CIKM 2021
1-5 NOVEMBER

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

数据智能系统研究中心
Data Intelligence System Research Center

Single Node Injection Attack against Graph Neural Networks

Shuchang Tao, Qi Cao, Huawei Shen, Junjie Huang, Yunfan Wu, Xueqi Cheng

Data Intelligence System Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

https://github.com/TaoShuchang/G-NIA

SIGIR
Special Interest Group on Information Retrieval

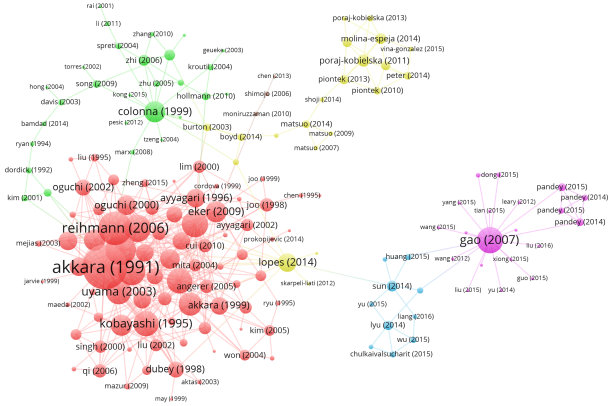
Content

- **Background and Motivation**
- Single Node Injection Attack
- Experiments
- Conclusions

Background: Graphs and GNNs

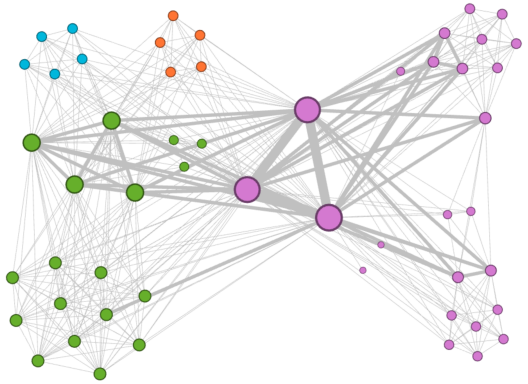
- **Graph data are ubiquitous**
  - Social networks
  - Citation networks
  - Co-purchasing networks
  - Biological networks

Social networks

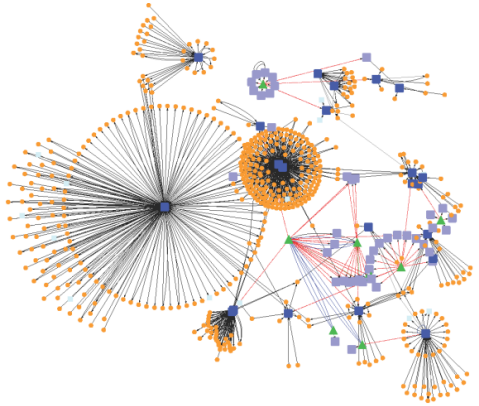


Citation networks

- **GNNs: state of the art on tasks:**
  - Node classification
  - Link prediction
  - Graph classification



