ICML
International Conference On Machine Learning

TMLR
TRUSTWORTHY MACHINE LEARNING AND REASONING

On *Strengthening* and *Defending* **Graph Reconstruction Attack** with *Markov Chain Approximation*

Zhanke Zhou
Hong Kong Baptist University
Supervised by Dr. Bo Han

2024 / 01 /18

Paper: https://openreview.net/pdf?id=Vcl3qckVyh Code: https://github.com/tmlr-group/MC-GRA

Outlines

- Background
- Problem Statement & Modeling
- Experiments
- Summary and Discussion

Outlines

- Background
    - model inversion attack: from image to graph
- Problem Statement & Modeling
- Experiments
- Summary and Discussion

Background

Forward pipeline of a neural network:

input

model

output

A neural network

Cat

Background

Question: What if we reverse the pipeline?

output | model | input

A neural network ← Cat

Can we recover the cat image from the trained model? 🤔

What if we reverse the pipeline?

Background

Question: What if we reverse the pipeline?

output | model | input

A neural network ← Cat

Can we recover the cat image from the trained model? 🤔

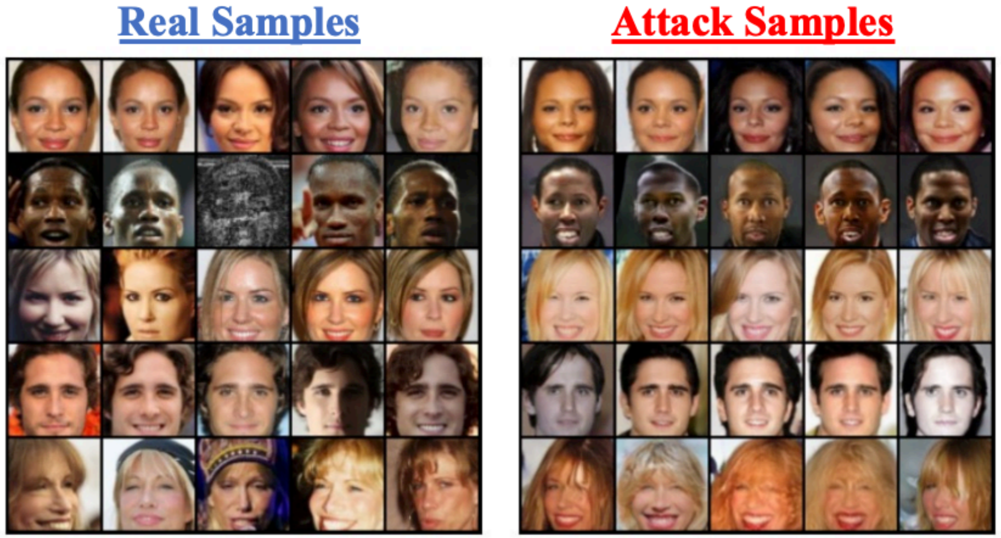
➔ **Yes!** we can recover the training data via **model inversion attack**

Background

Definition of model inversion attack

- a malicious user attempts to **recover** the private data that is used to **train** a neural network

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

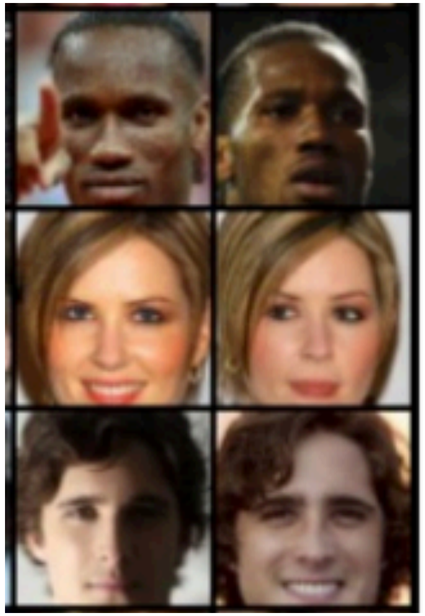


Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. USENIX Security 2014.
Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. CCS 2015.
Variational Model Inversion Attacks. NeurIPS 2021.

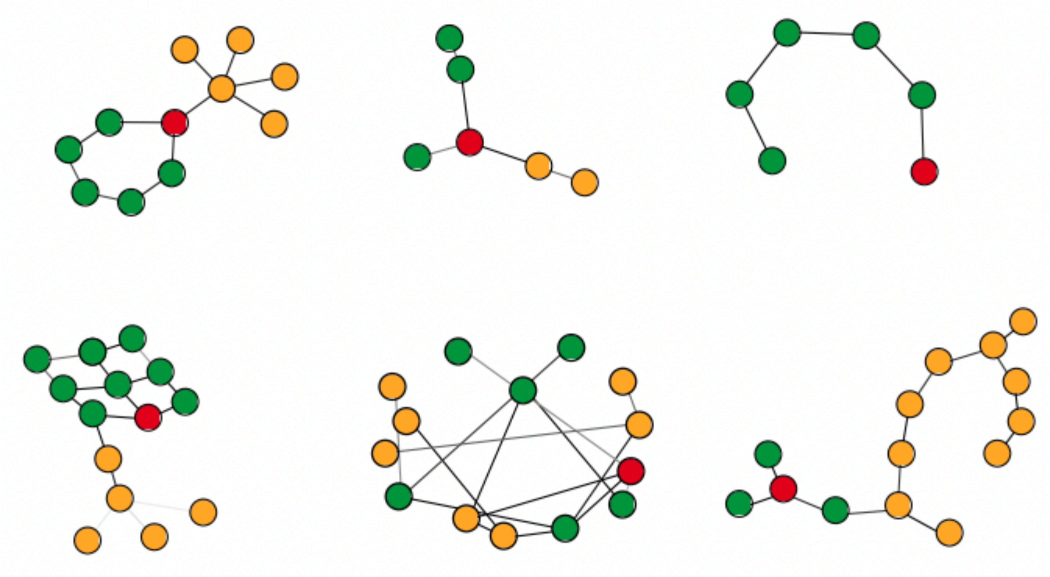
Background

Model inversion attack: from images to graphs

"human faces"



but, what about "graphs"?



