# More is Better (Mostly): On the Backdoor Attacks in Federated Graph Neural Networks

Jing Xu [1], Rui Wang [1], **Stefanos Koffas** [1], Kaitai Liang [1], Stjepan Picek [1, 2]

[1]Deflt University of Technology

[2]Radboud University

Annual Computer Security Applications Conference (ACSAC) 2022

# Outline

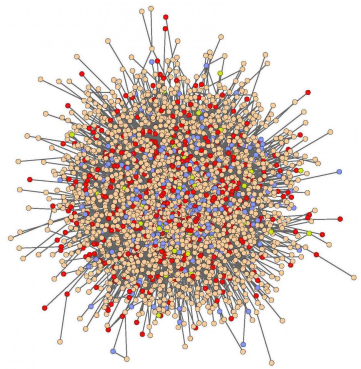
# Outline

# Graph Neural Netowrks (GNN)

## Networks are "everywhere"

- Physical networks

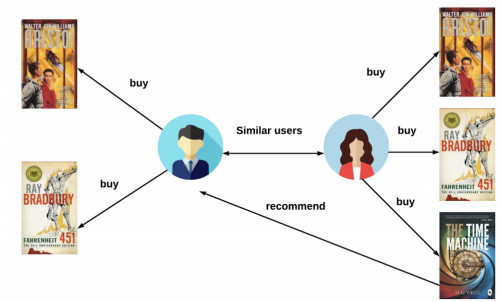
(a) Transportation Network



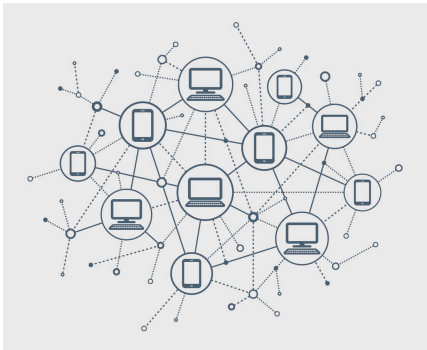
(b) Molecular Network

- Model complex relationships

(a) Social Network




(b) User-Item Network



(c) Web Network

