-prediction-based & (2) constractive-learning-based GFM