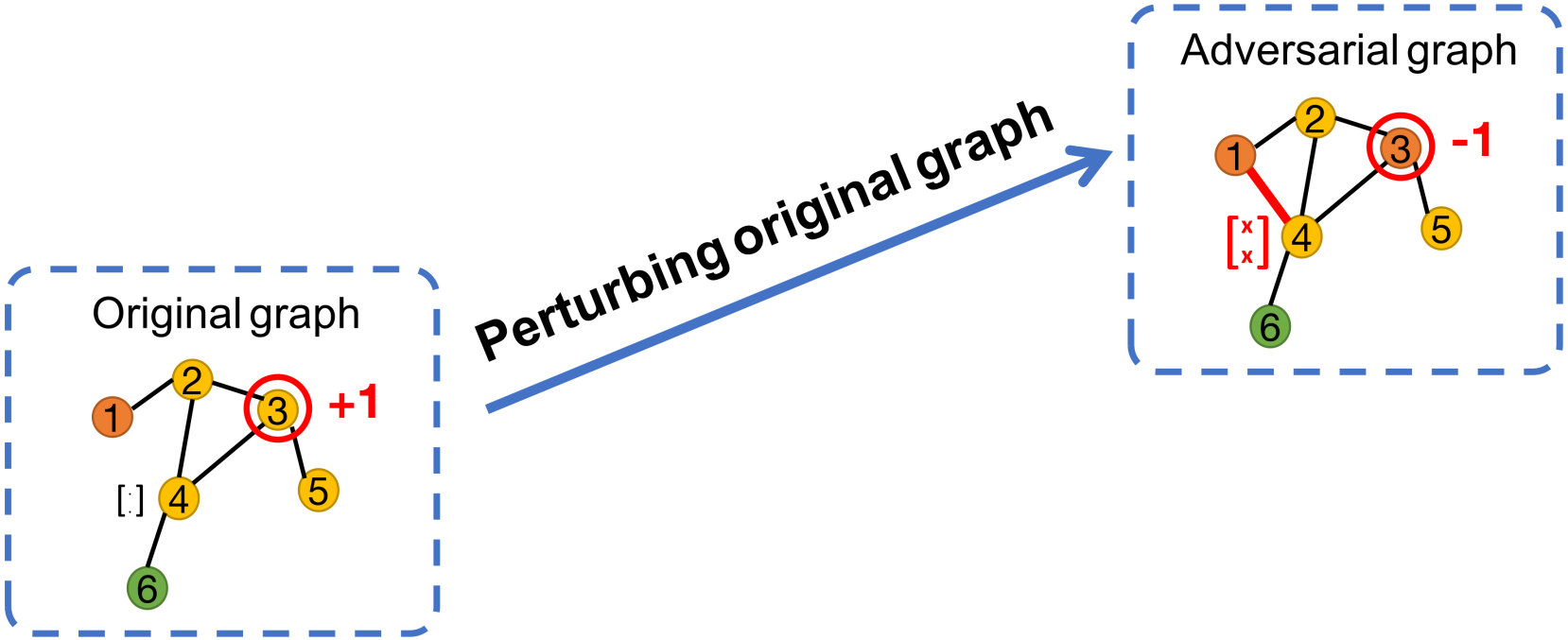
Co-purchasing networks



Biological networks

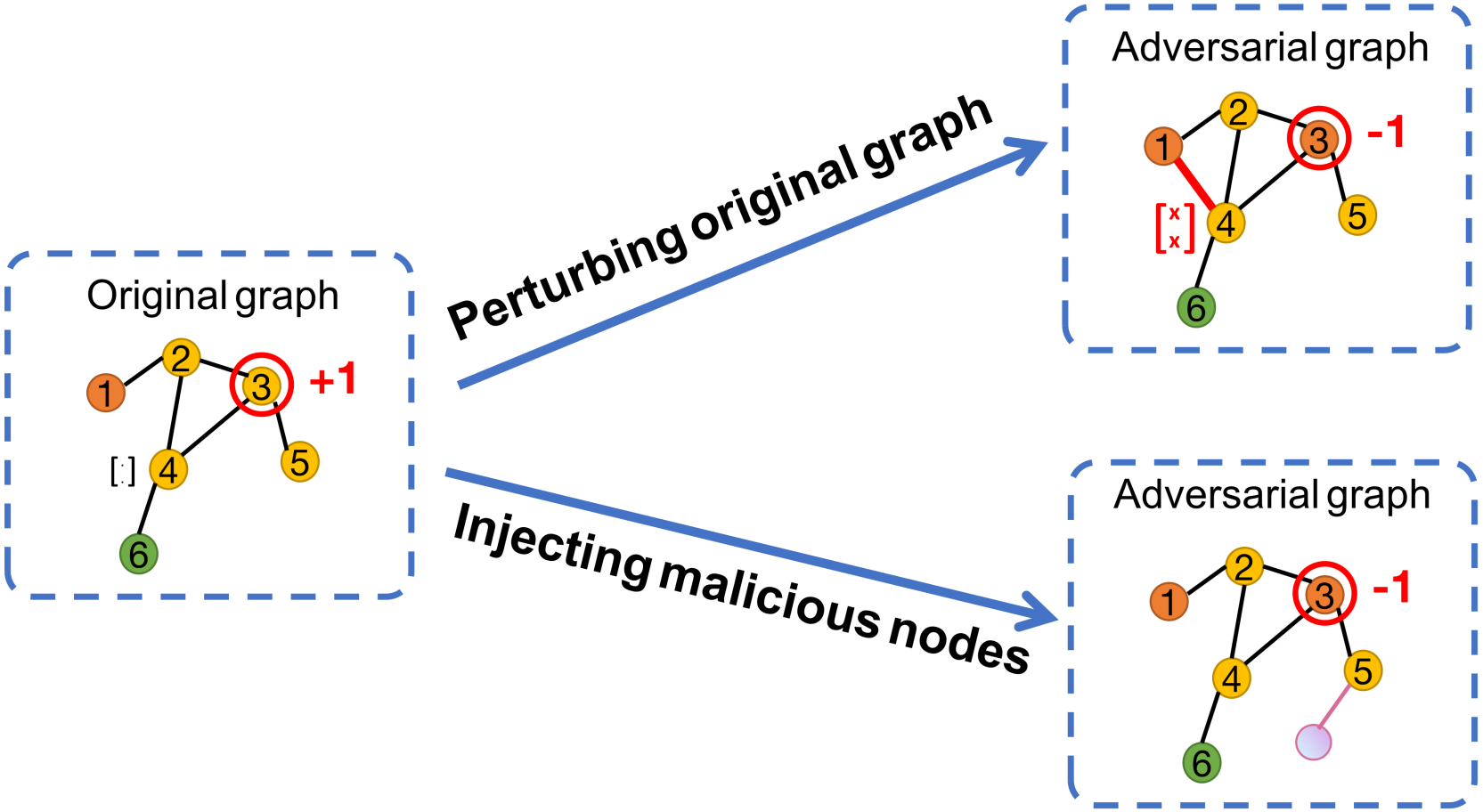
Background: Graph Adversarial Attacks

- GNNs are vurnerable to adversarial attacks



Background: Graph Adversarial Attacks

- GNNs vurnerable to adversarial attacks



Node Injection Attack: Do not modify original nodes or edges

Motivations

- Existing node injection attacks
  - Ignore extremely limited scenarios

    Too many injected nodes

    ⊗ Perceptible / Detected by { victim system / target nodes }

  - Focus on poisoning scenarios

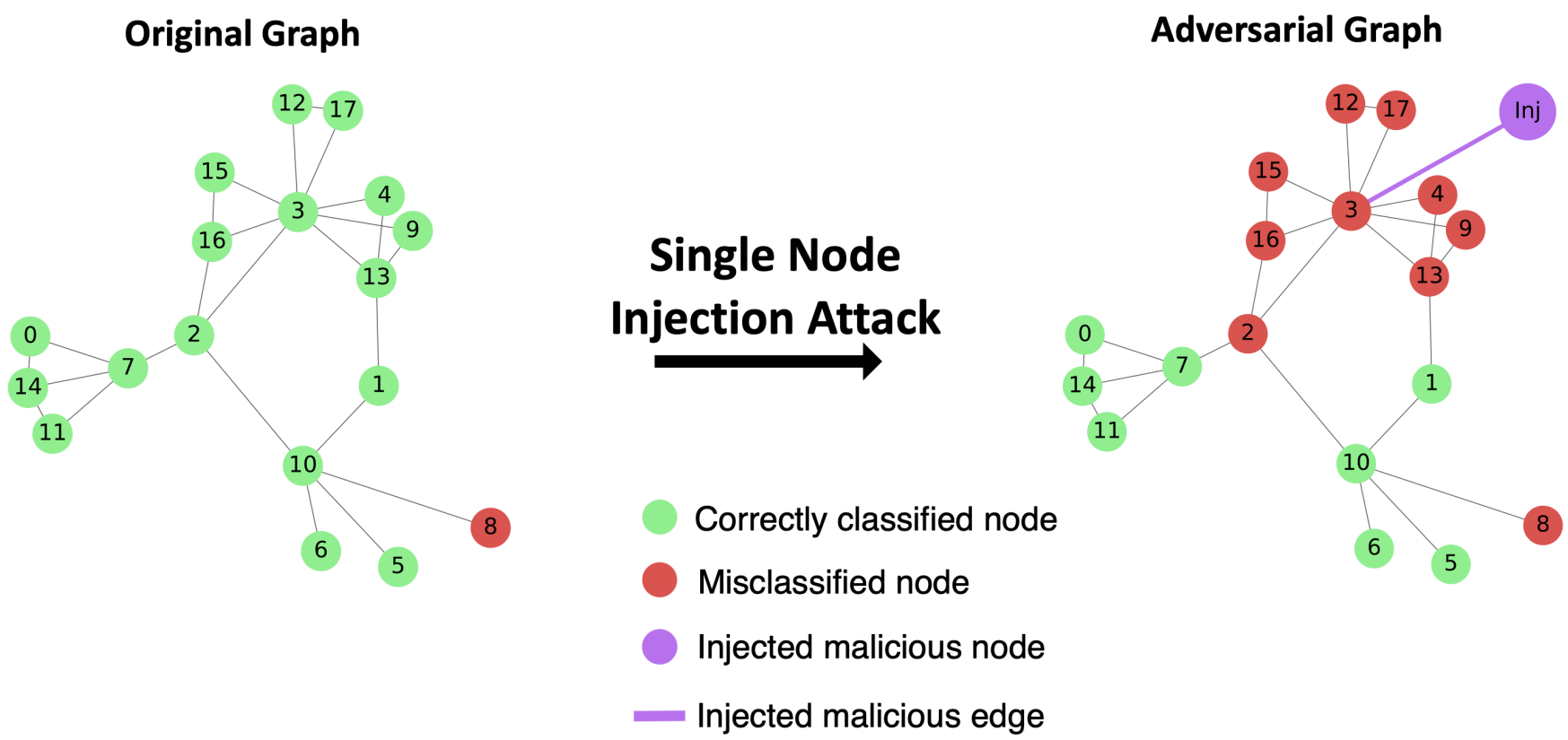
    Require GNN to be trained on injected malicious nodes

    ⊗ Hard to know label standard of victim system

Motivations

- Propose *Single Node Injection Evasion Attack*
  - Inject one single node during the test phase
  - More practical and extremely limited scenario
  - One single malicious follower
    - ✓ Difficult to detect
    - ✓ Low in cost
  - Does not force GNN to train on the unknown labels
    - ✓ More practical

Single Node Injection Attack Example



- Real dataset: Reddit social network
- Injecting 1 node and 1 edge

