ETH zürich

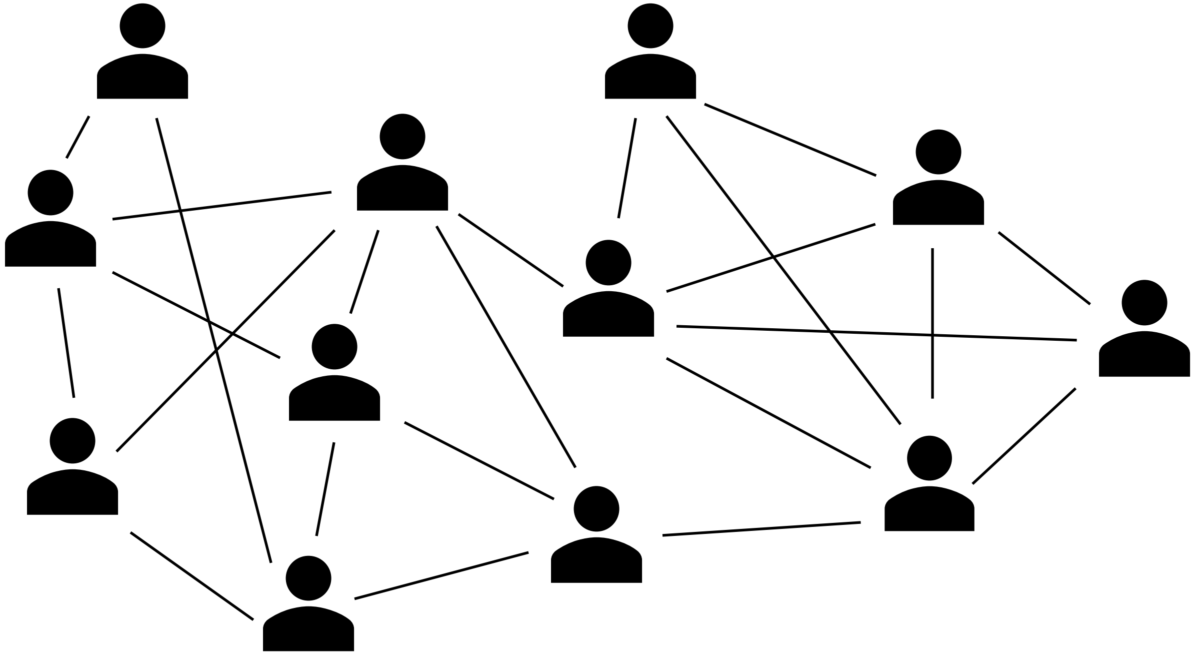
Distributed Computing Group (D-ITET)