# Graph Neural Networks

**Graph Neural Network** is a type of Neural Network which directly operates on the graph structure.

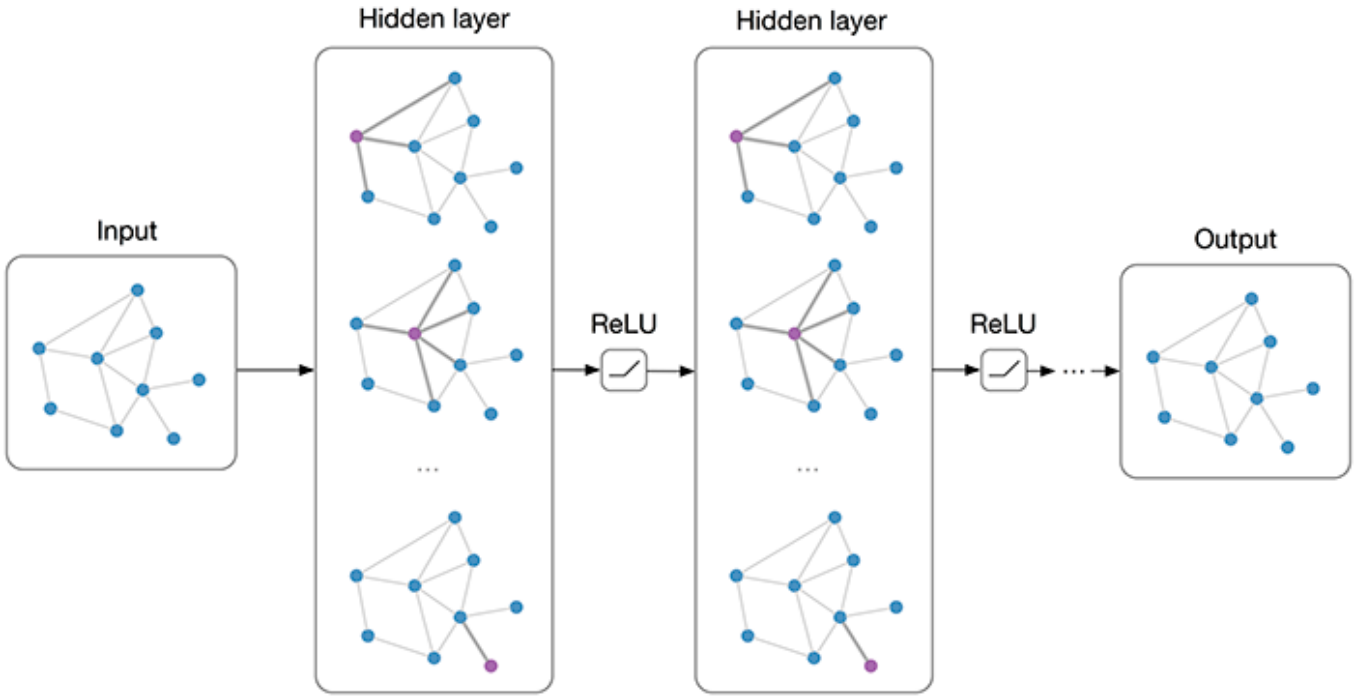



Figure 3: Multi-layer Graph Convolutional Network (GCN).

# Federated Learning

A distributed learning paradigm that enables multiple clients to train a global model collaboratively without revealing local datasets.

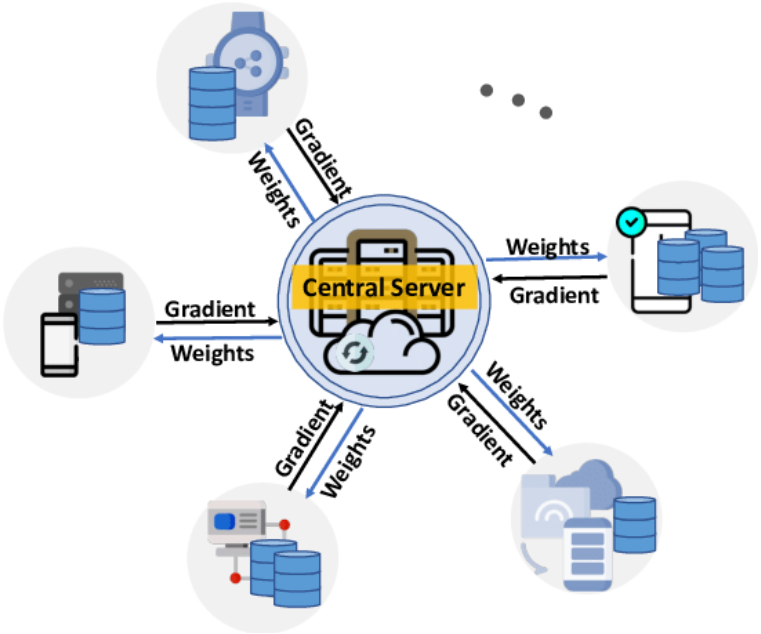



Figure 4: Federated Learning Framework.

- Ensures data privacy, data security, data access rights and allows usage of heterogeneous data.
- Cross-device setting (e.g., android keyboard), cross-silo setting (e.g., drug discovery from different pharmaceutical institutions).