9 more nodes are misclassified.

Content

- Background and Motivation
- **Single Node Injection Attack**
- Experiments
- Conclusions

Single Node Injection Attack

- **Problem Definition:**

  Max GNN misclassification by injecting one single node in test phase

$$\max_{G'} \sum_{t \in V_{tar}} \mathbb{I}\left(f_{\theta^*}\left(G', t\right) \neq y_t\right)$$

Single Node Injection Attack

- **Problem Definition:**

Max GNN misclassification by injecting one single node in test phase

$$\max_{G'} \sum_{t \in V_{tar}} \mathbb{I}\left(f_{\theta^*}(G', t) \neq y_t\right)$$

Target nodes

Victim GNN trained on original graph G

$$G' = \left(V \cup \boldsymbol{v}_{inj}, E \cup \boldsymbol{e}_{inj}, X \oplus \boldsymbol{a}_{inj}\right)$$

Perturbed graph G'

Original graph G

Single injected node v_{inj}
its edges e_{inj}
and node attributes a_{inj}

Single Node Injection Attack

- **Challenges**
  - **Discrete nature of edges**

    Non-differentiability hinders backpropagating gradients to guide the edge optimization

  - **Attributes of different network datasets vary greatly**
    - Continuous attributes: Most networks
    - Discrete attributes: Citation networks