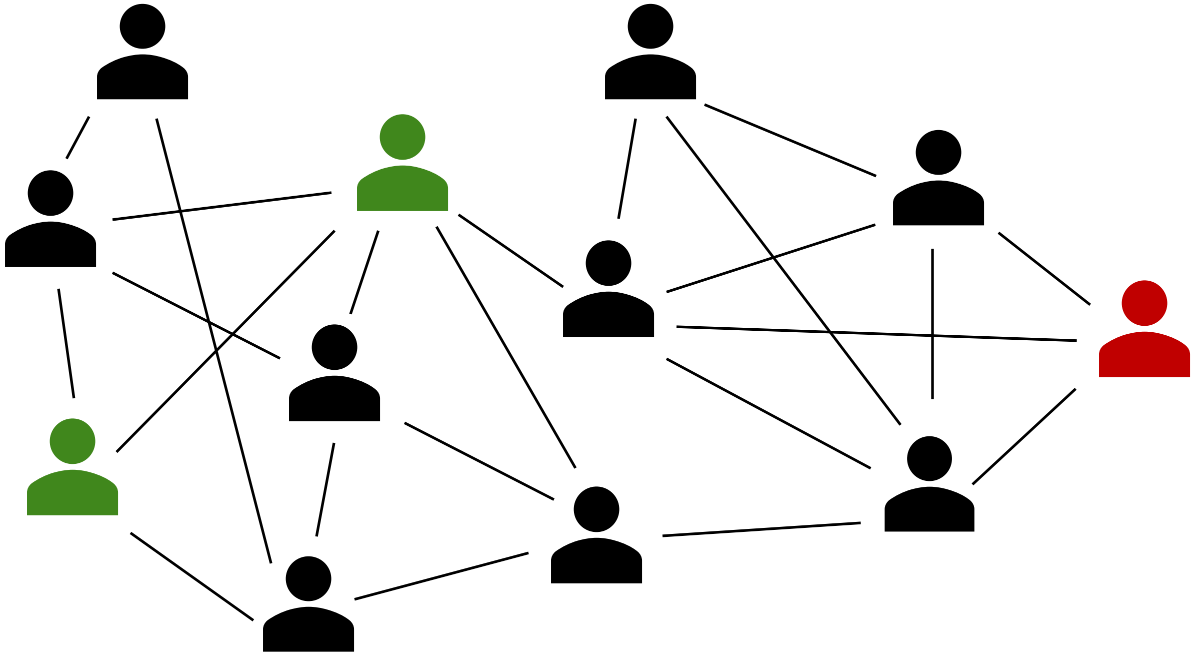
Sybil Detection using GNNs

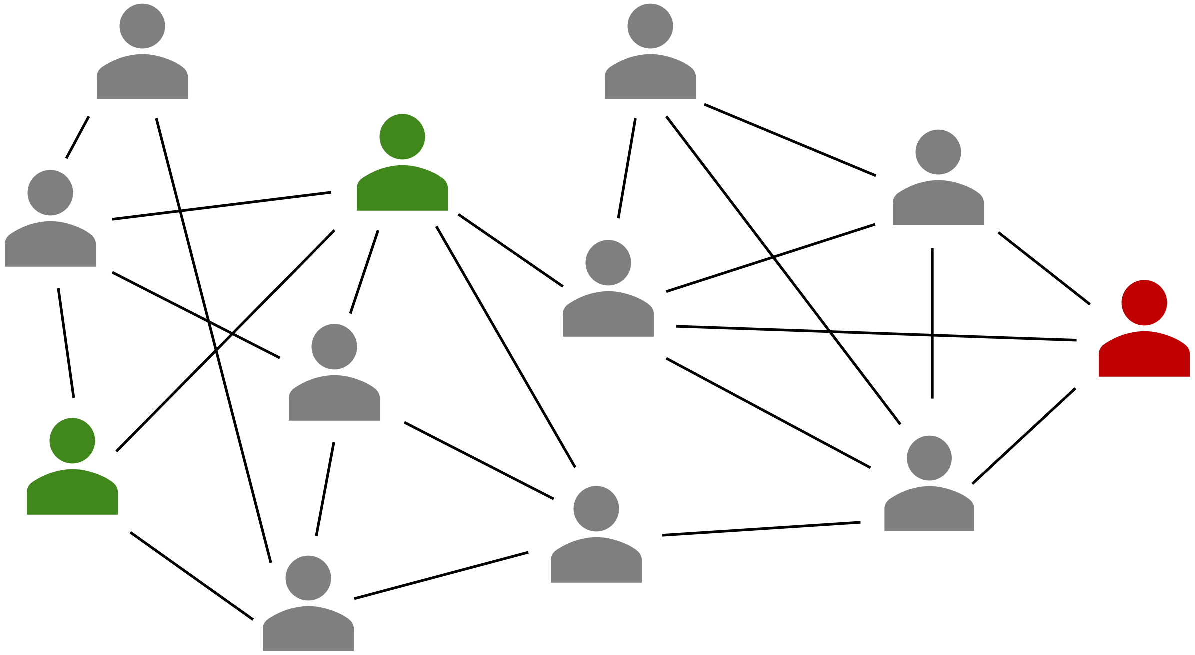
Master Thesis Presentation

Stuart Heeb

September 12th, 2024

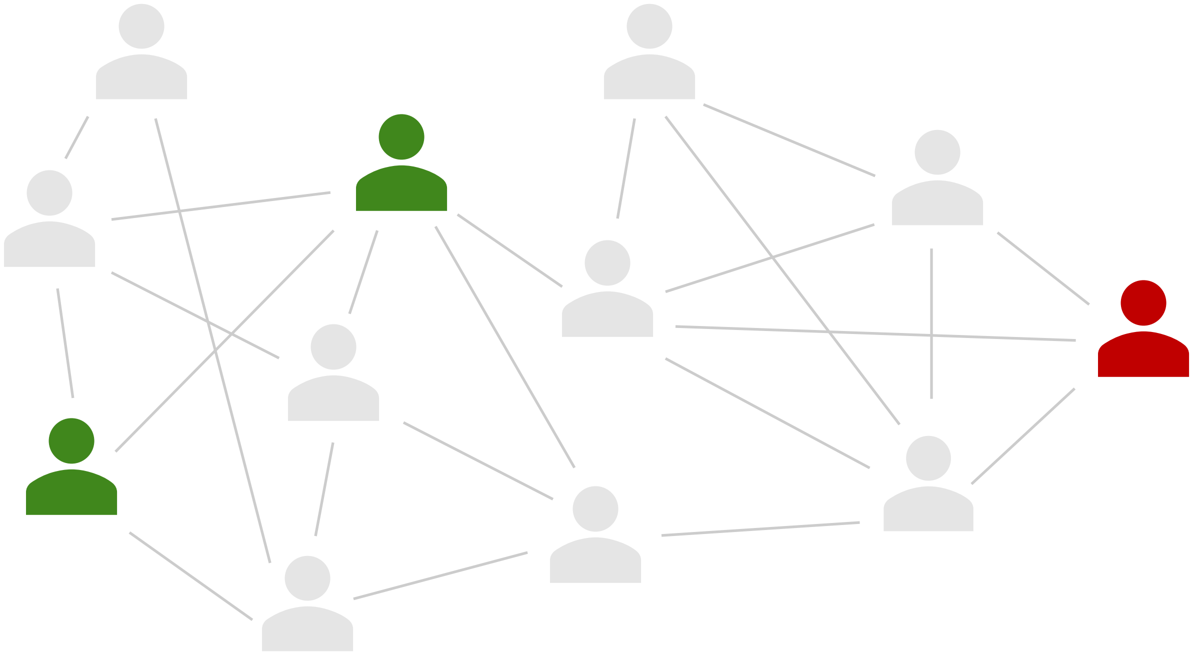




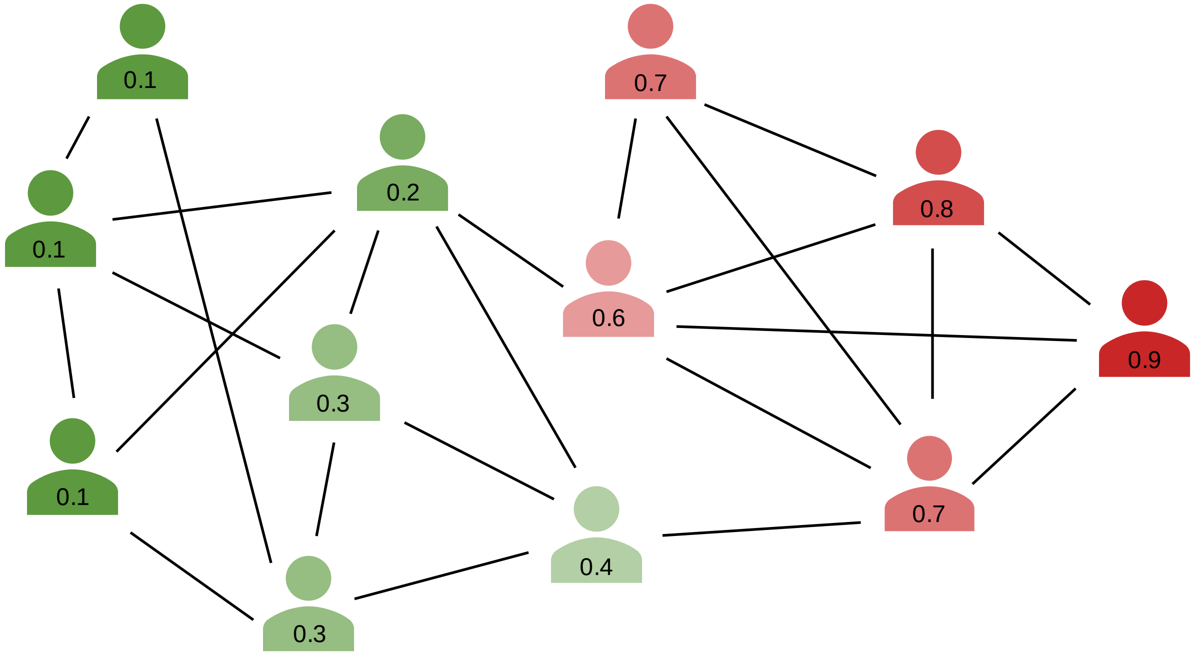


Known honest nodes

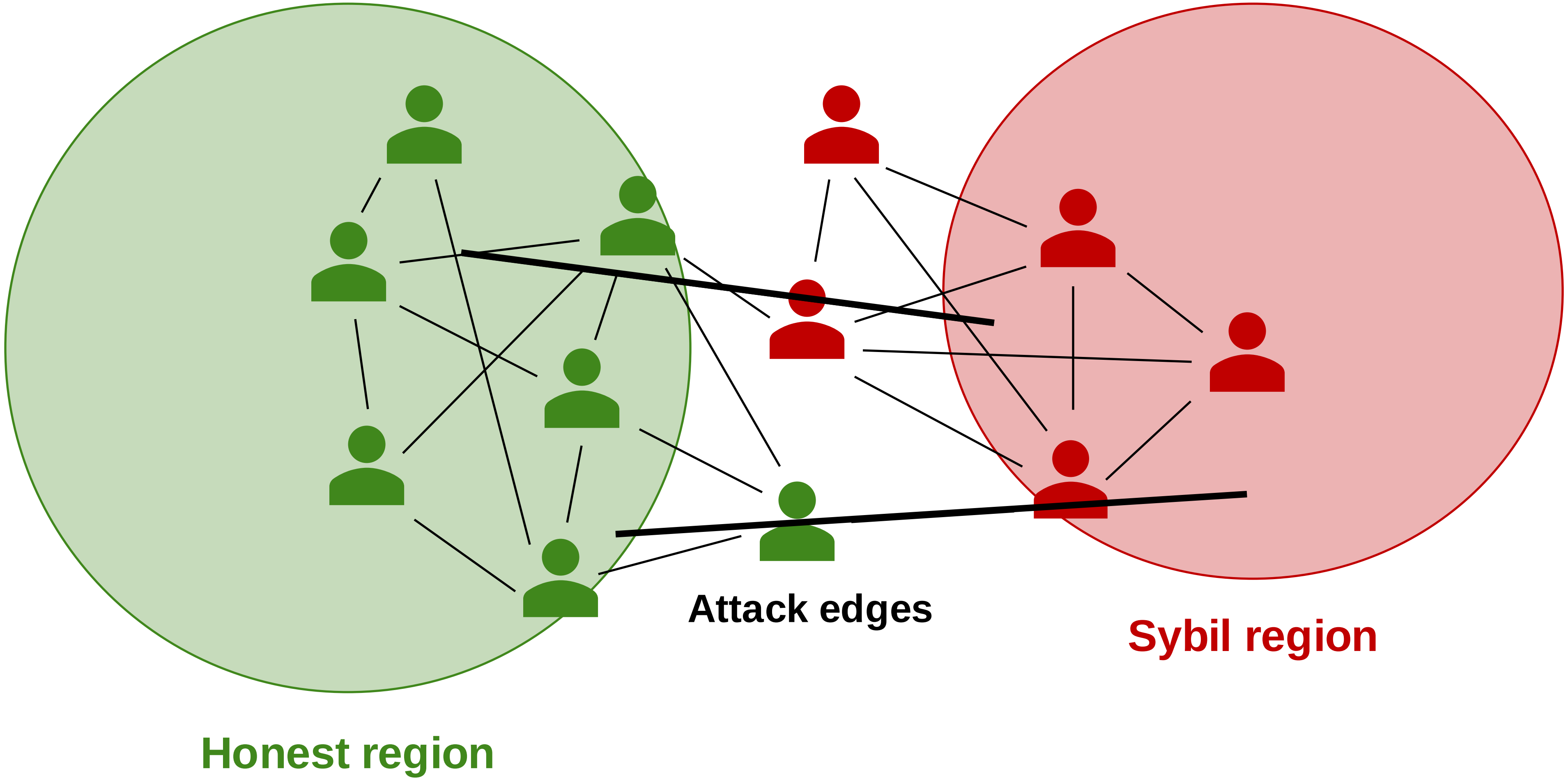
Known sybil nodes

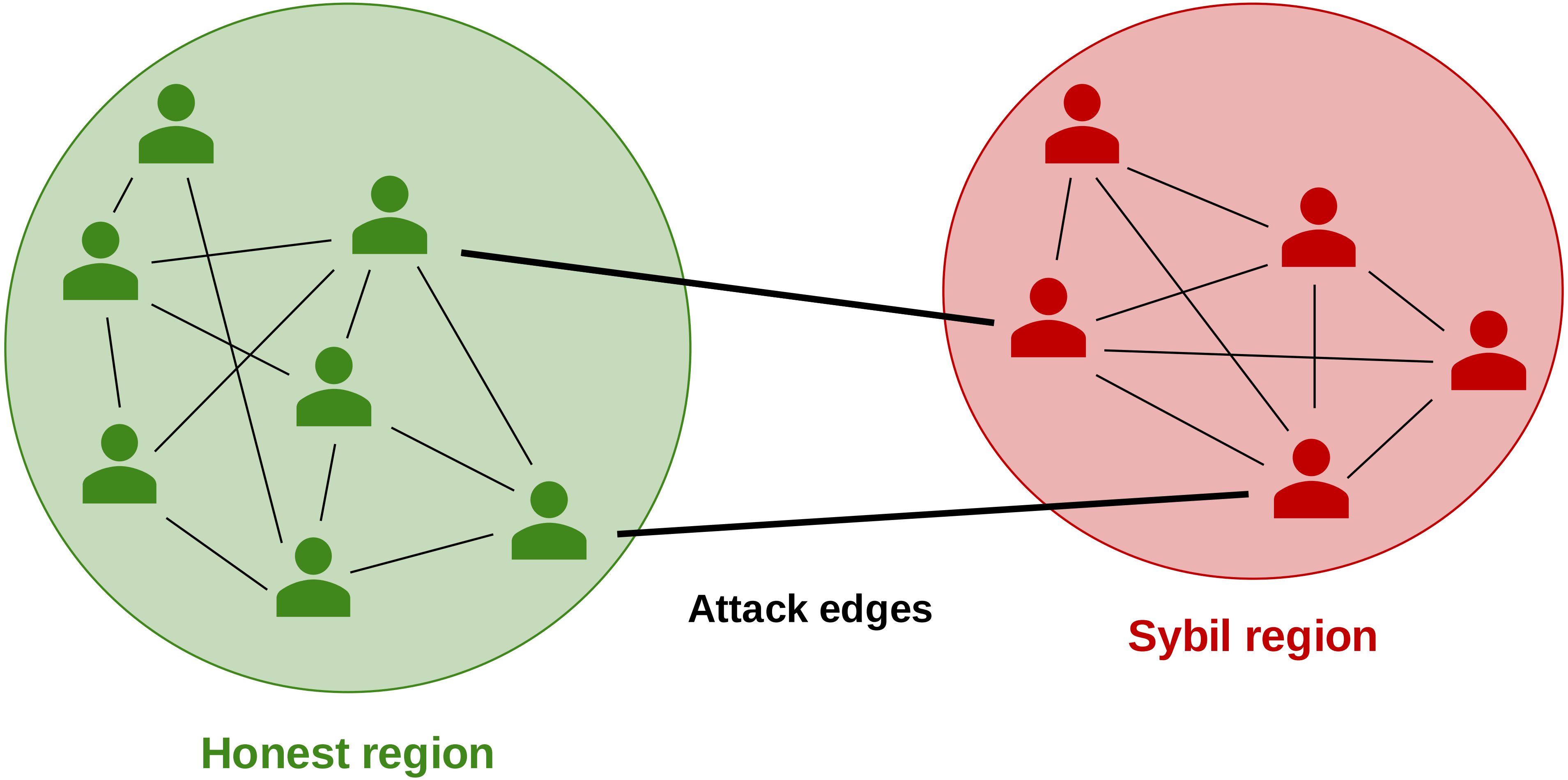


… Sybil Detection Algorithm



$$P(x_u = 1 \mid \mathbf{x}_V)$$


Attack edges
Sybil region
Honest region

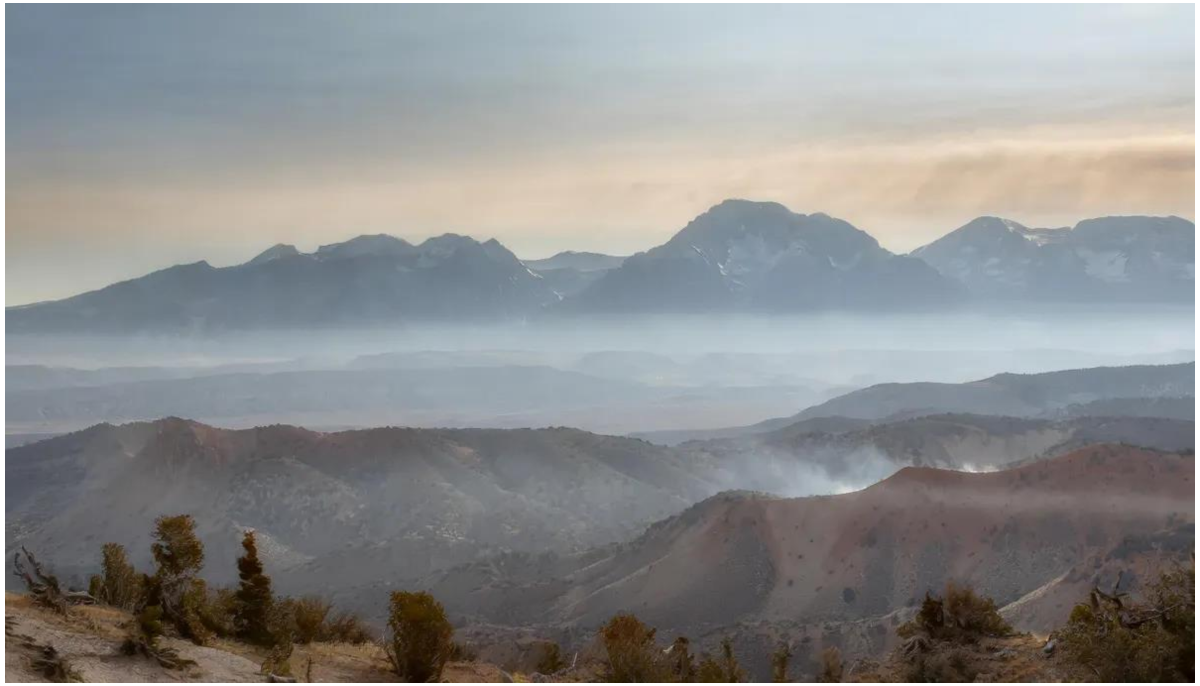

Attack edges
Sybil region
Honest region

Why are Sybils a Problem?

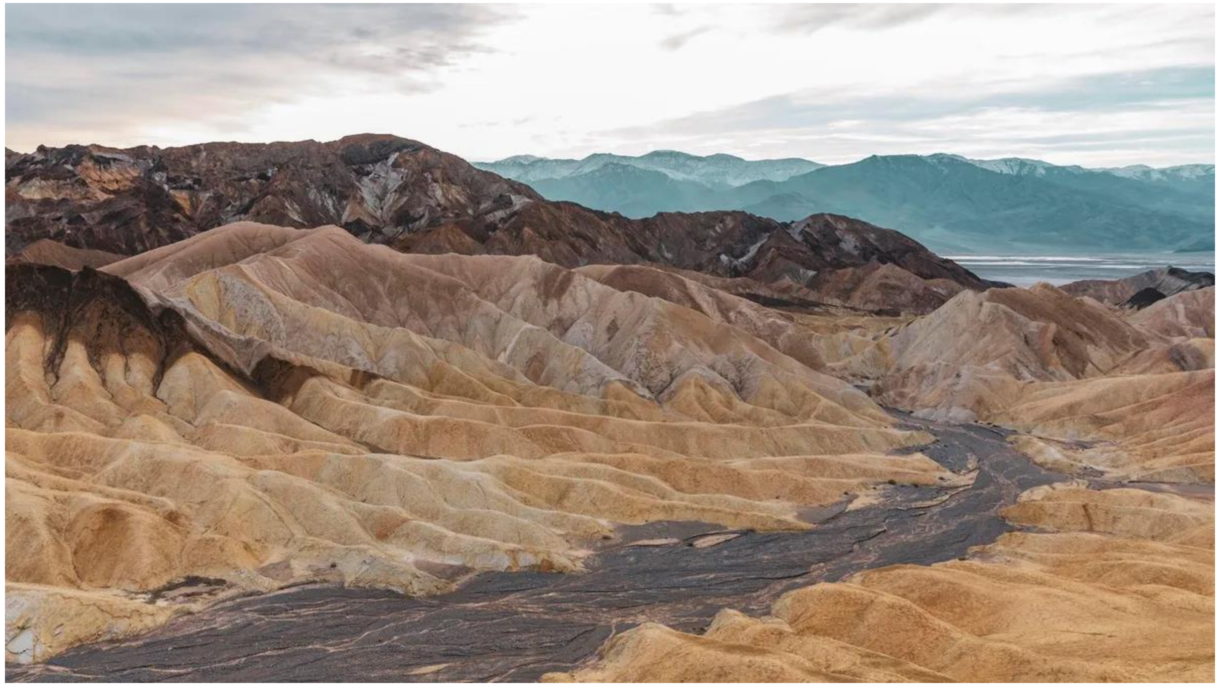
- Spread misinformation
- Degrade trust
- Influence public opinion

https://www.cnet.com/pictures/ai-or-not-ai-can-you-spot-the-real-photos/

Modern Sybils