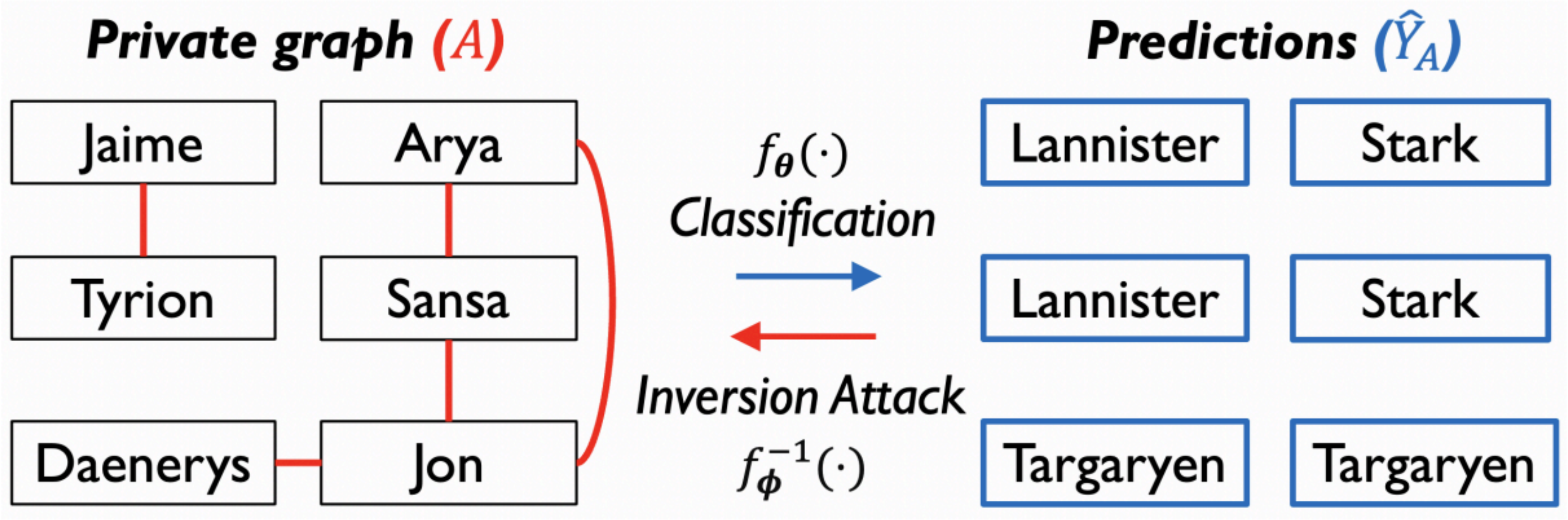
Only **limited** research has been conducted on MIA on graphs 🤔
The **general principles** for strengthening and defending MIA are **unknown**

Outlines

- Background
- Problem Statement & Modeling
  - [problem] graph reconstruction attack: model inversion attack on graphs
  - [modeling] analyze the problem with Markov chain
  - [methods] the corresponding attack and defense methods
- Experiments
- Summary and Discussion

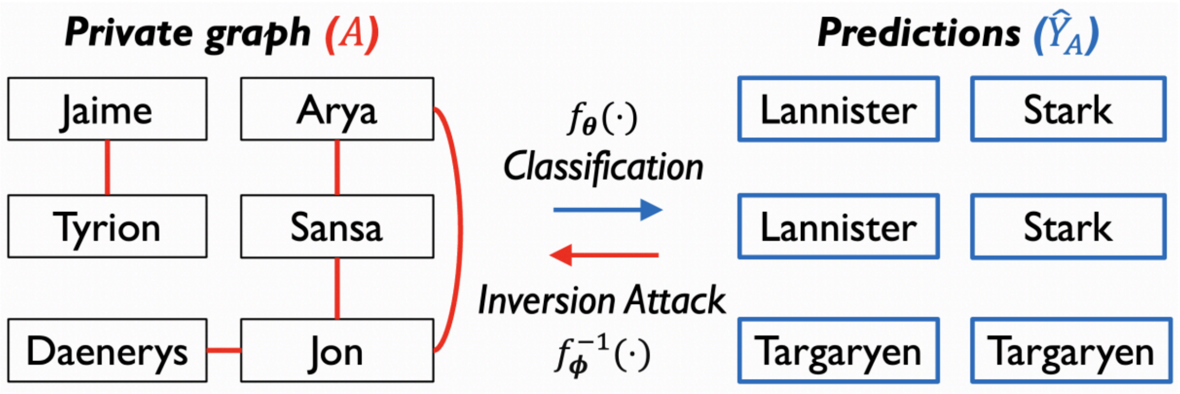
Problem Statement

Graph Reconstruction Attack (GRA):
to recover ***the original adjacency (A)*** *via attacking* ***a trained GNN model (f_θ)***



Problem Statement

Graph Reconstruction Attack (GRA):
to recover ***the original adjacency*** ***(A)*** *via attacking* ***a trained GNN model (f_θ)***



An illustration of GRA

Definition 2.1 (Graph Reconstruction Attack). Given a set of prior knowledge $\mathcal{K}$ and a trained GNN $f_{\theta^*}(\cdot)$, the graph reconstruction attack aims to recover the original linking relations $\hat{\boldsymbol{A}}^*$ of the training graph $\mathcal{G}_{\text{train}} = (A, X)$, namely,

$$\text{GRA:} \quad \hat{\boldsymbol{A}}^* = \arg\max_{\hat{\boldsymbol{A}}} \mathbb{P}(\hat{\boldsymbol{A}} | f_{\theta^*}, \mathcal{K}). \tag{1}$$

Here, $\mathbb{P}(\cdot)$ is the attack method to generate $\hat{\boldsymbol{A}}$, and $\mathcal{K}$ can be any subset of $\{X, Y, \boldsymbol{H}_A, \hat{\boldsymbol{Y}}_A\}$. Note that GRA is conducted in a post-hoc manner, *i.e.*, after the training of GNNs $f_\theta(\cdot)$.

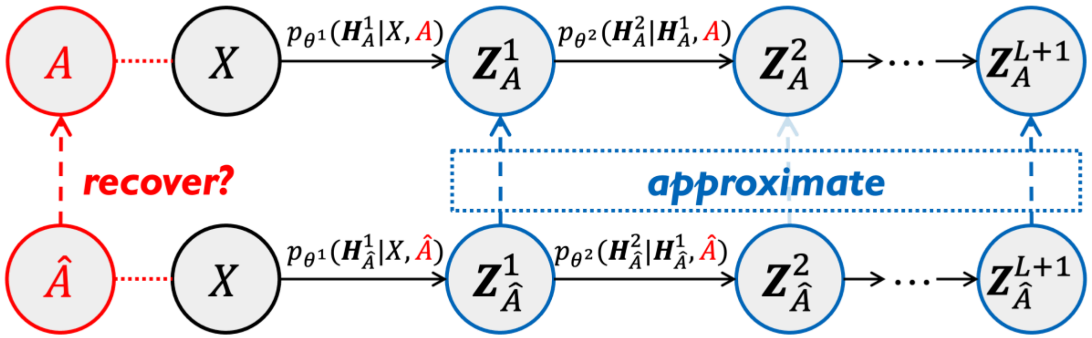
A formal definition

Modeling & Main Results

Markov Chain Modeling:

$$\texttt{ORI-chain}: \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{A} \boldsymbol{Z}_A^1 \xrightarrow[\boldsymbol{\theta}^2]{A} \boldsymbol{Z}_A^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{A} \boldsymbol{Z}_A^{L+1},$$