  - **Coupling effect between network structure and node attributes**

Optimization-based Node Injection Attack

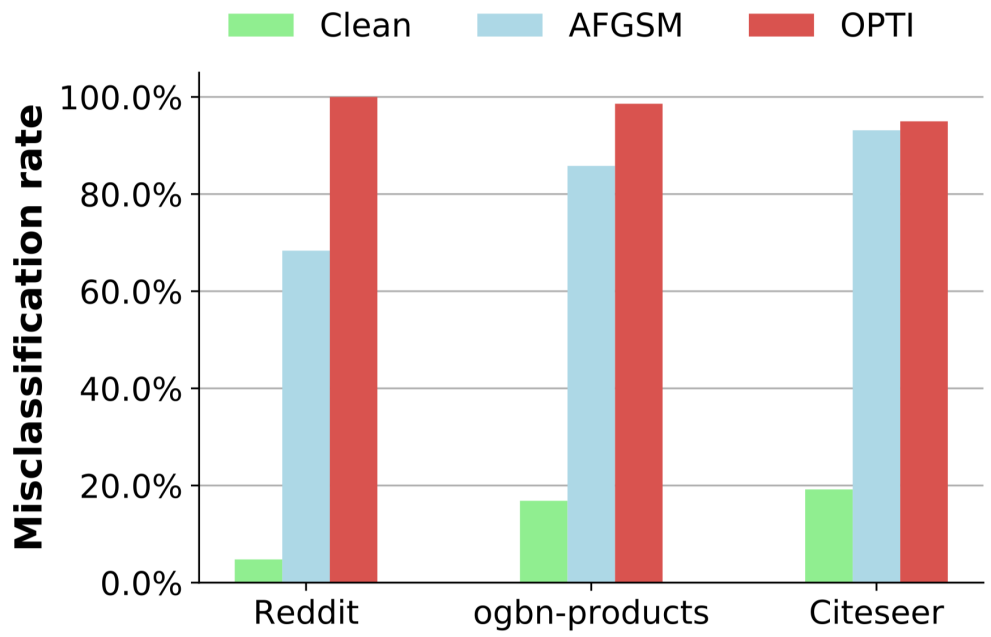
- **Optimization-based Node Injection Attack (OPTI)**
  - **Directly optimize the attributes and edges of the malicious injected node**

Optimization-based Node Injection Attack

- **Optimization-based Node Injection Attack (OPTI)**
  - **Directly optimize the attributes and edges of the malicious injected node**

- **Excellent Performance**

  Inject one node and one edge to attack one target node

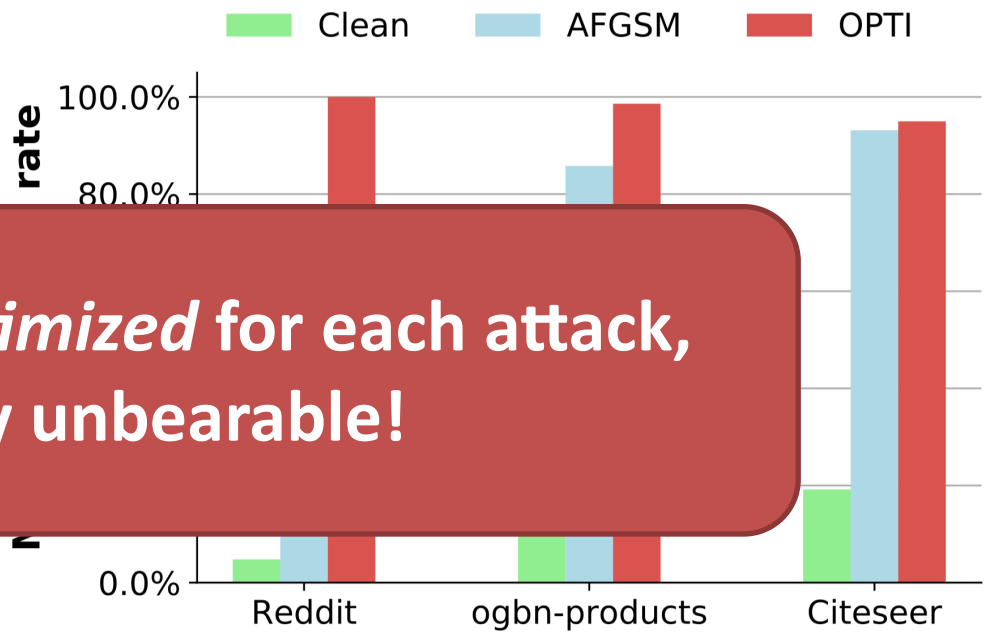


Optimization-based Node Injection Attack

- **Optimization-based Node Injection Attack (OPTI)**
  - **Directly optimize the attributes and edges of the malicious injected node**

- **Excellent Performance**

  Inject one node and one edge to attack one target node



However, OPTI needs to be *re-optimized* for each attack, which is computationally unbearable!

Generalizable Node Injection Attack (G-NIA)