AI-generated



Actual photograph

Which picture is real?

As in, an image of the real world taken by a physical camera?

Structure-based Sybil Detection

- Content becoming harder to (even manually) distinguish
- Other advantages
  - Privacy concerns
  - Enhanced generalizability

Background & Related Work

SybilRank

2012

- Random Walks (RW)
- Uses only honest labels
- Computationally efficient

SybilBelief

2014

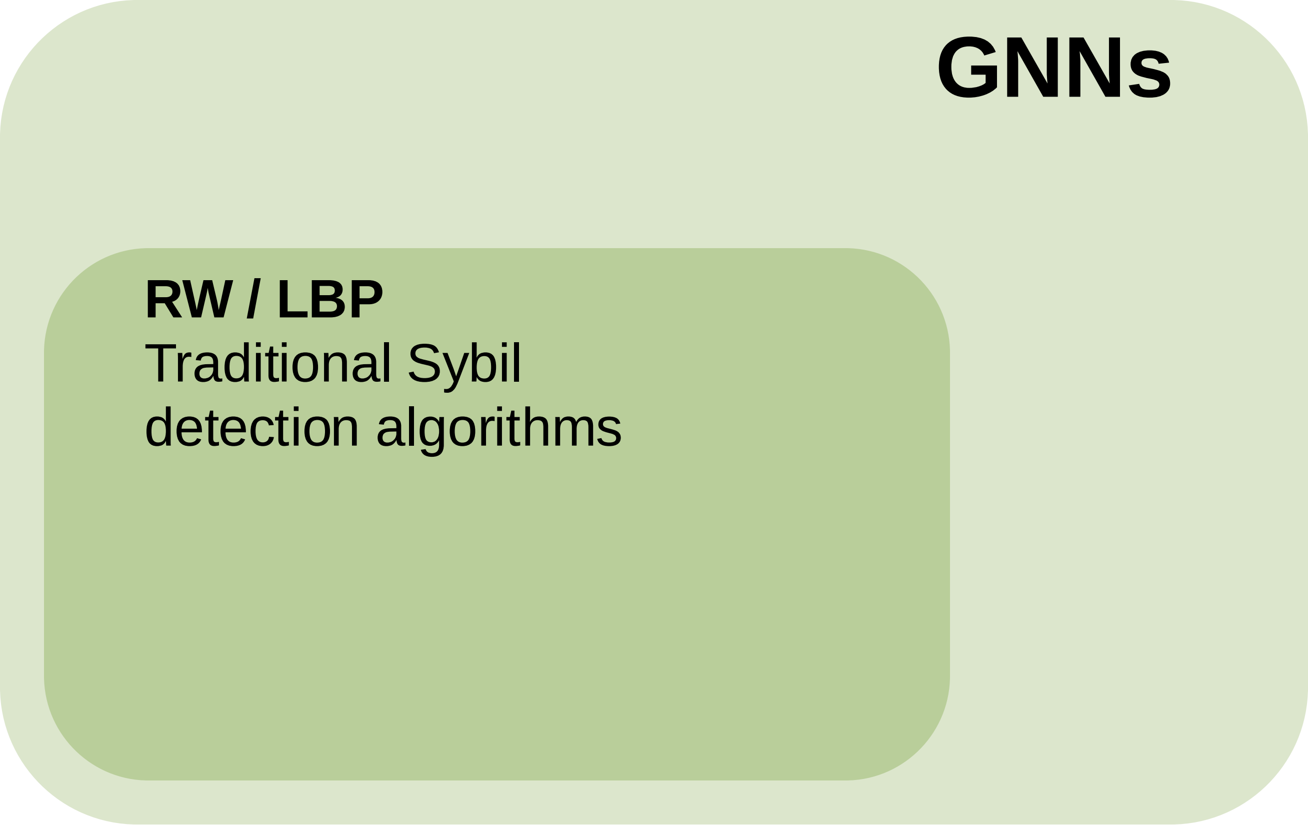
- Loopy Belief Propagation (LBP)
- Uses both honest and sybil labels

SybilSCAR

2019

- Local rule-based propagation
- Uses both honest and sybil labels
- Combines benefits of RW and LBP

Why use GNNs?

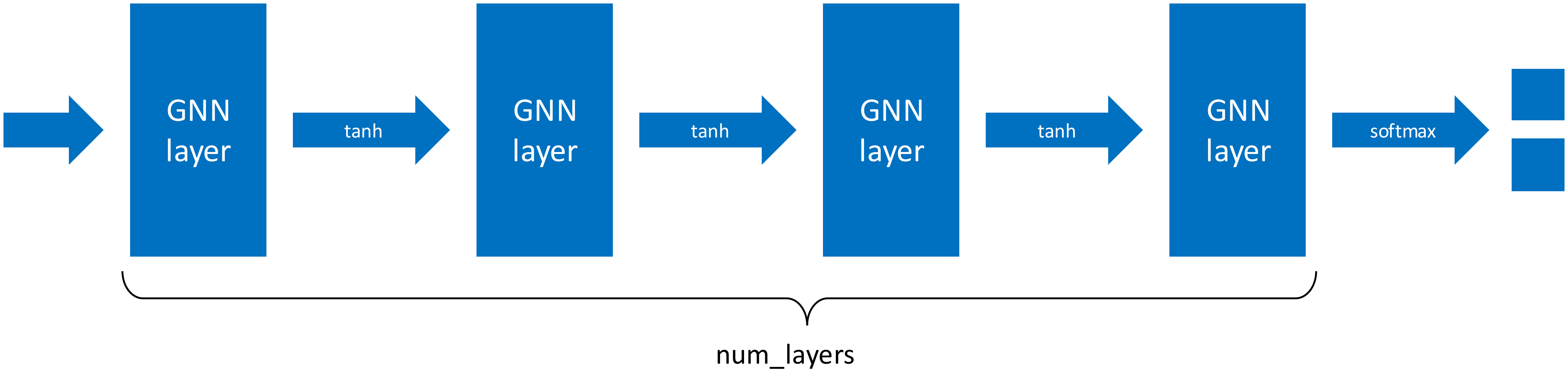


GNNs

RW / LBP
Traditional Sybil detection algorithms

GNNs as a «one size fits all»?