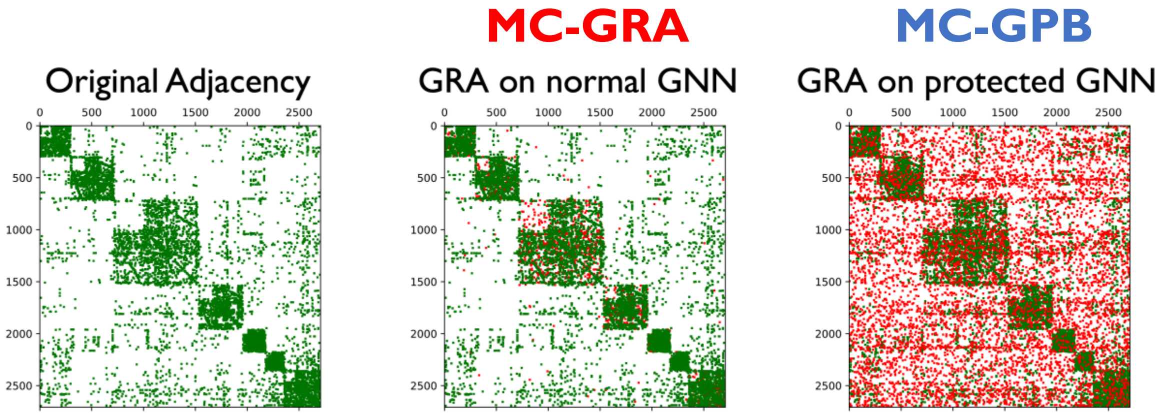
$$\texttt{GRA-chain}: \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{\hat{A}} \boldsymbol{Z}_{\hat{A}}^1 \xrightarrow[\boldsymbol{\theta}^2]{\hat{A}} \boldsymbol{Z}_{\hat{A}}^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{\hat{A}} \boldsymbol{Z}_{\hat{A}}^{L+1},$$



Modeling the GRA problem as approximating the original Markov chain (upper) by the attack chain (lower)

The main results:

- **MC-GRA (a new attack method)**
- **MC-GPB (a new defense method)**



Recovered adjacency on Cora dataset. Green dots are correctly predicted edges while red dots are wrong ones.

A Comprehensive Study of GRA

Based on the **Markov Chain modeling:**

$$\texttt{ORI-chain:}\, \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{A} \boldsymbol{Z}_A^1 \xrightarrow[\boldsymbol{\theta}^2]{A} \boldsymbol{Z}_A^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{A} \boldsymbol{Z}_A^{L+1}$$

Observation 1: a single variable in ORI-chain can recover the original adjacency to some extent

Table 1: Quantitative analysis of $I(A;\boldsymbol{Z})$ with AUC metric under range $[0,1]$. A higher AUC value means a severer privacy leakage. "—" indicates that nodes in this dataset do not have features. Besides, the **boldface** numbers mean the best results, while the underlines indicate the second-bests. The target model $f_{\boldsymbol{\theta}}$ is a two-layer GCN by default.

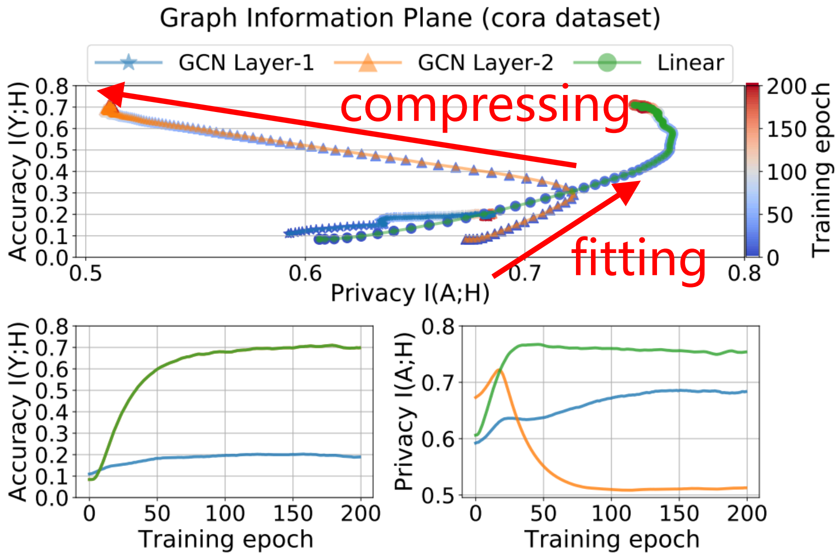
MI	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
$I(A;X)$	.781	**.881**	—	—	—	.521
$I(A;\boldsymbol{H}_A)$	.766	.760	**.763**	**.850**	**.758**	**.584**
$I(A;\hat{\boldsymbol{Y}}_A)$	.712	.743	.772	.826	.732	.561
$I(A;Y)$	**.815**	.779	.705	.728	.613	.536

Observation 2: the linear combination of informative terms only brings marginal improvements in recovering

Table 2: An ensemble study on the prior knowledge with AUC metric. For a generic evaluation, it is assumed that node feature X is accessible (if exists), based on which we evaluate all the possible 8 combinations with 2, 3, or 4 components, where "✓" means accessible to this variable.

X	$\boldsymbol{H}_A$	$\hat{\boldsymbol{Y}}_A$	Y	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
✓	✓			.781	.881	.763	.850	.758	.521
✓		✓		.781	.881	.772	.826	.732	.521
✓			✓	.849	.907	.705	.728	.613	.522
✓	✓	✓		.781	.881	.763	.848	.756	.521
✓	✓		✓	.849	.907	.779	.850	.743	.522
✓		✓	✓	.842	.907	.785	.842	.730	.522
✓	✓	✓	✓	.849	.907	.781	.852	.717	.522

Observation 3: the training procedure contains two main phases, i.e., fitting and compressing



please refer to Sec. 4 in our paper

For enhancing the attack:
To recover better, you must extract more

A chain-based attack method *MC-GRA*
- extract the knowledge stored in target model
- utilize all the prior knowledge simultaneously