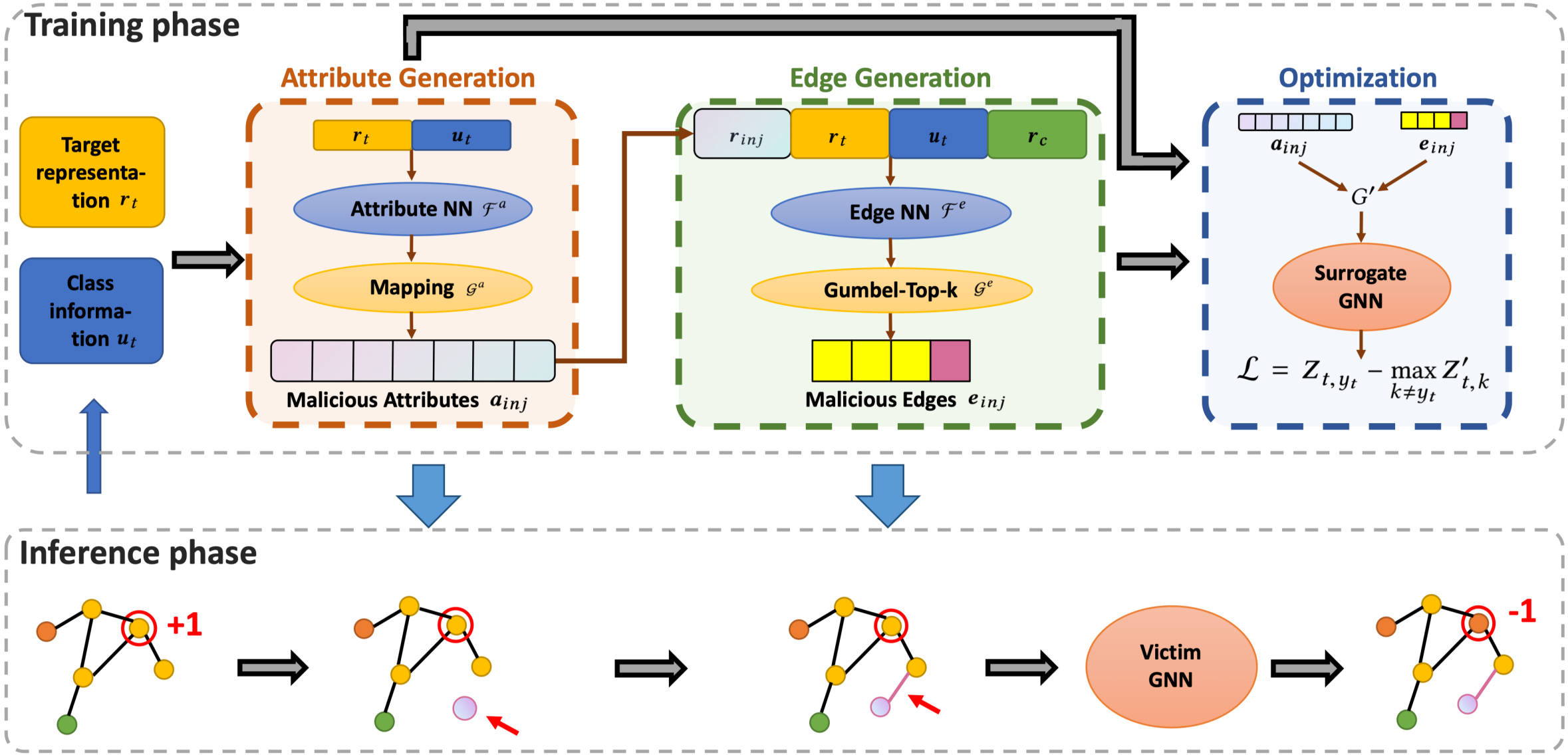
- To solve the dilemma, we further propose a *Generalizable Node Injection Attack* model, namely G-NIA

- Aim:

  Improve the efficiency while ensuring attack performance

- Advantages:
  - Representation ability <= Jointly modeling
  - High efficiency <= Model-based framework

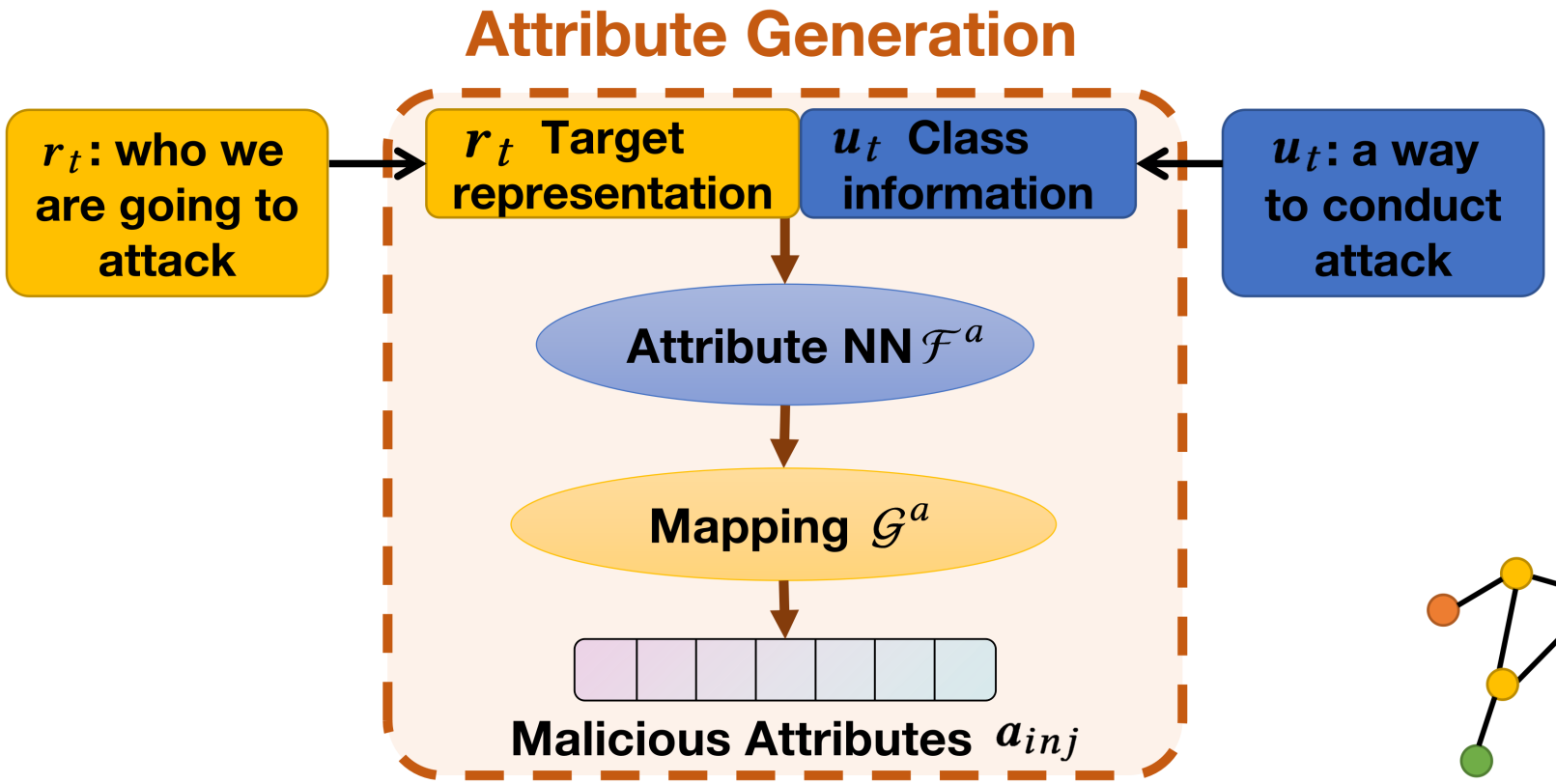
Generalizable Node Injection Attack (G-NIA)