# Why Federated GNNs?

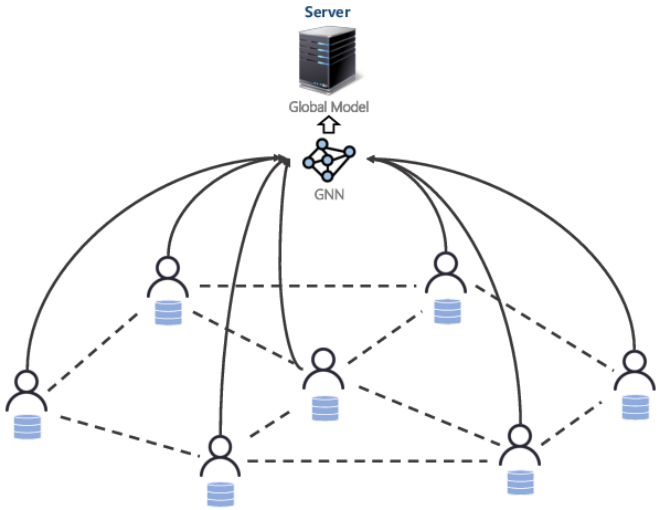



Figure 5: Federated GNNs.

- E.g., when pharmaceutical institutions want to collaborate for drug discovery but cannot share their data
- FL is a promising solution for training GNNs over isolated graph data[1][2][3]

[1] "Fedgraphnn: A federated learning system and benchmark for graph neural networks" (2021). In: *arXiv*

[2] "Peer-to-peer federated learning on graphs" (2019). In: *arXiv*

[3] "Federated Graph Learning–A Position Paper" (2021). In: *arXiv*

# Backdoor Attacks

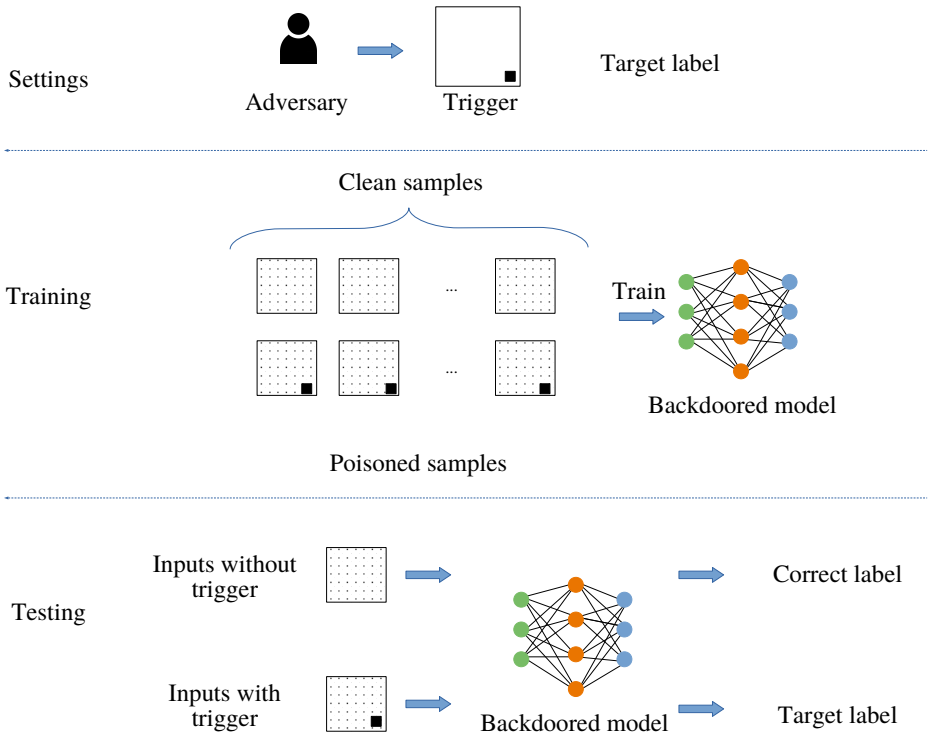



Figure 6: Backdoor attack framework.

- Backdoor attacks aim to make a model misclassify its inputs to a preset-specific label without affecting its original task.
- Attackers poison the model by injecting triggers into the training data that activate the backdoor in the test phase.