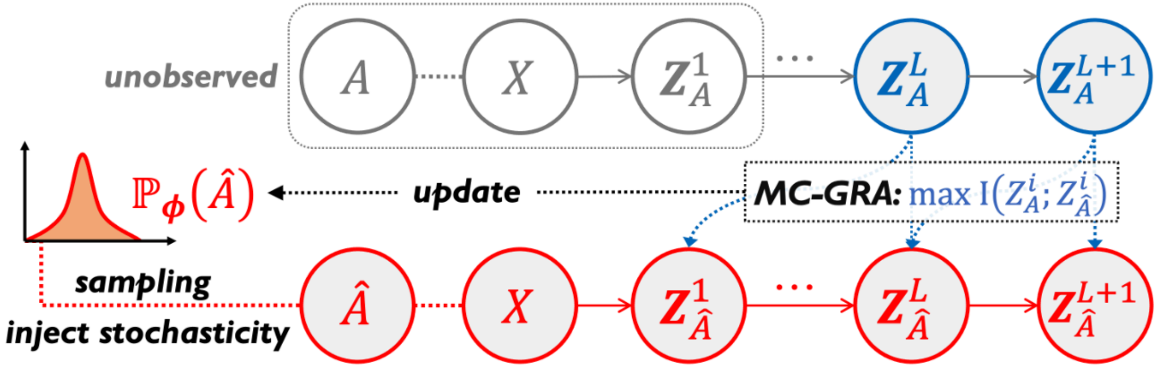
Technical designs
- the objective of enhanced attack
- parametrization of the recovered adjacency
- optimize with injected stochasticity

$$\text{MC-GRA: } \hat{\boldsymbol{A}}^* = \arg\max_{\hat{\boldsymbol{A}}} \sum_{i=1}^{L} \underbrace{\alpha_1^i I(\boldsymbol{H}_A; \boldsymbol{H}_{\hat{\boldsymbol{A}}}^i)}_{\text{propagation approximation}}$$

$$+ \underbrace{\alpha_2 I(\boldsymbol{Y}_A; \boldsymbol{Y}_{\hat{\boldsymbol{A}}}) + \alpha_3 I(Y; \boldsymbol{Y}_{\hat{\boldsymbol{A}}})}_{\text{outputs approximation}} - \underbrace{\alpha_4 H(\hat{\boldsymbol{A}})}_{\text{complexity}}.$$

$$\texttt{ORI-chain:}\, \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{A} \boldsymbol{Z}_A^1 \xrightarrow[\boldsymbol{\theta}^2]{A} \boldsymbol{Z}_A^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{A} \boldsymbol{Z}_A^{L+1}$$

$$\texttt{GRA-chain:}\, \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^1 \xrightarrow[\boldsymbol{\theta}^2]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^{L+1}$$



please refer to Sec. 5 in our paper

For defending the attack:
To learn safer, you must forget more

A chain-based defense method *MC-GPB*

- make the learned representations $\boldsymbol{H}$ contain less information about adjacency A

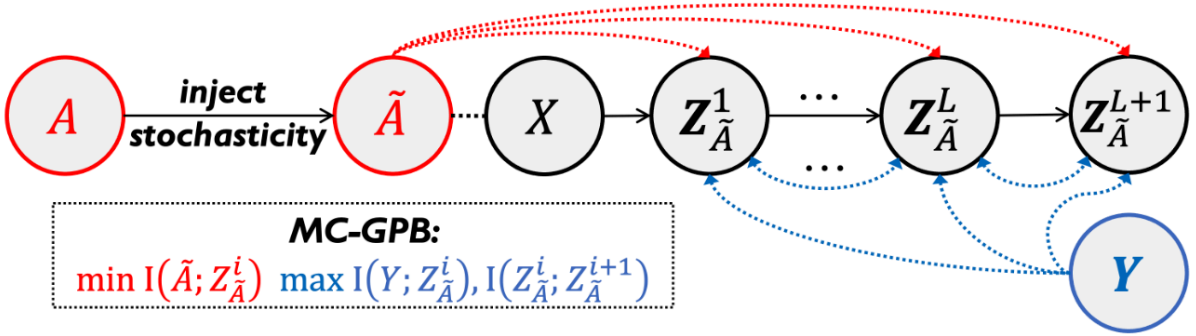
Technical designs

- the objective of defensive training
- differentiable similarity measurements
- optimize with injected stochasticity

$$\texttt{ORI-chain:}\, \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{A} \boldsymbol{Z}_A^1 \xrightarrow[\boldsymbol{\theta}^2]{A} \boldsymbol{Z}_A^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{A} \boldsymbol{Z}_A^{L+1}$$

$$\texttt{GRA-chain:}\, \boldsymbol{Z}^0 \xrightarrow[\boldsymbol{\theta}^1]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^1 \xrightarrow[\boldsymbol{\theta}^2]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^2 \rightarrow \cdots \xrightarrow[\boldsymbol{\theta}^{L+1}]{\hat{\boldsymbol{A}}} \boldsymbol{Z}_{\hat{\boldsymbol{A}}}^{L+1}$$

$$\text{MC-GPB:}\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1} \underbrace{-I(Y;\boldsymbol{H}_A^i)}_{\text{accuracy}} + \underbrace{\beta^i I(A;\boldsymbol{H}_A^i)}_{\text{privacy}} + \sum_{i=1}^{L-1} \underbrace{\beta_c I(\boldsymbol{H}_A^i;\boldsymbol{H}_A^{i+1})}_{\text{complexity}}.$$



please refer to Sec. 6 of our paper

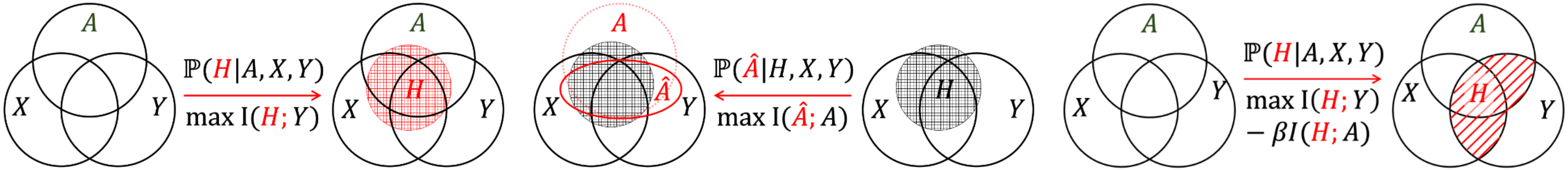
To what extent can we recover or defend?