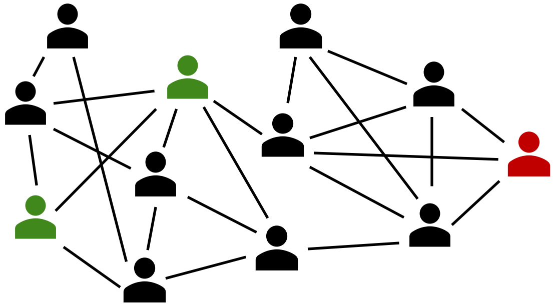
GNNs for Sybil Detection



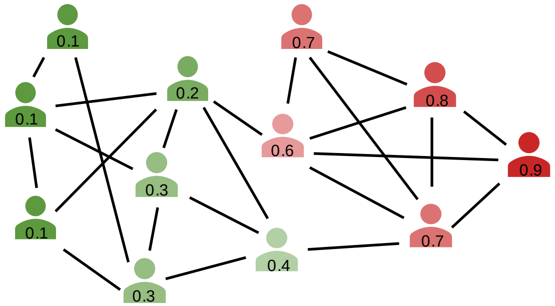
SybilGCN SybilRGCN SybilGAT

GNNs for Sybil Detection

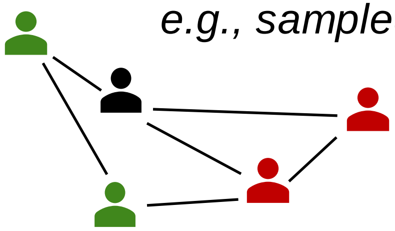
Approaches



Train and **directly** predict



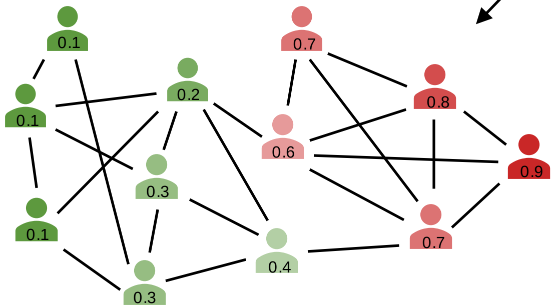
Without pre-training



e.g., sampled subgraph

Train, **then** predict

could be multiple graphs



With pre-training

Data

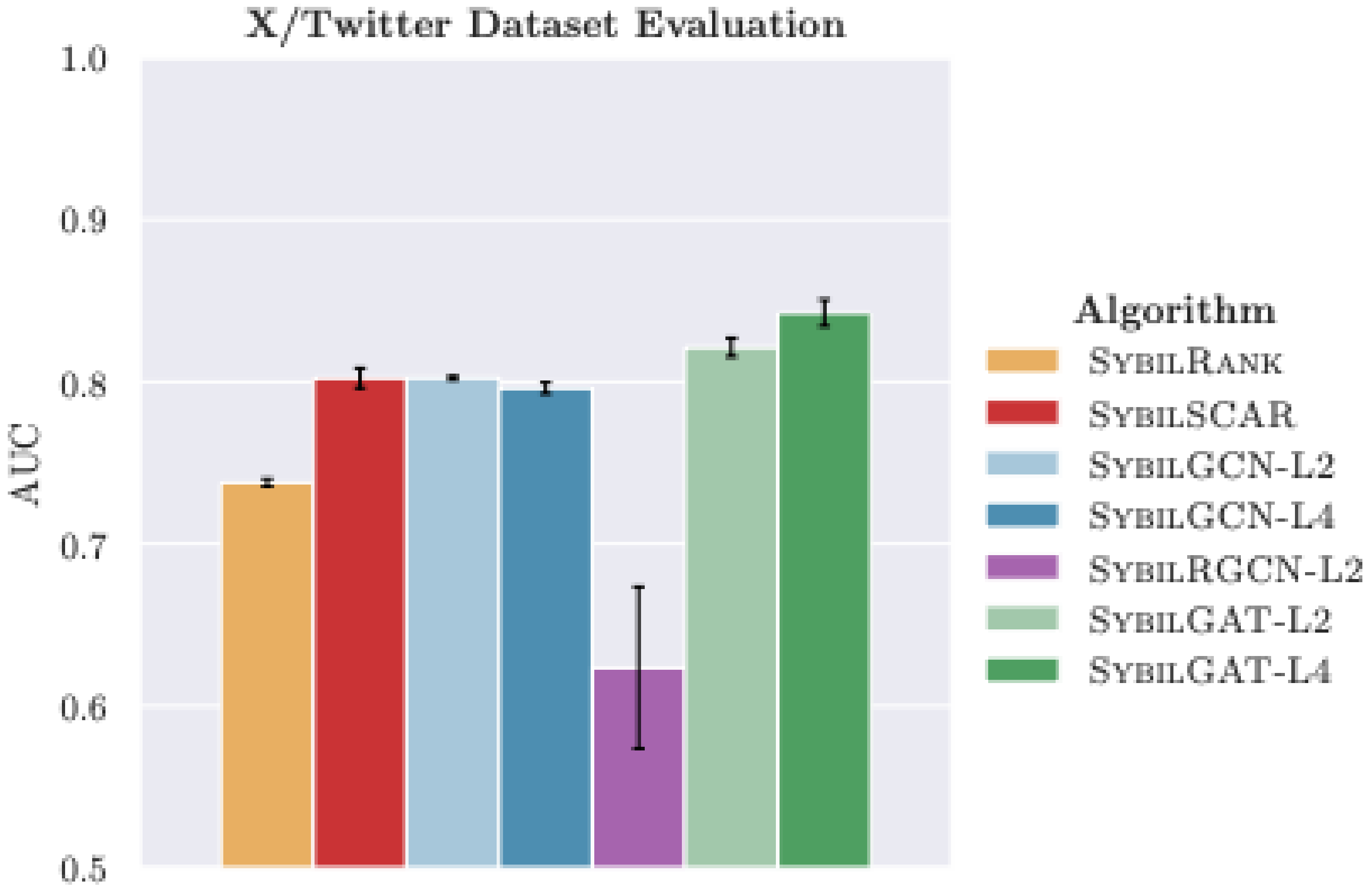
270K nodes
6.8M edges
labeled

4K nodes
88K edges
unlabeled

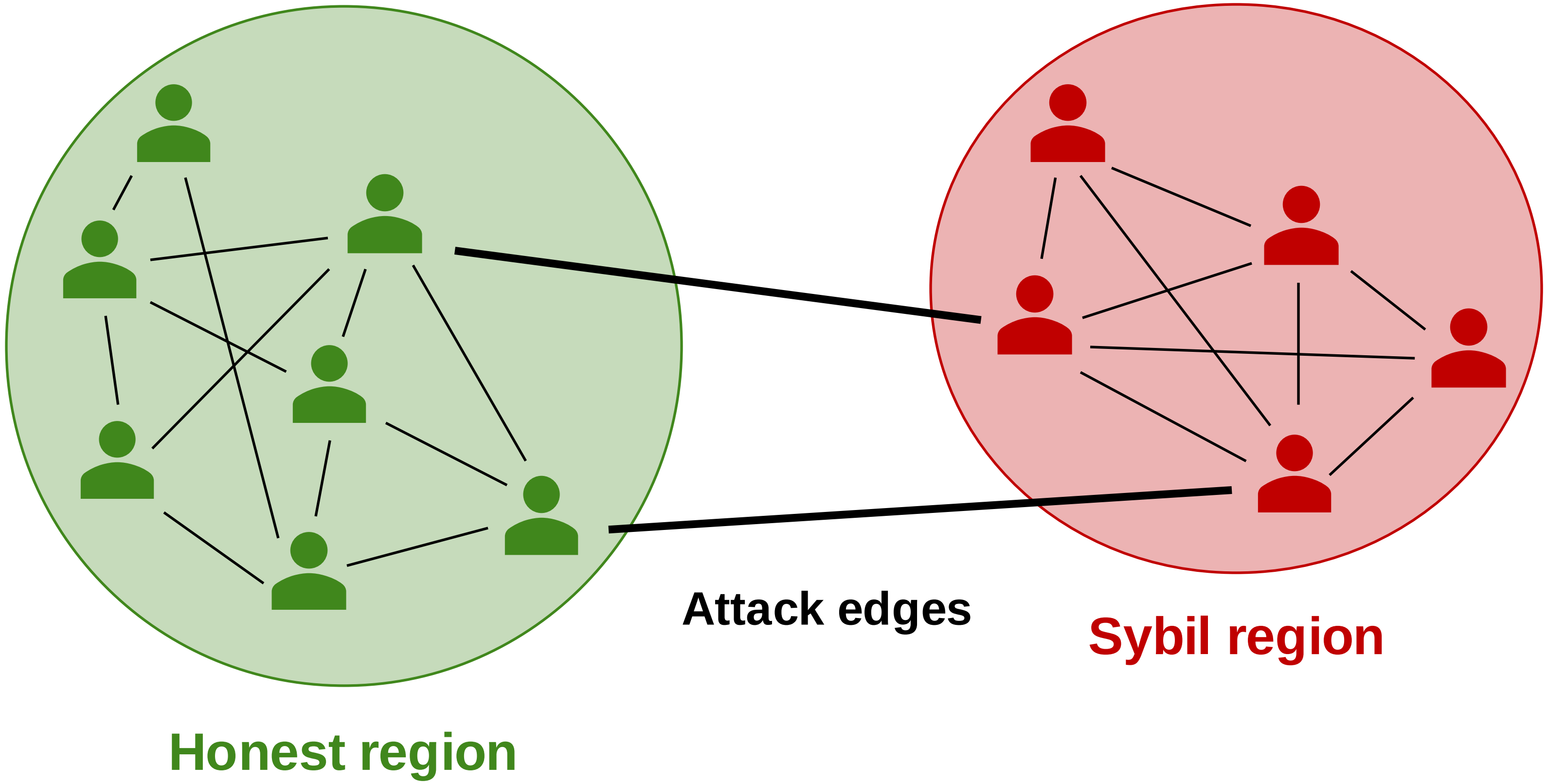
Synthetic
networks

https://www.vecteezy.com/free-vector/facebook-logo-black
https://www.freepik.com/free-photos-vectors/x-logo