# Prior Works on Backdoor Attack in GNNs and Federated Learning

- Backdoor attack in GNNs [4] [5] [6]
  - Focus on GNN models in centralized learning
  - The trigger is a subgraph which is defined by:
    - **Trigger size**: number of nodes.
    - **Trigger density**: the complexity of the subgraph (from 0 to 1).
- Backdoor attack in federated learning
  - Euclidean data, e.g., images and words [7] [8] [9]

---

[4] Z. Zhang, J. Jia, B. Wang, and N. Z. Gong (2021). "Backdoor attacks to graph neural networks". In: *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*

[5] Z. Xi, R. Pang, S. Ji, and T. Wang (2021). "Graph backdoor". In: *USENIX Security*

[6] J. Xu, M. Xue, and S. Picek (2021). "Explainability-based backdoor attacks against graph neural networks". In: *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*

[7] "How to backdoor federated learning" (2020). In: *AISTATS*

[8] "Analyzing federated learning through an adversarial lens" (2019). In: *ICML*

[9] "Dba: Distributed backdoor attacks against federated learning" (2019). In: *ICLR*

# Our Work

## Backdoor Attacks in Federated GNNs

**Challenges**

- The malicious updates will be weakened in the aggregation function.
- The backdoor trigger generation methods and injecting position are different between graph data and images/words.
  - There is no position information we can exploit in the graph data because it's a non-Euclidean data.
- Current defenses may not be effective in backdoor attacks in Federated GNNs.

# Our Work: Contributions

- Explore two types of backdoor attacks in Federated GNNs, i.e., distributed backdoor attack (DBA) and centralized backdoor attack (CBA).
- Perform ablation study to illustrate the impact of many different parameters on the backdoor attack performance.
- Explore the robustness of backdoor attacks in Federated GNNs against state-of-the-art defenses.

# Outline

# The Threat Model

- **Adversary's capability**
  - The adversary $\mathcal{A}$ can corrupt $M$ ($M \leq K$) clients to perform DBA.
  - A complete attack in every round [10]
  - The adversary cannot impact the aggregation process on the central server nor the training or model updates of other clients.
- **Adversary's knowledge**
  - The compromised clients' training dataset.
- **Adversary's goal**
  - Make the global model misclassify the backdoored data samples into specific pre-determined labels
  - Without affecting the accuracy on clean data.

[10] "Dba: Distributed backdoor attacks against federated learning" (2019). In: *ICLR*

# Outline

# How to Design Backdoor Attacks in Federated GNNs

**Definition (Local Trigger & Global Trigger.)**

The local trigger is the specific graph trigger for each malicious client in DBA. The global trigger is the combination of all local triggers.[a]

[a] Since it is an NP-hard problem to decompose a graph into subgraphs [11], we first generate local triggers and then compose them to get the global trigger used in CBA.

[11] S. Dasgupta, C. H. Papadimitriou, and U. V. Vazirani (2008). *Algorithms*. McGraw-Hill Higher Education New York

# How to Design Backdoor Attacks in Federated GNNs

**Definition (Distributed Backdoor Attack (DBA).)**

There are multiple malicious clients, and each of them has its local trigger. An adversary $\mathcal{A}$ conducts DBA by compromising at least two clients in FL.

**Definition (Centralized Backdoor Attack (CBA).)**

A global trigger consisting of local triggers is injected into one client's local training dataset. An adversary $\mathcal{A}$ conducts CBA by usually compromising only one client in FL.