An information-theoretical analysis

Theorem 5.3. *The layer-wise transformations $Z_A^i \to Z_A^{i+1}$ are non-invertible, e.g., $Z_A^{i+1} = \sigma(\psi(A) \cdot Z_A^i \cdot \theta^i)$, where $\psi(A)$ is the graph convolution kernel, as in Eq.* (2). *It leads to a lower MI between the two Markov chains, i.e., $I(Z_A^i; Z_{\hat{A}}^i) - I(Z_A^{i+1}; Z_{\hat{A}}^{i+1}) \geq 0$. Proof. See Appendix.A.3.*

Theorem 5.4 (Tractable Lower Bound of Fidelity). *The attack fidelity satisfies $I(A; \hat{A}) \geq H(H_A) - H_b(e) - P(e)\log(|\mathcal{H}|)$, where $P(e) \triangleq P(H_A \neq H_{\hat{A}})$ is the probability of approximation error, $\mathcal{H}$ denotes the support of H_A, and $H_b(\cdot)$ is the binary entropy. Proof. See Appendix. A.4.*

Theorem 5.5 (The Optimal Fidelity). *The recovering fidelity satisfies $I(A; X, Y, H_A) - I(A; \hat{A}) \geq 0$. Solving MC-GRA sufficiently yields a solution to achieve the optimal case, i.e., $I(A; \hat{A}^*) = I(A; X, Y, H_A)$. Proof. See Appendix. A.5.*



(a) Standard training (b) Reconstruction attack by MC-GRA (c) Defensive training by MC-GPB

Theorem 6.2 (Maximum Adjacency Information). *The MI between representations H_A and adjacency A satisfies that $I(A; H_A) \leq I(A; A) = H(A)$. Proof. See Appendix. A.6.*

Theorem 6.4 (Minimum Adjacency Information). *For any sufficient graph representations H_A of adjacency A w.r.t. task Y, its MI with A satisfies that $I(A; H_A) \geq I(A; Y)$. The minimum information $I(A; H_A) = I(A; Y)$ can be achieved iff $I(A; H_A|Y) = 0$. Proof. See Appendix. A.7.*

Theorem 6.5. *When degenerating $\beta_c = 0$ and $\beta^i = \beta$, MC-GPB Eq.* (4) *is equivalent to minimizing the Information Bottleneck Lagrangian, i.e., $\mathcal{L}(p(Z|A)) = H(Y|Z) + \beta I(Z; A)$. It yields a sufficient representation Z of data A for task Y, that is an approximation to the optimal representation Z^* in Proposition 6.3. Proof. See Appendix. A.8.*

Outlines

- Background
- Problem Statement & Modeling
- Experiments
- Summary and Discussion

Experiments | quantitative results

Table 3: Results of MC-GRA with standard GNNs. Relative promotions (in %) are computed *w.r.t.* results in Tab. 2.

X	$\boldsymbol{H}_A$	$\hat{\boldsymbol{Y}}_A$	Y	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
✓	✓			.864 (10.6%↑)	.912 (3.5%↑)	.831 (8.9%↑)	.883 (3.8%↑)	.771 (1.7%↑)	.574 (10.1%↑)
✓		✓		.839 (7.4%↑)	.902 (2.3%↑)	.836 (8.2%↑)	.913 (10.5%↑)	.800 (9.2%↑)	.567 (8.8%↑)
✓			✓	.896 (5.5%↑)	.918 (1.2%↑)	.837 (18.7%↑)	.825 (13.3%↑)	.753 (22.8%↑)	.574 (9.9%↑)
✓	✓	✓		.866 (10.8%↑)	.921 (4.5%↑)	.839 (9.9%↑)	.878 (3.5%↑)	.776 (2.6%↑)	.572 (9.7%↑)
✓	✓		✓	.905 (6.5%↑)	.930 (2.5%↑)	.832 (6.8%↑)	.878 (3.5%↑)	.758 (2.0%↑)	.603 (15.5%↑)
✓		✓	✓	.897 (5.6%↑)	.928 (2.3%↑)	.839 (6.8%↑)	.870 (3.3%↑)	.758 (3.7%↑)	.567 (8.6%↑)
✓	✓	✓	✓	.904 (6.4%↑)	.931 (2.6%↑)	.853 (9.2%↑)	.870 (1.9%↑)	.760 (5.9%↑)	.588 (12.6%↑)

Table 4: Results of GRA with MC-GPB protected GNNs. Relative reductions are computed *w.r.t.* results in Tab. 1. $I(A;\boldsymbol{H}_A), I(A;\hat{\boldsymbol{Y}}_A)$ are non-learnable GRA (He et al., 2021a) while $I(A;\boldsymbol{H}^1_{\hat{A}})$ is the learnable GRA (Zhang et al., 2021b).

MI	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
$I(A;\boldsymbol{H}_A)$	.706 (7.8%↓)	.750 (1.3%↓)	.724 (5.1%↓)	.716 (15.8%↓)	.745 (1.7%↓)	.564 (3.4%↓)
$I(A;\hat{\boldsymbol{Y}}_A)$	.704 (0.1%↓)	.730 (1.7%↓)	.705 (8.7%↓)	.587 (28.9%↓)	.692 (5.5%↓)	.559 (0.4%↓)
$I(A;\boldsymbol{H}^1_{\hat{A}})$	.625 (9.9%↓)	.691 (9.8%↓)	.506 (26.3%↓)	.300 (64.5%↓)	.609 (25.1%↓)	.514 (10.6%↓)
Acc.	.734 (3.0%↓)	.602 (4.4%↓)	.830 (1.1%↓)	.391 (16.8%↓)	.808 (5.1%↑)	.668 (0.0%↑)

MC-GRA is better than baseline methods

MC-GPB can defend all the baselines

Experiments | quantitative results

Table 5: Results of MC-GRA with MC-GPB protected GNNs. Relative reductions are computed *w.r.t.* results in Tab. 3.

X	$\boldsymbol{H}_A$	$\hat{\boldsymbol{Y}}_A$	Y	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
✓	✓			.816 (5.5%↓)	.871 (4.4%↓)	.748 (9.9%↓)	.841 (4.7%↓)	.752 (2.4%↓)	.503 (12.3%↓)
✓		✓		.817 (9.7%↓)	.843 (6.5%↓)	.707 (15.4%↓)	.844 (7.5%↓)	.747 (6.6%↓)	.458 (19.2%↓)
✓			✓	.892 (0.4%↓)	.888 (3.2%↓)	.699 (16.4%↓)	.738 (10.5%↓)	.700 (7.0%↓)	.490 (14.6%↓)
✓	✓	✓		.804 (7.1%↓)	.894 (2.9%↓)	.706 (15.8%↓)	.754 (14.1%↓)	.636 (16.7%↓)	.546 (3.7%↓)
✓	✓		✓	.890 (1.6%↓)	.881 (5.2%↓)	.731 (12.1%↓)	.808 (5.6%↓)	.705 (6.9%↓)	.507 (15.9%↓)
✓		✓	✓	.858 (4.3%↓)	.903 (2.6%↓)	.791 (5.7%↓)	.768 (11.7%↓)	.656 (13.4%↓)	.511 (9.8%↓)
✓	✓	✓	✓	.864 (4.4%↓)	.891 (4.2%↓)	.757 (11.2%↓)	.853 (1.9%↓)	.637 (16.1%↓)	.547 (6.9%↓)

MC-GRA and MC-GPB can be generalized to different scenarios

Table 6: MC-GRA with various architectures on Cora.