Results: Performance Comparison



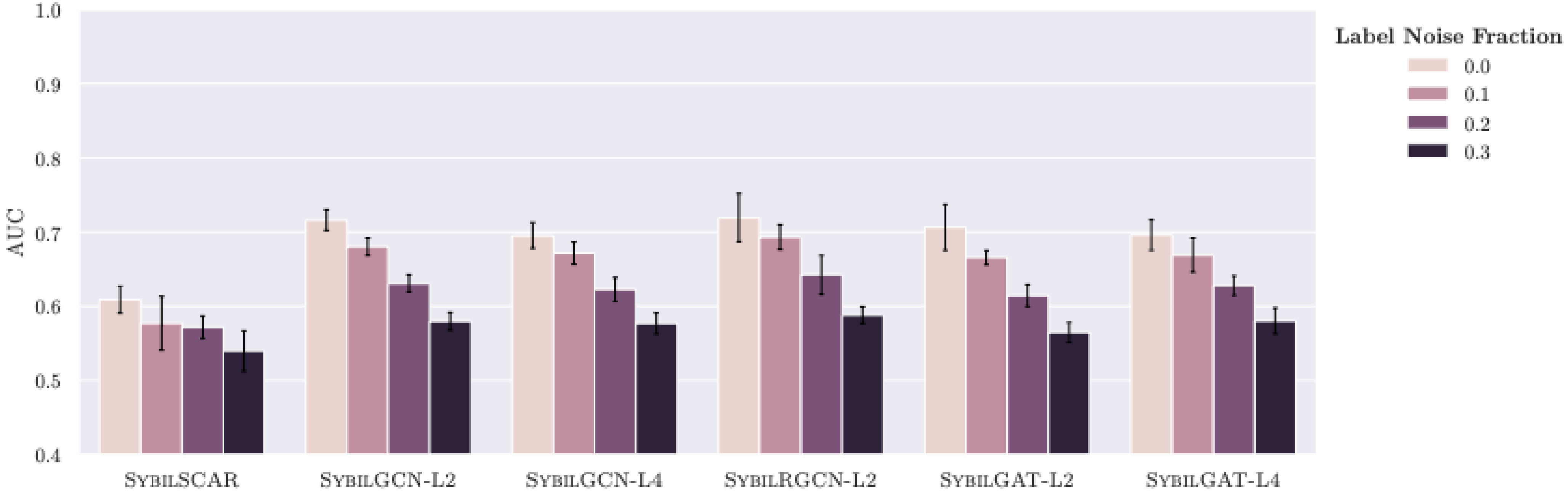
Results: Social Networks Synthetization



How to synthesize attack edges?

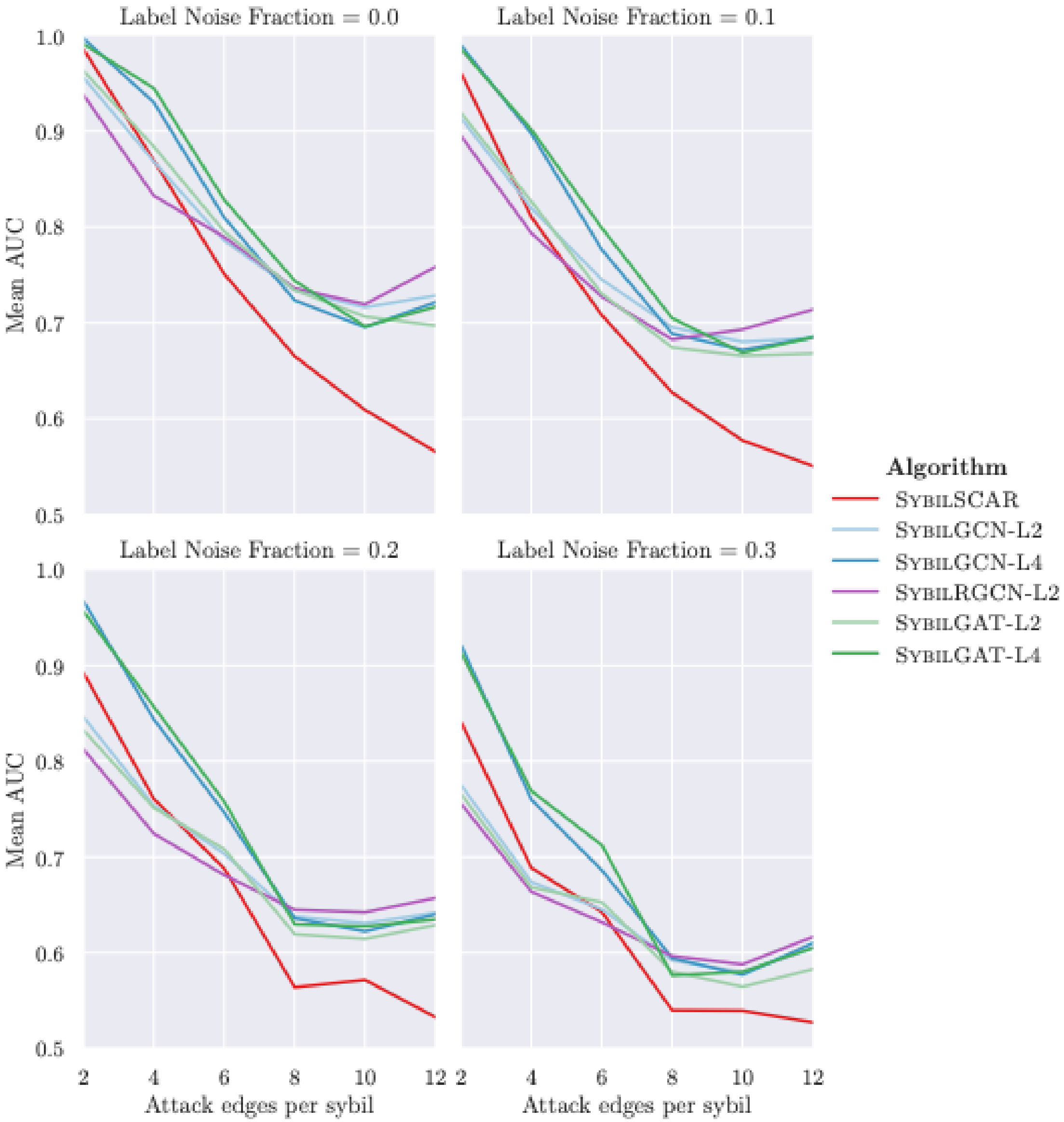
Results: Robustness

Label Noise Level

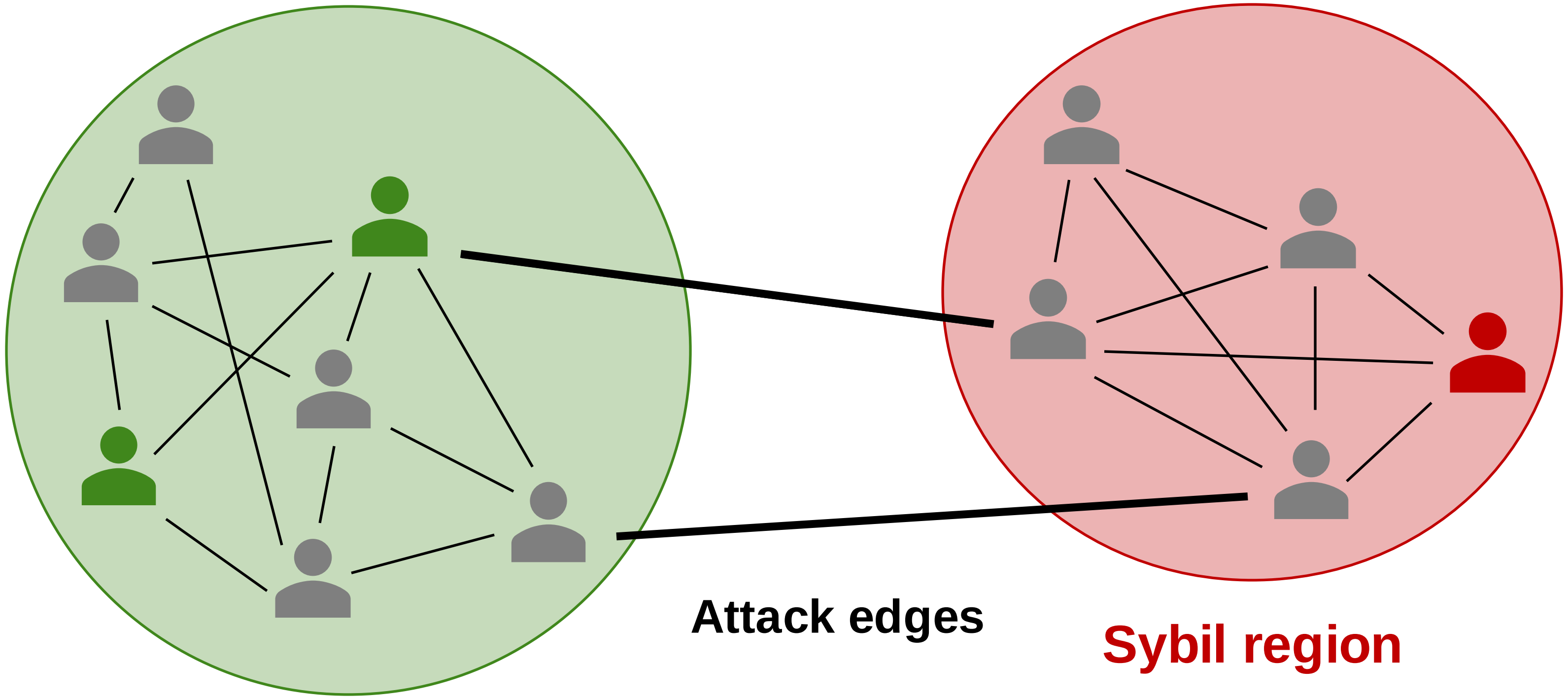


Results: Robustness

Label Noise Level



Adversarial Attack



Attack edges

Sybil region

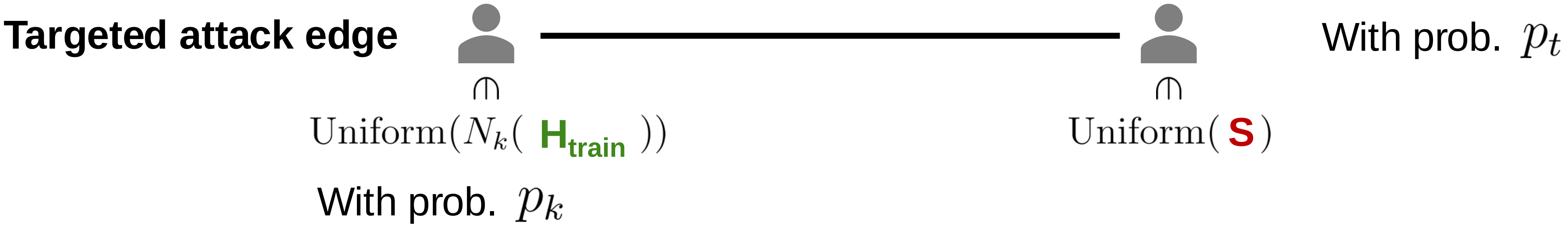
Honest region

How to place attack edges in a targeted fashion?