# Attack Framework

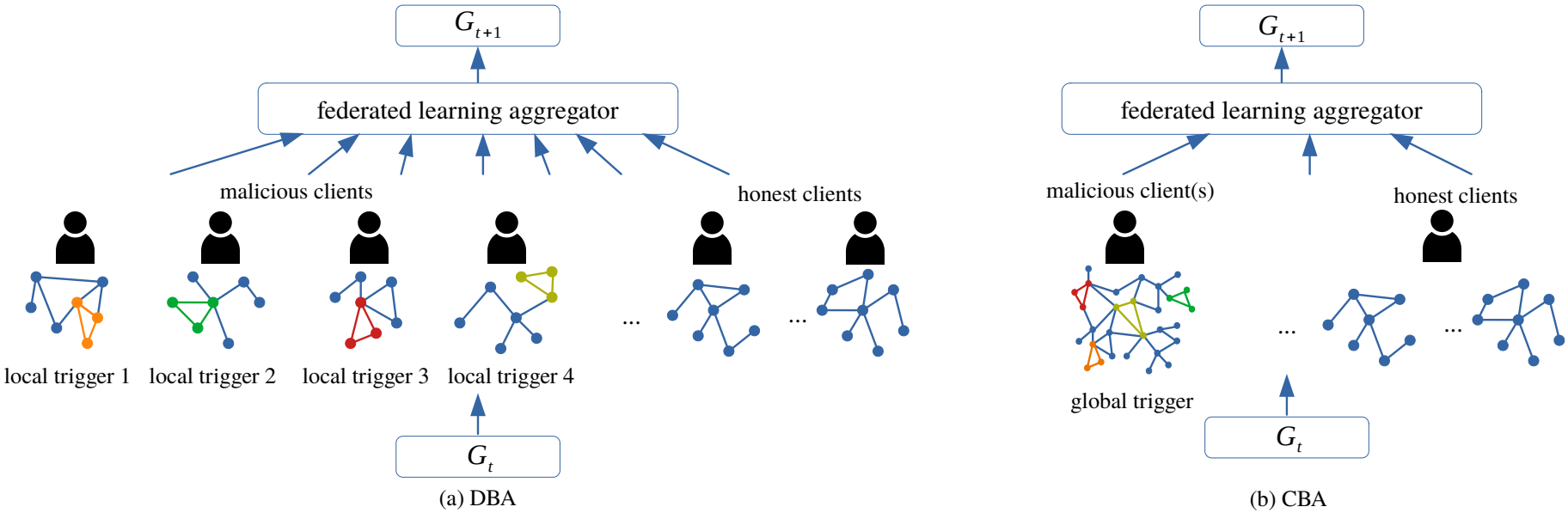



Figure 7: Attack Framework

# Outline

# Metrics

- *Clean accuracy drop (CAD)*: the classification accuracy difference between global models with and without malicious clients over the clean testing dataset.
- *Attack Success Rate (ASR)*: $\frac{\#successful\_backdoors}{\#attempts}$

# Experimental Setting

- Datasets

| Dataset | # Graphs | Avg. # nodes | Avg. # edges | Classes | Class Distribution |
|---|---|---|---|---|---|
| NCI1 | $4,110$ | 29.87 | 32.30 | 2 | $2,053[0], 2,057[1]$ |
| PROTEINS_full | $1,113$ | 39.06 | 72.82 | 2 | $663[0], 450[1]$ |
| TRIANGLES | $45,000$ | 20.85 | 32.74 | 10 | $4,500[0-9]$ |

- Models: GCN, GAT, GraphSage

# Tested Scenarios

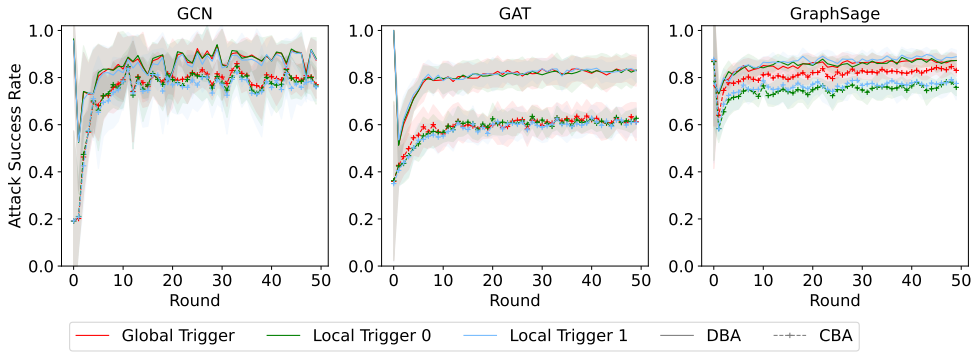
Table 1: Summary of the experimental setting ($K$: number of clients, $M$: number of malicious clients).[12]