$\mathcal{K}$	GCN $L=2$	GCN $L=4$	GCN $L=6$	GAT $L=2$	GAT $L=4$	GAT $L=6$	GraphSAGE $L=2$	GraphSAGE $L=4$	GraphSAGE $L=6$
$\{X,Y\}$	.895	.892	.878	.883	.878	.876	.889	.872	.840
$\{X,Y,\boldsymbol{H}_A\}$	.904	.900	.884	.897	.885	.874	.892	.8881	.873
$\{X,Y,\boldsymbol{H}_A,\hat{\boldsymbol{Y}}\}$	.905	.895	.892	.913	.887	.879	.909	.893	.865
Acc.	.792	.661	.248	.637	.651	.630	.614	.443	.145

Table 7: MC-GPB with various architectures on Polblogs.

MI	GCN $L=2$	GCN $L=4$	GCN $L=6$	GAT $L=2$	GAT $L=4$	GAT $L=6$	GraphSAGE $L=2$	GraphSAGE $L=4$	GraphSAGE $L=6$
$I(A;\boldsymbol{H}_A)$	.724	.790	.810	.901	.808	.854	.805	.808	.813
$I(A;\hat{\boldsymbol{Y}}_A)$	.705	.650	.650	.654	.623	.673	.803	.668	.652
$I(A;\boldsymbol{H}_{\hat{\boldsymbol{A}}})$	.506	.577	.532	.542	.656	.536	.599	.769	.468
Acc.	.830	.822	.512	.855	.880	.869	.830	.869	.801

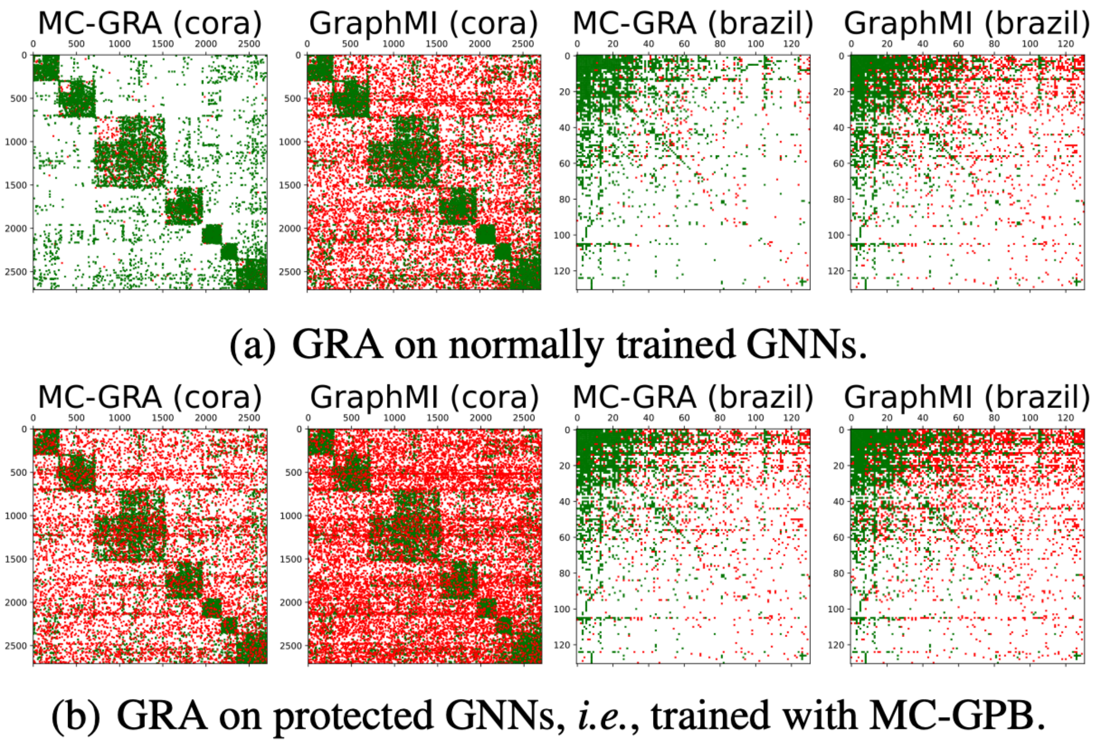
Table 8: Ablation study of two algorithms *w.r.t.* the approximation (*appr.*) and constraint (*cons.*) terms.

variant	Cora	USA	AIDS
MC-GRA (full)	.905	.904	.572
- w/o encoding appr.	.829 (8.3%↓)	.870 (3.7%↓)	.536 (6.2%↓)
- w/o decoding appr.	.854 (5.6%↓)	.849 (6.0%↓)	.490 (14.3%↓)
- w/o complexity cons.	.889 (1.7%↓)	.858 (5.0%↓)	.537 (11.3%↓)
MC-GPB (full)	.745	.391	.668
- w/o accuracy cons.	.681 (8.6%↓)	.369 (5.6%↓)	.625 (6.4%↓)
- w/o privacy cons.	.707 (5.1%↓)	.249 (36.3%↓)	.480 (28.1%↓)
- w/o complexity cons.	.705 (5.4%↓)	.251 (35.8%↓)	.448 (32.9%↓)

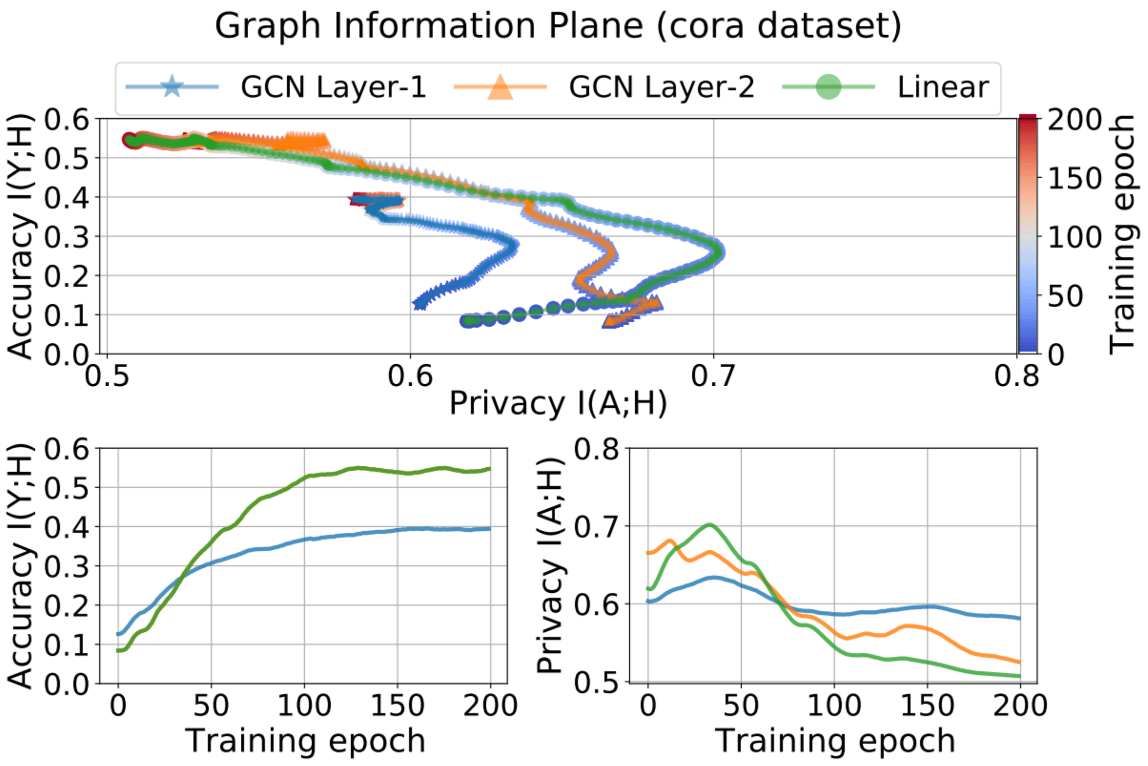
Table 9: Results of removing injecting stochasticity.