$$p_t \qquad \boldsymbol{p}$$

ETH zürich

Adversarial Attack



Targeted attack edge

With prob. p_t

$\in$ $\mathrm{Uniform}(N_k($ **H_{train}** $))$

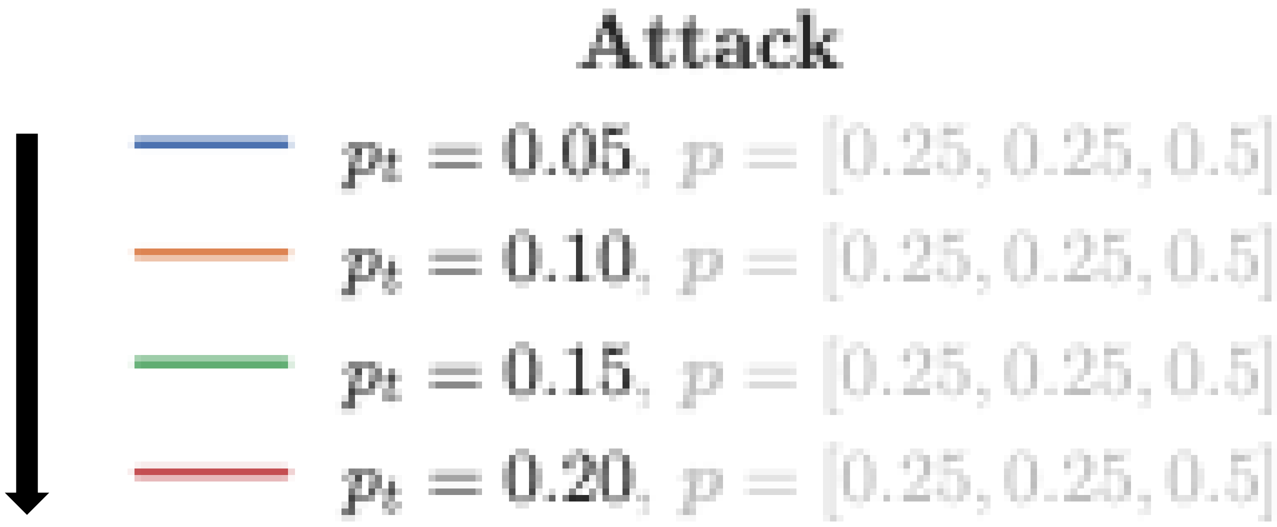
With prob. p_k

$\in$ $\mathrm{Uniform}($ **S** $)$

$$\boldsymbol{p} = [0.25, 0.25, 0.5] \in [0,1]^K$$

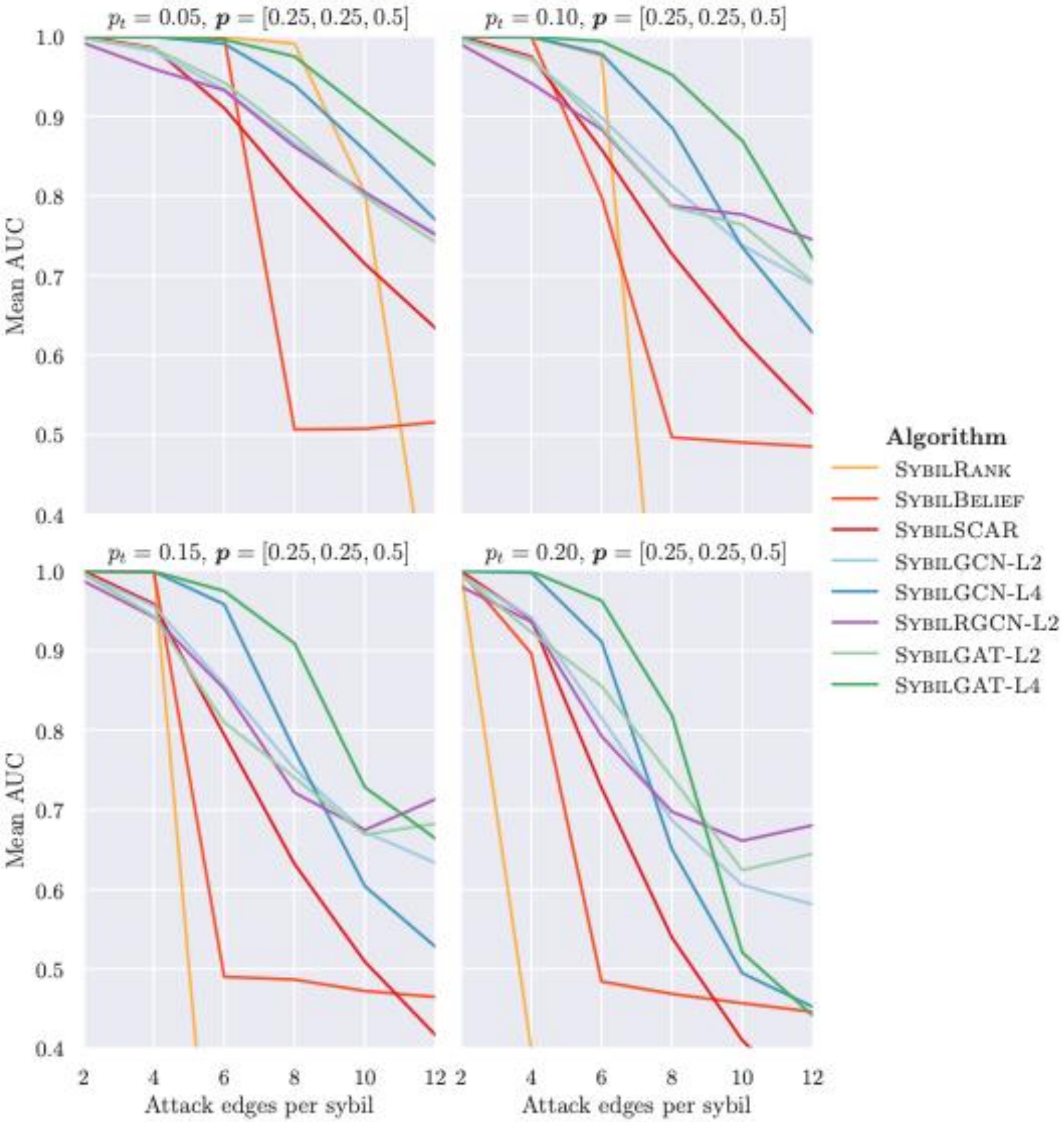
direct hit neighbor 2-hop neighbor

Results: Adversarial Attacks


Attack
$p_t = 0.05$, $p = [0.25, 0.25, 0.5]$
$p_t = 0.10$, $p = [0.25, 0.25, 0.5]$
$p_t = 0.15$, $p = [0.25, 0.25, 0.5]$
$p_t = 0.20$, $p = [0.25, 0.25, 0.5]$

Increasing expected number of targeted attack edges

Results: Adversarial Attacks



Conclusion

- $\textsc{SybilGCN}$ and $\textsc{SybilGAT}$ algorithms outperform the baselines in almost all scenarios, including real-world X/Twitter dataset
- $\textsc{SybilRGCN}$ excels when attack complexity is high
- Synthesized social networks

Limitations & Future Work

- More advanced GNN architecture somewhat unexplored
- Robustness to lack of training data and label noise
- More data
- Theoretical analysis

ETH zürich

Thank you for listening!