| Experiment | Dataset | $K$ | $M$ |
|---|---|---|---|
| Exp. I | NCI1, PROTEINS_full, TRIANGLES | 5 | 2 |
| Exp. II | NCI1, PROTEINS_full, TRIANGLES | 5 | 3 |
| Exp. III | TRIANGLES | 10<br>20 | 4, 6<br>8, 12 |
| Exp. IV | TRIANGLES | 100 | 5, 10, 15, 20 |
| Exp. V | NCI1, PROTEINS_full, TRIANGLES | 5 | 2, 3 |

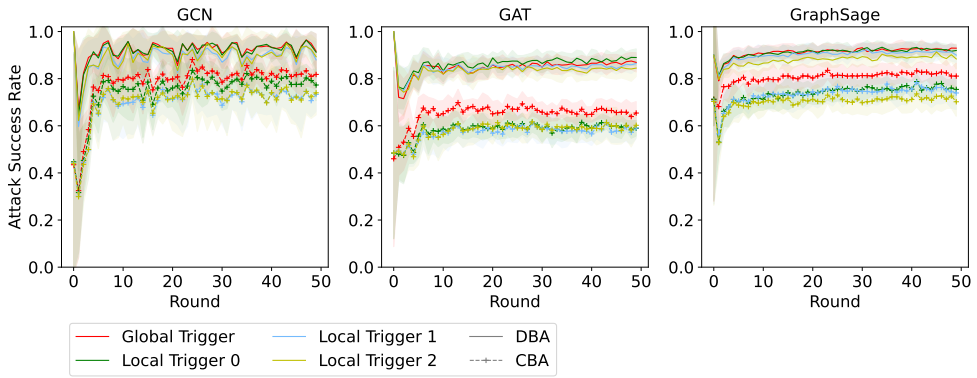
[12] Exp. I, Exp. II, Exp. III, and Exp. IV represent the experiments of honest majority attack scenario, malicious majority attack scenario, the impact of the number of clients, and the impact of percentage of malicious clients, respectively.

# Results

## Exp. I, II: experiments of honest majority & malicious majority attack scenarios




(a) Honest majority attack scenario



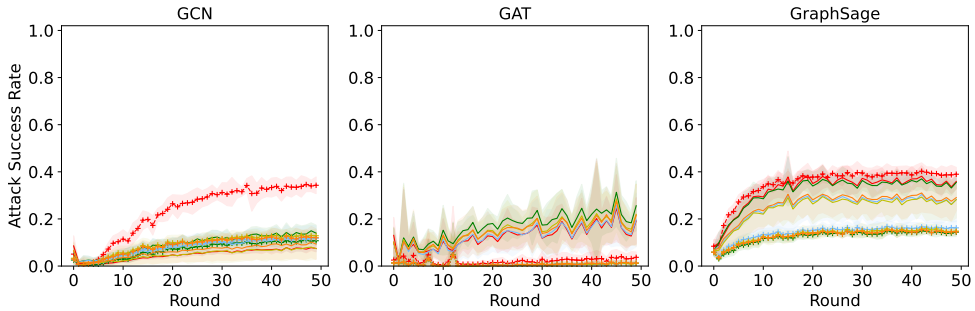

(b) Malicious majority attack scenario

Figure 8: Backdoor attack results on NCI1 with 5 clients.