type	case	USA	Brazil	AIDS
attack	$\mathcal{K}=\{X,Y\}$	.802 (2.7%↓)	.713 (5.3%↓)	.567 (1.2%↓)
	$\mathcal{K}=\{X,Y,\boldsymbol{H}_A\}$	.856 (1.3%↓)	.740 (2.3%↓)	.572 (5.1%↓)
	$\mathcal{K}=\{X,Y,\boldsymbol{H}_A,\hat{\boldsymbol{Y}}\}$	.864 (0.4%↓)	.730 (3.9%↓)	.567 (3.5%↓)
defense	$I(A;\boldsymbol{H}_A)$	.861 (16.2%↑)	.758 (1.7%↑)	.564 (0.0%↑)
	$I(A;\hat{\boldsymbol{Y}}_A)$	.309 (47.4%↓)	.722 (4.3%↑)	.548 (2.0%↓)
	$I(A;\boldsymbol{H}_{\hat{\boldsymbol{A}}})$	.389 (29.7%↑)	.796 (30.7%↑)	.539 (4.9%↑)
	Acc.	.259 (33.8%↓)	.538 (33.4%↓)	.628 (6.0%↓)

Experiments | qualitative results



Examples of recovered adjacency



Graph information plane: defensive training with MC-GPB

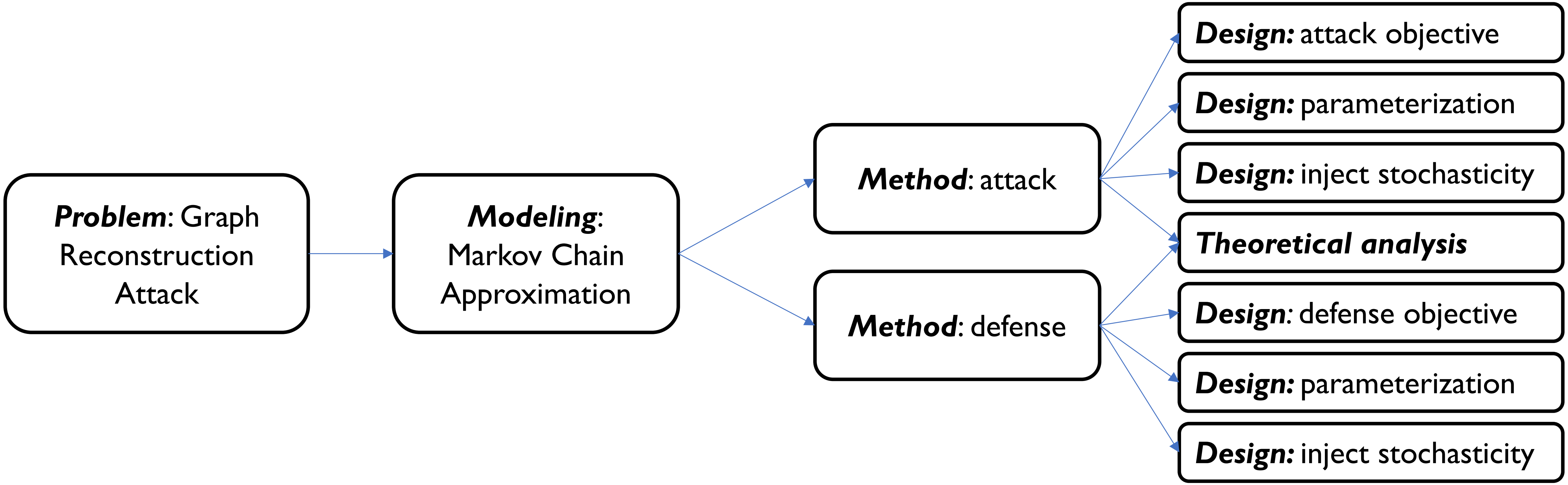
please refer to Sec. 7 of our paper

Outlines

- Background
- Problem Statement & Modeling
- Experiments
- Summary and Discussion

Summary

1. We are the first to conduct a ***systematic*** study of ***GRA*** (***G***raph ***R***econstruction ***A***ttack)
2. We propose a ***attack*** and a ***defense*** method based on ***Markov chain***
3. We provide a ***information-theoretical analysis*** on how to strengthen and defend GRA
4. Both the two proposed methods achieve the ***best results*** on 6 datasets and 3 common GNNs



Potential risk and values

- The MI attack approaches can be **misused** to attack real-world targets

- However, it is important to **raise the awareness** of such an attack
  - inform the community about the risk of privacy leaks, especially the user side
  - e.g., the attack manners and patterns

- More importantly, the inversion attacks can inspire **robust methods**
  - to develop the defensing strategies and to better protect privacy
  - to make the AI products more safe and trustworthy

Potential risk and values

"The gun is not guilty, the person who pulled the trigger is."