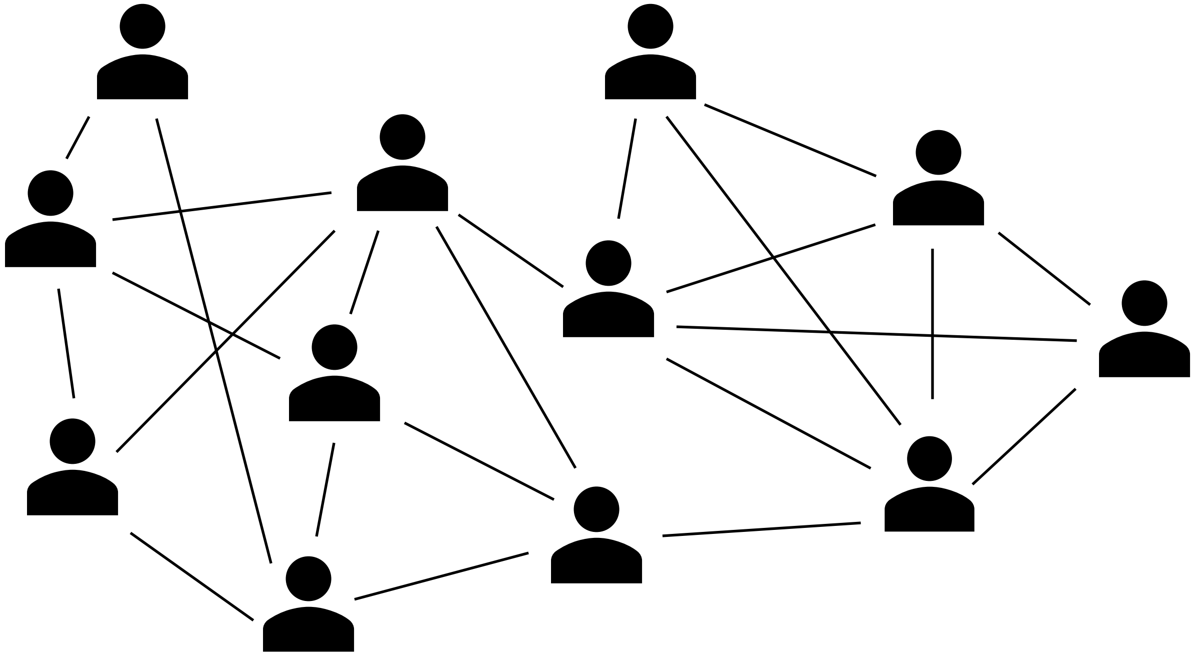
DITET DINFK

Stuart Heeb

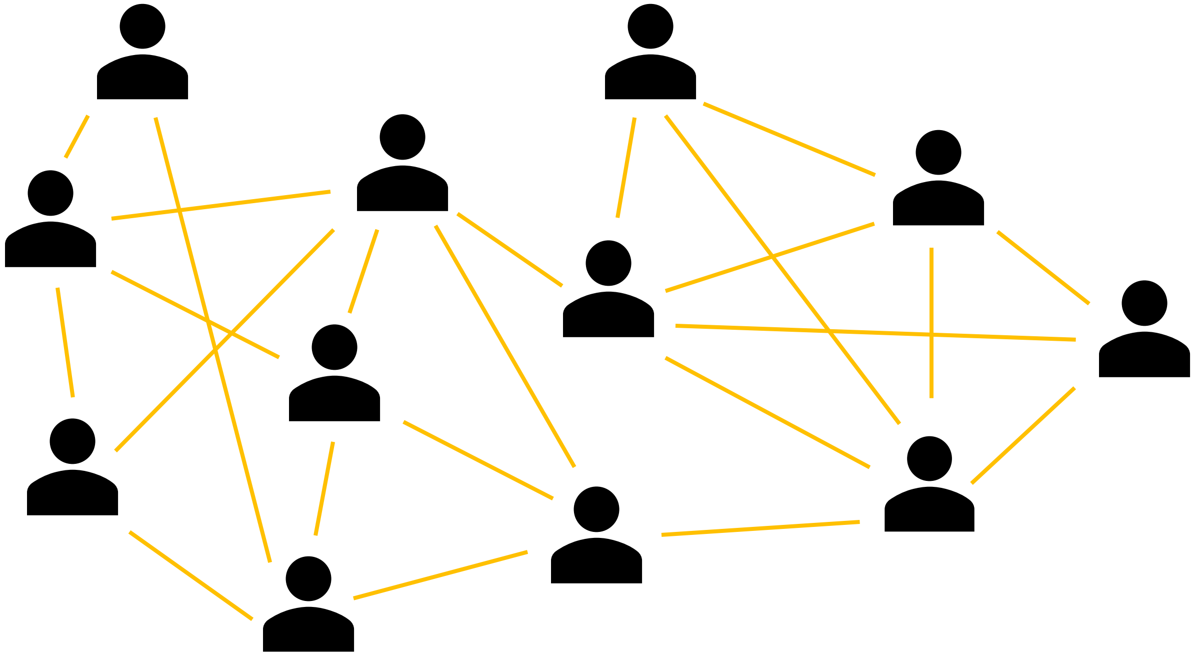
stuart.heeb@inf.ethz.ch

Backup Slides

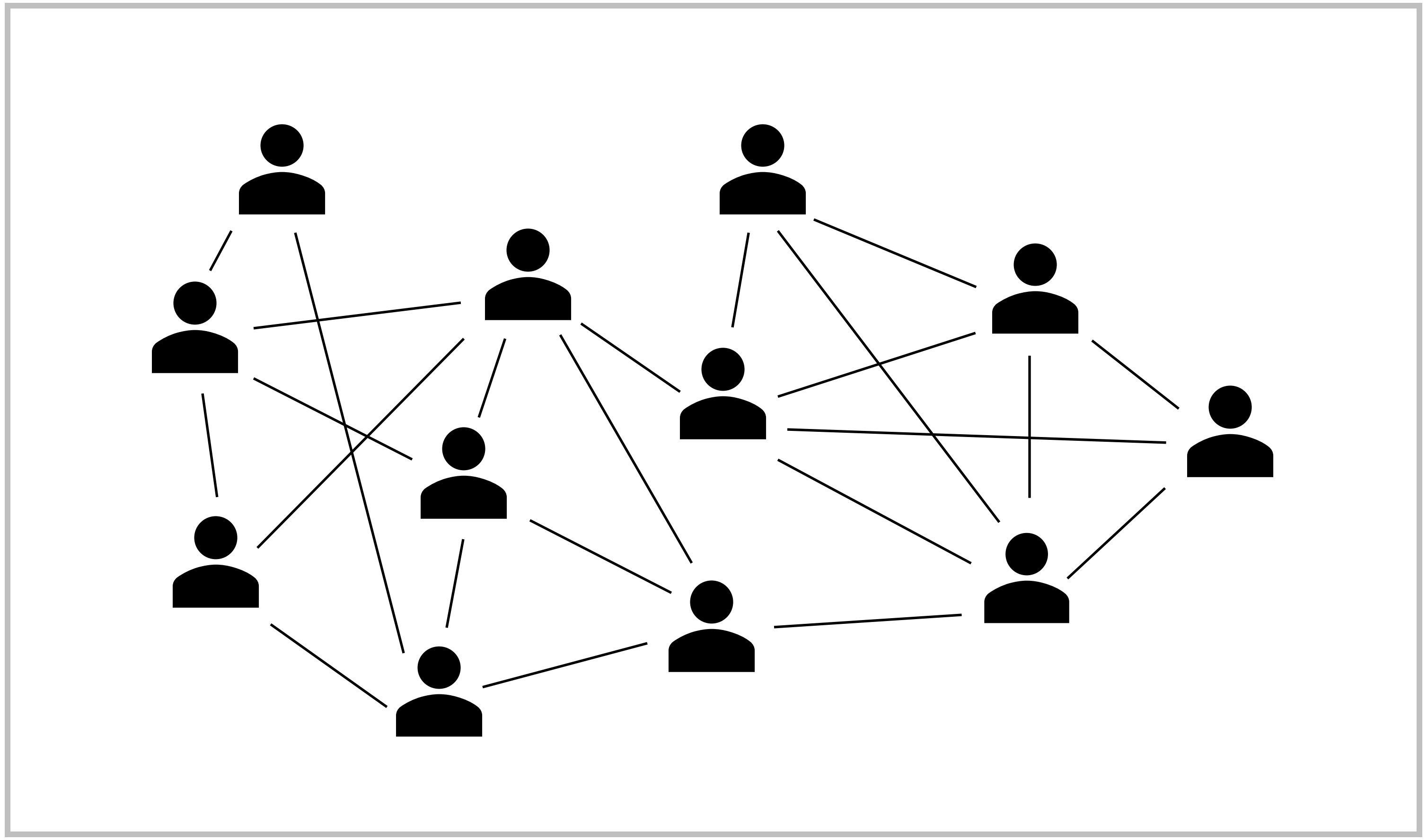
trust



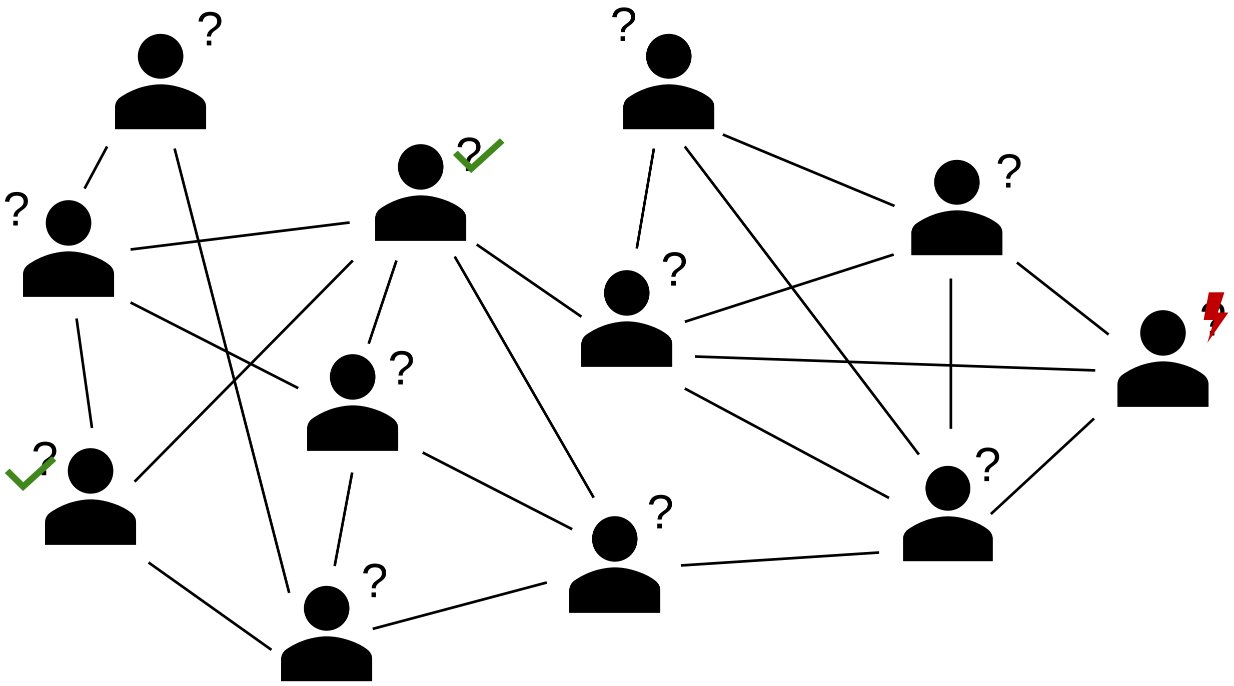
trust

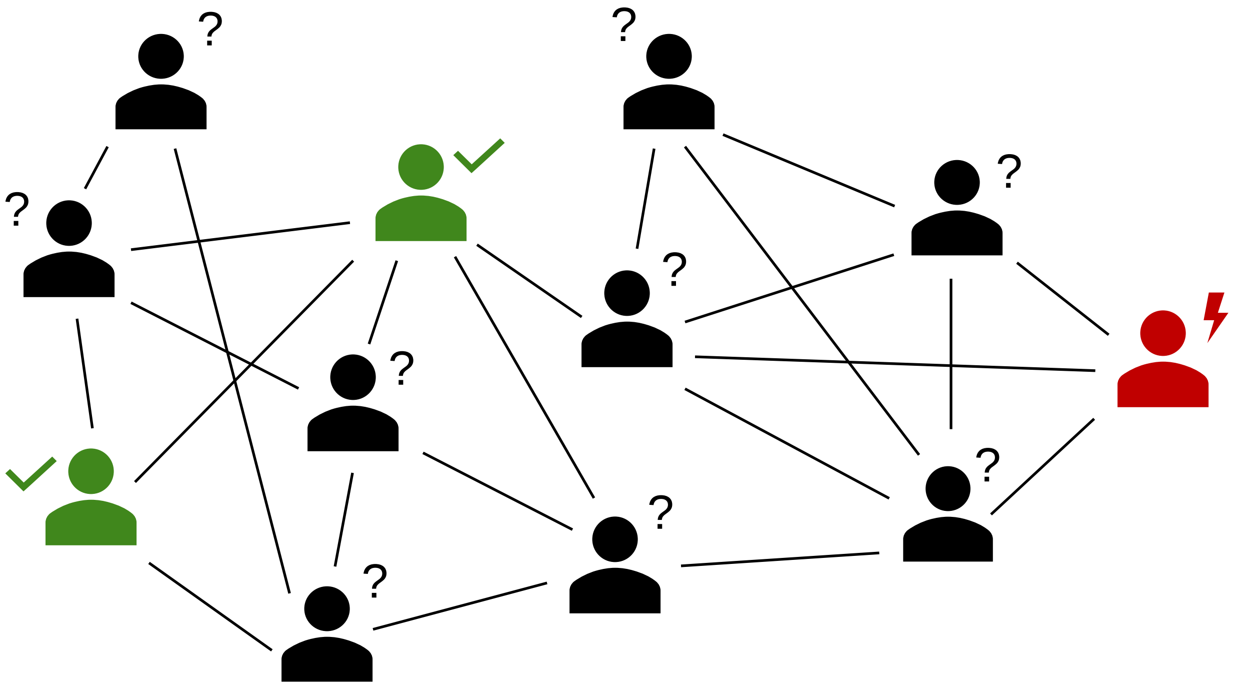


Online Social Network (OSN)



Users on a social media platform





Background & Related Work

	SybilRank (2012)	SybilBelief (2014)	SybilSCAR (2019)
Approach	Random Walks (RW)	Loopy Belief Propagation (LBP)	Local rule-based propagation
Labels (honest / Sybil)	Only honest	Both	Both
Guaranteed convergence	Yes	No	Yes
Computational complexity	O(n log n)	O(n) per iteration	O(n log n)
Main advantage	Computationally efficient	Can use both label types	Combines benefits of RW and LBP