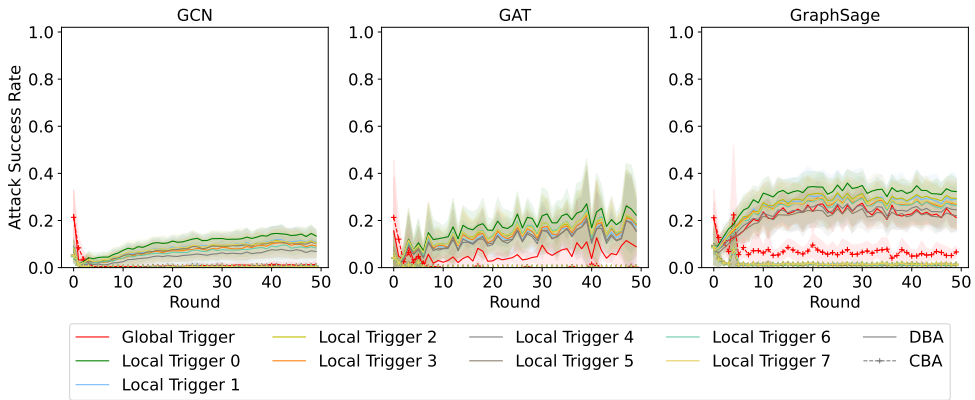
- For CBA, surprisingly, the ASR of all local triggers can be as high as the global trigger even if the centralized attacker embeds only the global trigger into the model
- In most cases, the ASR of DBA and CBA increases with more malicious clients. And the increase in DBA is more significant than in CBA.

# Results

## Exp. III: experiment to explore the impact of the number of clients




(a) 10 clients




(b) 20 clients

- With more clients, the attack success rate of CBA decreases dramatically while the attack performance of DBA keeps steady.

Figure 9: Backdoor attack results on TRIANGLES with more clients (honest majority scenario).

# Results

## Exp. IV: experiment to explore the impact of percentage of malicious clients

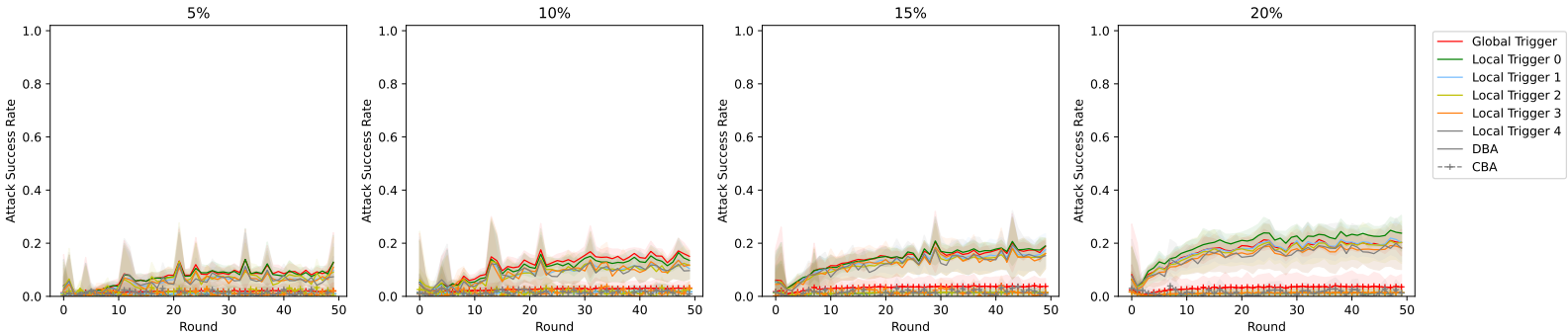
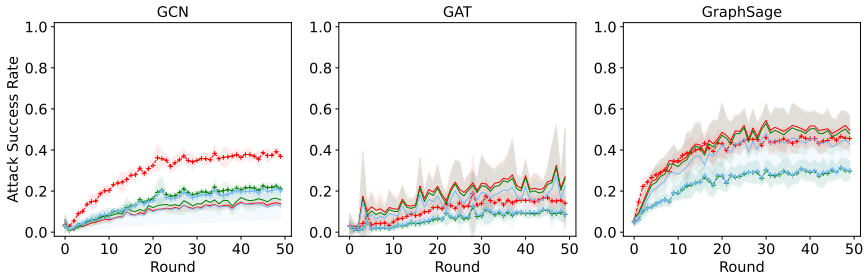



Figure 10: Backdoor attack results of TRIANGLES with less percentage of malicious clients ($K = 100$, GraphSage).