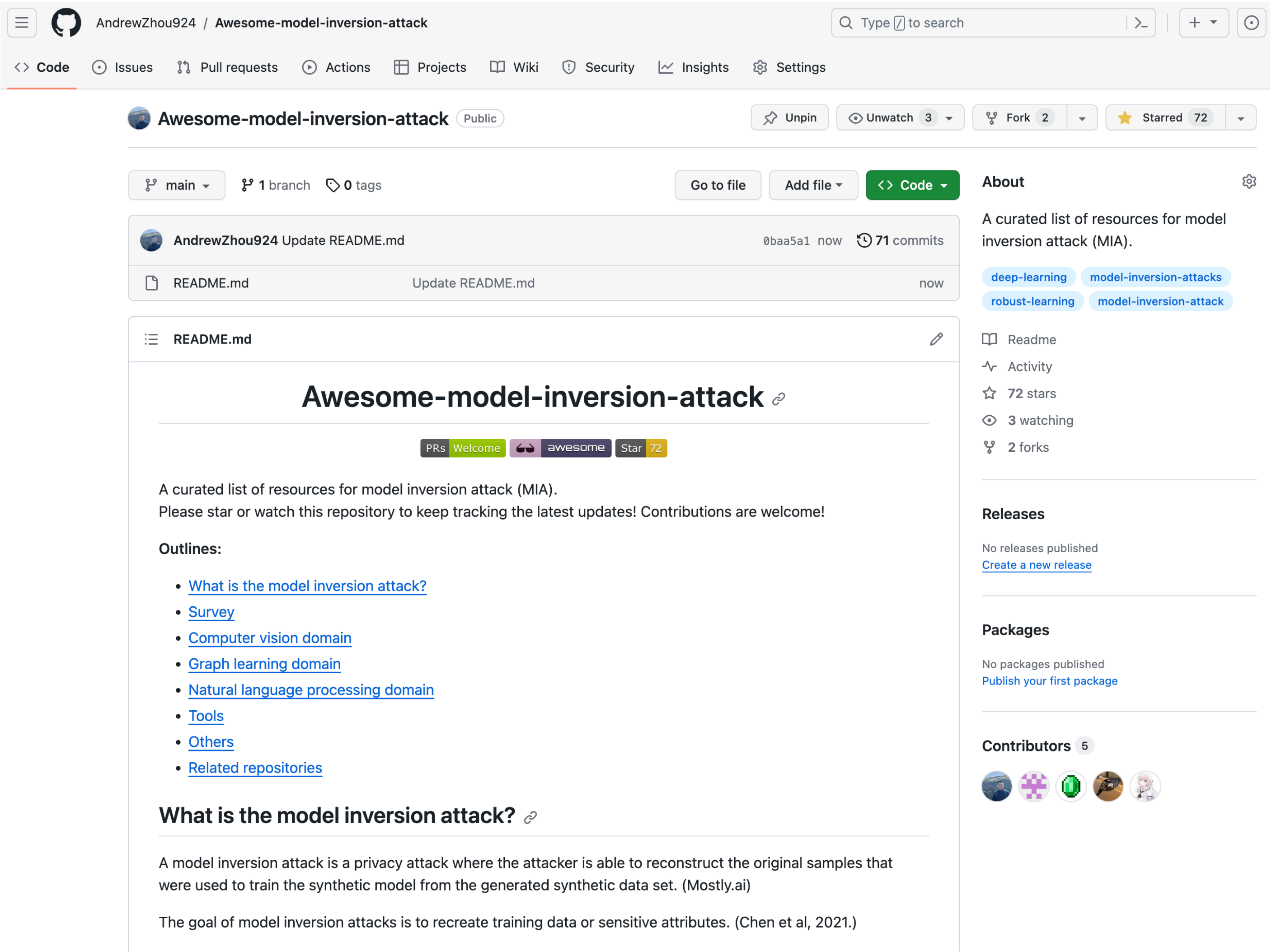
------ by Mikhail Kalashnikov, farther of AK-47

A curated list of resources

include 100+ papers

- computer vision
- natural language processing
- graph learning

https://github.com/AndrewZhou924/Awesome-model-inversion-attack

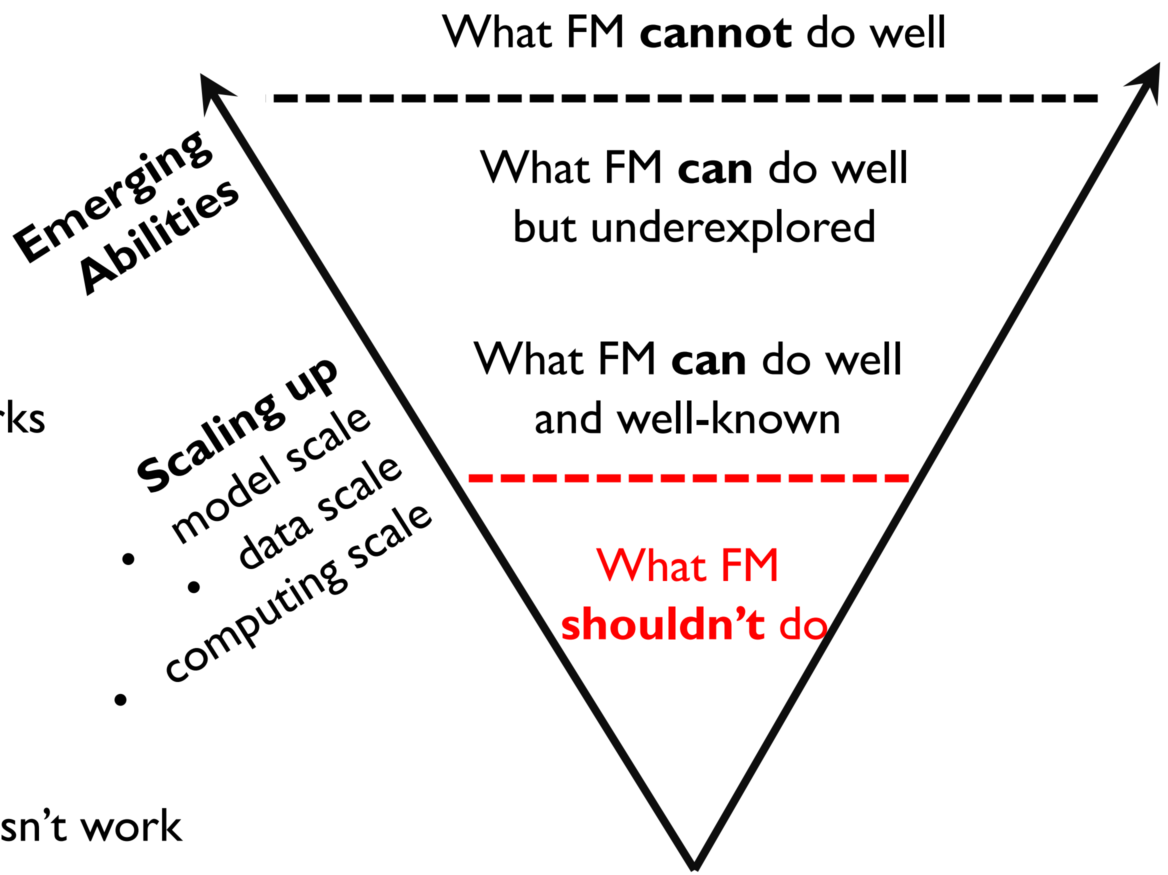


Research scope | Foundation Models

The idea still not works (yet)

The idea works

The idea doesn't work



e.g., algorithmic/complex reasoning

e.g., multi-agent collaboration, predicting future events

e.g., zero/few-shot with in-context learning, traditional supervised learning tasks

e.g., jailbreak, privacy leakage

[1]FM: Foundation Models, including LLM, VLM, etc.

Paper

Code

Q & A

Thanks for your listening!

Email: cszkzhou@comp.hkbu.edu.hk

WeChat: zhouzhanke924