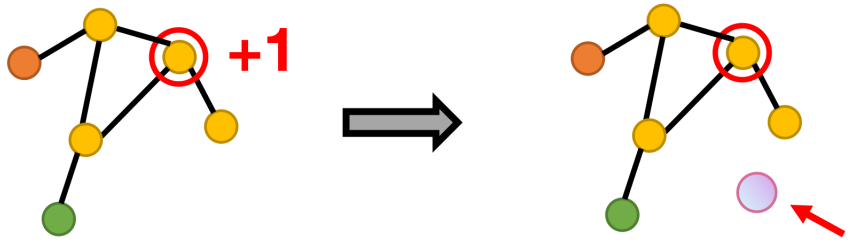


Framework of G-NIA model

Attribute Generation

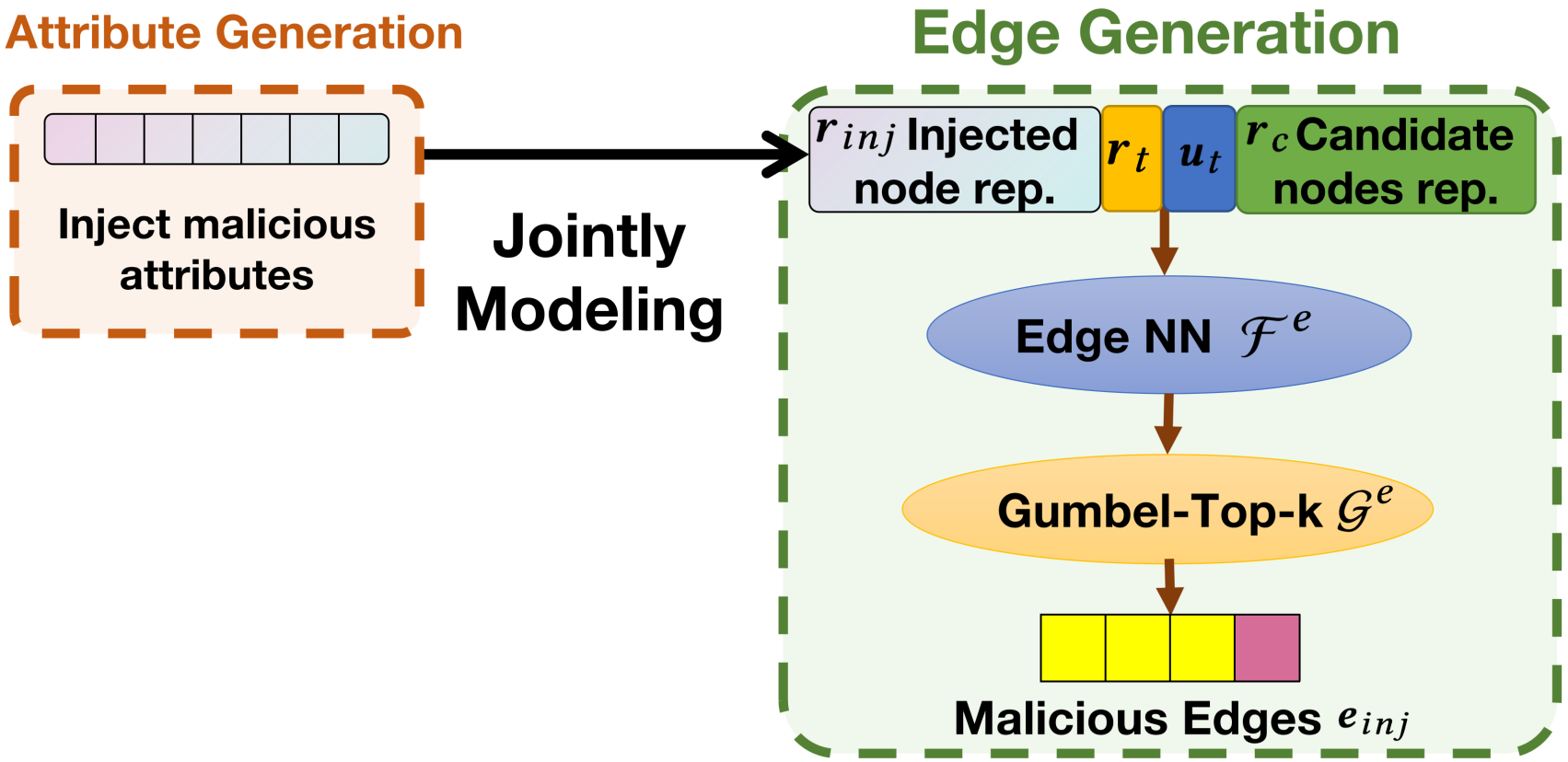
- Generate malicious attributes based on target node representation and class information