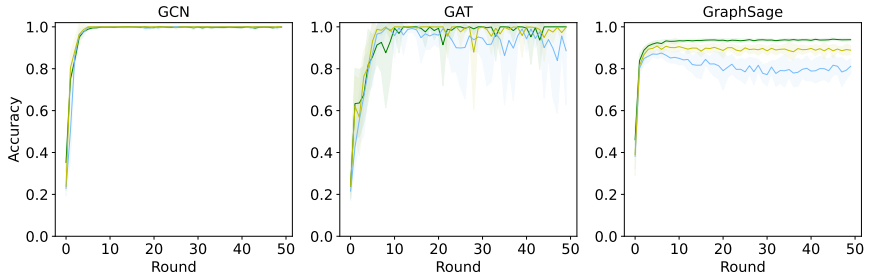
- The increase in $M$ has a more positive impact on DBA than CBA.

# Results

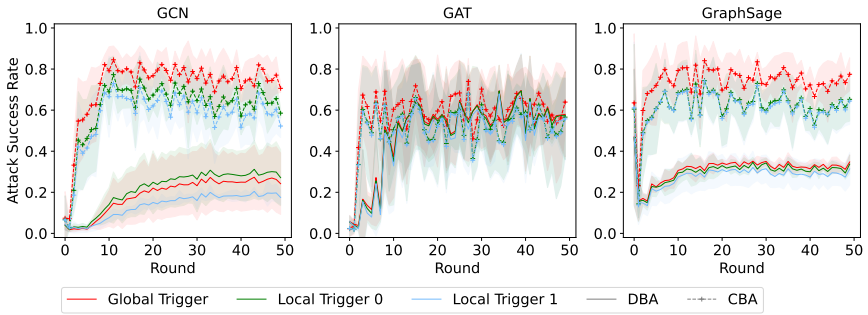
## Exp. V: experiment of backdoor attacks against potential defenses



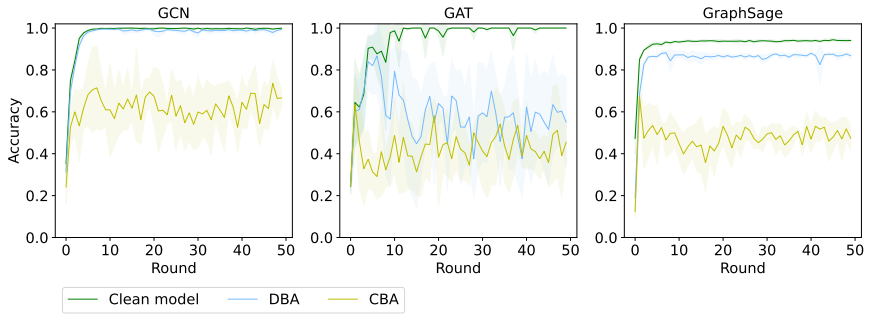

(a) FLAME




(b) FLAME




(c) FoolsGold




(d) FoolsGold

Figure 11: Backdoor attack results of TRIANGLES on two defenses (left: ASR, right: testing accuracy).

- Both defenses cannot detect malicious updates successfully.

# Outline

# Conclusions & Future Work

- Conclusions:
  - Backdoor attacks (both DBA and CBA) are a practical threat for the federated GNNs under the cross-silo threat model.
  - Current defenses are not effective.
- Future work:
  - Explore backdoor attacks in Federated GNNs for the node classification task.
  - Propose a defense specifically for the backdoor attacks in Federated GNNs.

*The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.*

Contact:

Jing Xu j.xu-8@tudelft.nl
Rui Wang r.wang-8@tudelft.nl
Stefanos Koffas s.koffas@tudelft.nl
Kaitai Liang kaitai.liang@tudelft.nl
Stjepan Picek s.picek@tudelft.nl