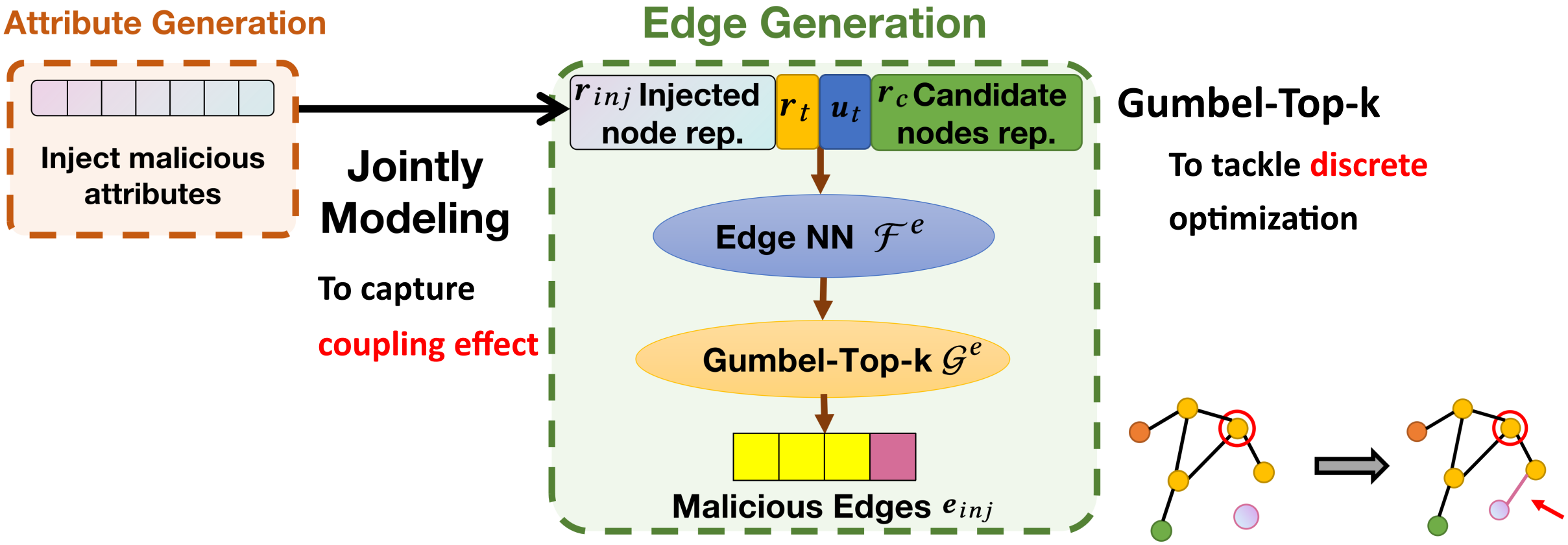
Edge Generation

- Generate malicious edges via **jointly modeling** and **Gumbel-Top-k technique**.



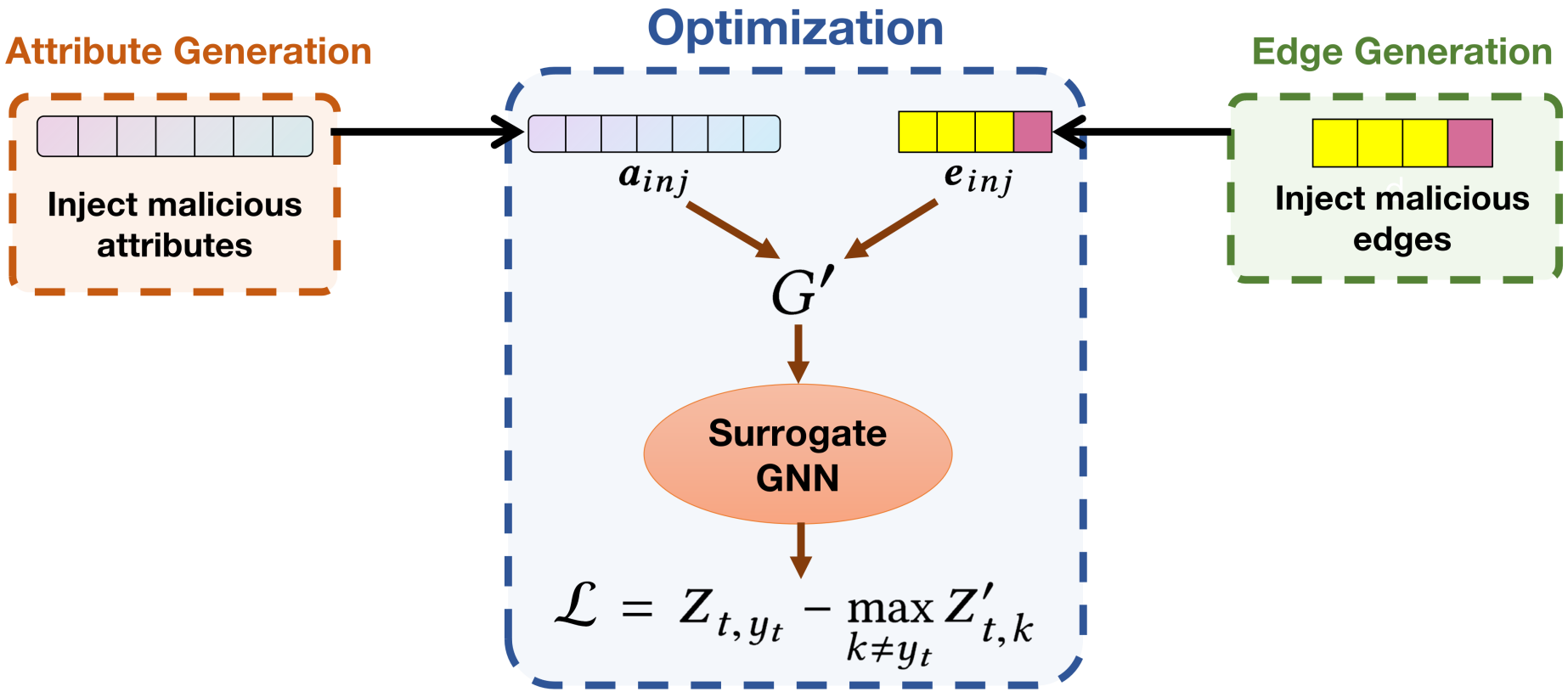
Edge Generation

- Generate malicious edges via **jointly modeling** and **Gumbel-Top-k technique**.



Optimization

- Inject malicious node and edges to obtain perturbed graph
- Feed perturbed graph into surrogate GNN to compute attack loss and optimize G-NIA.



Content

- Background and Motivation
- Single Node Injection Attack
- **Experiments**
- Conclusions

Experiments

- **Datasets:**
  - Social network: Reddit
  - Products co-purchasing network: ogbn-products
  - Citation network: Citeseer
- **Baselines:**
  - Random and Heuristic-based methods
  - SOTA: NIPA, AFGSM
- **Targeted attack**
- **Victim GNN:**
  - GCN, GAT, APPNP

Attack Performance (Effectiveness)

- G-NIA shows tremendous attack performance
- G-NIA outperforms all the baselines on all datasets attacking all the three GNNs,

	GCN			GAT			APPNP		
	Reddit	ogbn-products	Citeseer	Reddit	ogbn-products	Citeseer	Reddit	ogbn-products	Citeseer
Clean	8.15%	21.01%	22.04%	8.45%	24.34%	21.09%	6.65%	18.58%	19.91%
Random	11.33%	27.65%	24.05%	10.45%	29.60%	22.23%	8.05%	22.44%	20.83%
MostAttr	13.53%	31.82%	27.80%	12.30%	33.18%	24.41%	9.95%	26.56%	22.80%
PrefEdge	8.53%	23.05%	22.75%	8.70%	26.76%	21.64%	6.77%	20.11%	20.21%
NIPA	10.54%	22.20%	22.99%	10.19%	25.49%	21.56%	9.55%	20.10%	19.91%
AFGSM	68.57%	85.76%	84.60%	74.11%	80.28%	57.11%	45.43%	74.18%	**41.47%**
G-NIA	**99.90%**	**98.76%**	**85.55%**	**99.85%**	**93.23%**	**73.70%**	**91.05%**	**98.67%**	**41.47%**

Misclassification rate of Single Target Attack across various GNNs on Reddit, ogbn-products and Citeseer

Efficiency Analysis

- **G-NIA is much more efficient than OPTI**
  - **OPTI takes 1.17 seconds**
  - **G-NIA takes 0.00253 seconds, nearly 500 times faster than OPTI.**
  - **On all test data, G-NIA takes 5 seconds, while OPTI needs 2333 seconds.**

	AFGSM	NIPA	OPTI	G-NIA
Time (s)	2.72×10^{-2}	1.47×10^{-3}	1.17	2.53×10^{-3}

Running time comparison on Single Target Attack on GCN on Reddit

Flexible Attack Scenario

- **G-NIA is capable for various attack scenarios, such as Multiple Targets Attack and Black-box Attack.**

	Reddit	ogbn-products	Citeseer
Clean	3.89%	16.28%	19.70%
Random	5.43%	21.69%	16.49%
MostAttr	6.40%	24.08%	18.19%
PrefEdge	5.59%	22.02%	16.24%
NIPA	13.96%	21.56%	16.09%
AFGSM	51.86%	87.34%	63.22%
G-NIA	**92.62%**	**95.42%**	**63.51%**

Misclassification rate of Multiple Targets Attack on GCN on Reddit, ogbn-products, and Citeseer

		Reddit	ogbn-products	Citeseer
GAT	Clean	8.45%	24.34%	21.09%
	NIPA	12.54%	25.92%	21.09%
	AFGSM	45.83%	74.85%	**38.63%**
	G-NIA	**94.60%**	**95.00%**	36.49%
APPNP	Clean	6.65%	18.58%	19.91%
	NIPA	9.90%	21.15%	20.14%
	AFGSM	33.38%	67.65%	**30.09%**
	G-NIA	**99.85%**	**99.14%**	28.91%

Black-box Attack:
Misclassification rate of Single Target Attack on GCN on three datasets

Content

- Background and Motivation
- Single Node Injection Attack
- Experiments
- **Conclusions**

Conclusions

- **Focus on extremely limited scenario, and propose *single node injection evasion attack scenario*.**
- **Propose OPTI to explore attack performance upper bound.**
- **Propose *Generalizable Node Injection Attack model (G-NIA)***
  - **Effectiveness:**

    G-NIA shows tremendous attack performance, and significantly outperforms SOTA.
  - **Efficiency:**

    G-NIA is 500 times faster than OPTI in the test phase.
  - **Flexibility:**
    - Experiments conducted on three kinds of network datasets across three well-known GNN models shows the superiorityof G-NIA in two attack scenarios.
    - G-NIA is capable in black-box scenario.

Thank you!
Welcome to contact us!

Shuchang Tao
taoshuchang18z@ict.ac.cn

Qi Cao

Huawei Shen

Junjie Huang

Yunfan Wu

Xueqi Cheng

Bingbing Xu

Data Intelligence System Research Center, ICT, CAS

- **Research field**
  - **Data Intelligence**
    - Graph Neural Networks
  - **Social Media Computing**
  - **Trustable Machine Learning**
    - Adversarial Robustness

SIGIR
Special Interest Group on Information Retrieval