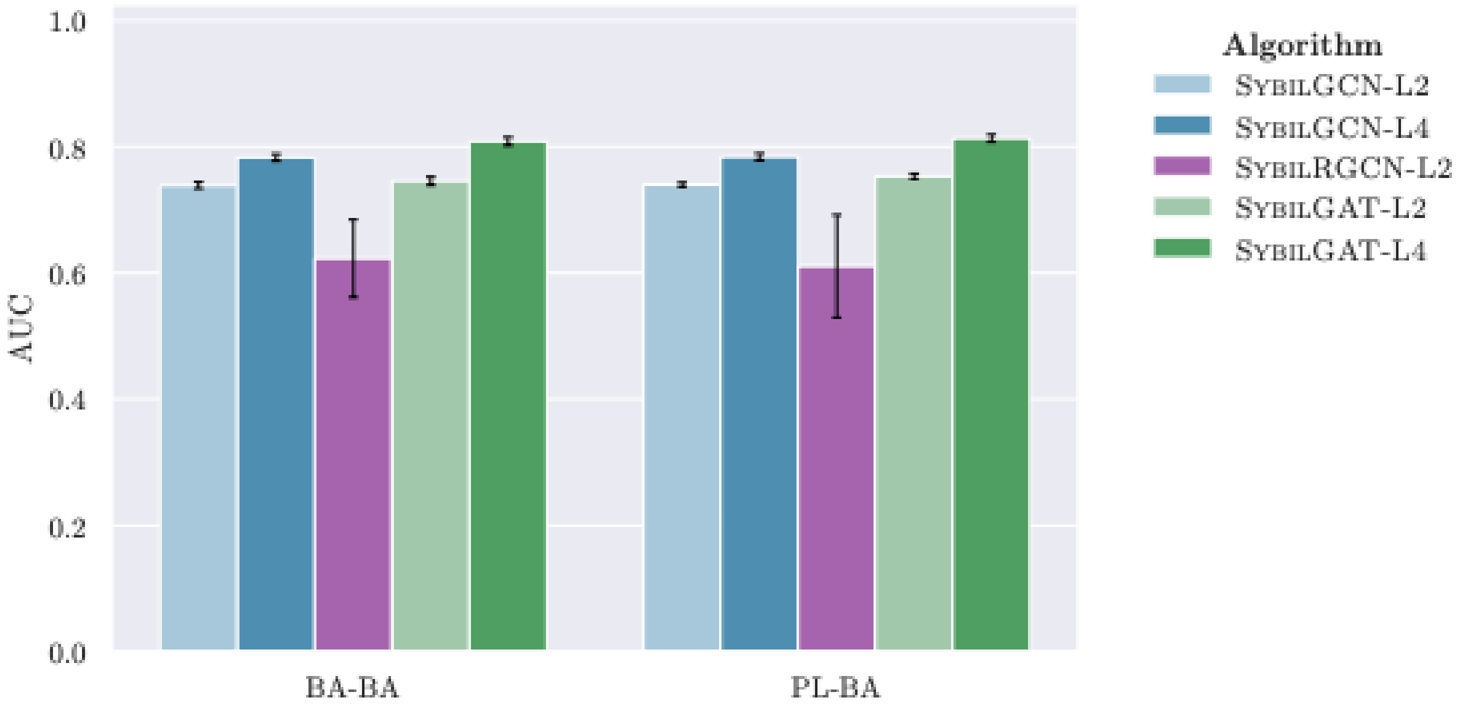
Modern Sybils

«The Eiffel Tower is a famous landmark in Paris, France. It stands tall at 330 meters (1,083 feet) with a square base. It is named after the engineer Gustav Eiffel.»

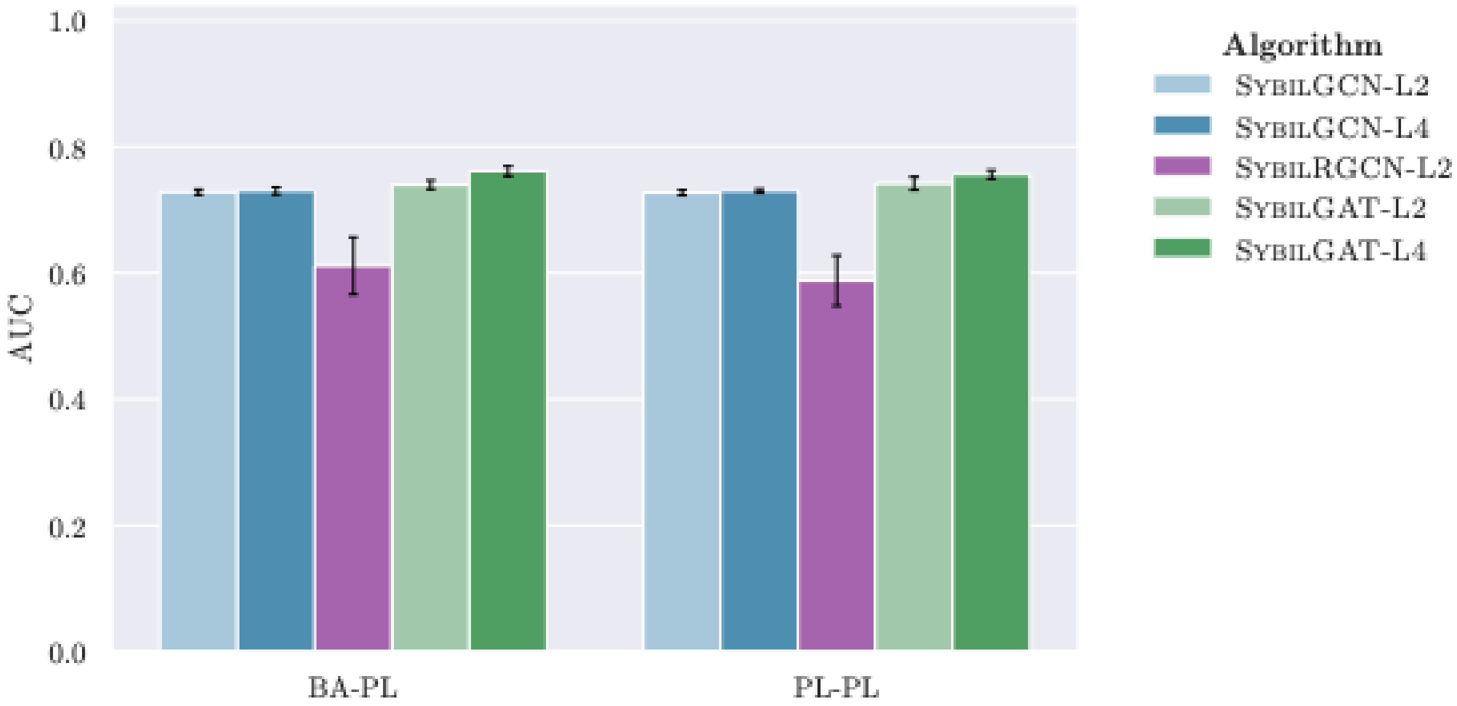
«The Eiffel Tower stands in Paris, France. It reaches a height of 330 meters (1,083 feet). The tower is named after Gustave Eiffel, the engineer whose company designed and built it.»

Claude 3.5 Sonnet
«write exactly 3 very short sentences about the eiffel tower, including its location, height and name origin»

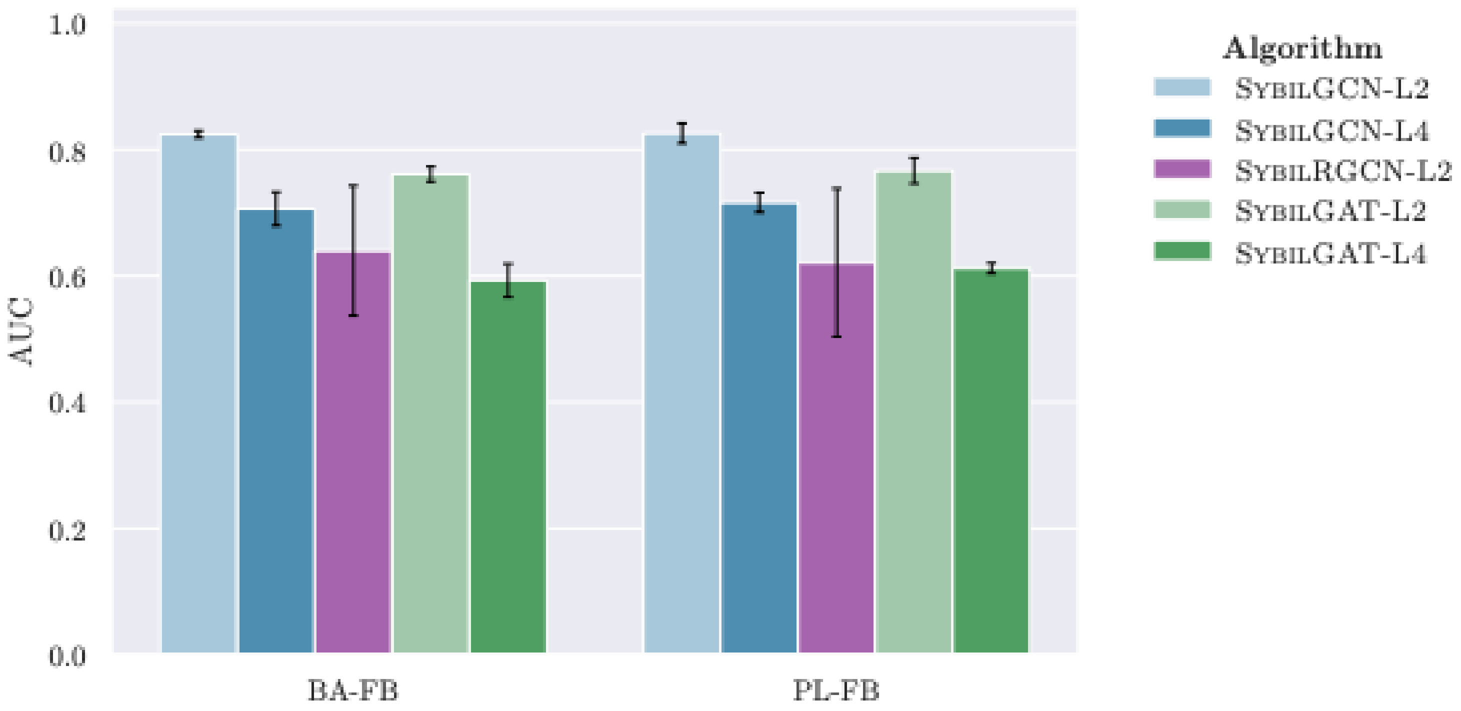
Results: Pre-training on Small Network



Results: Pre-training on Small Network

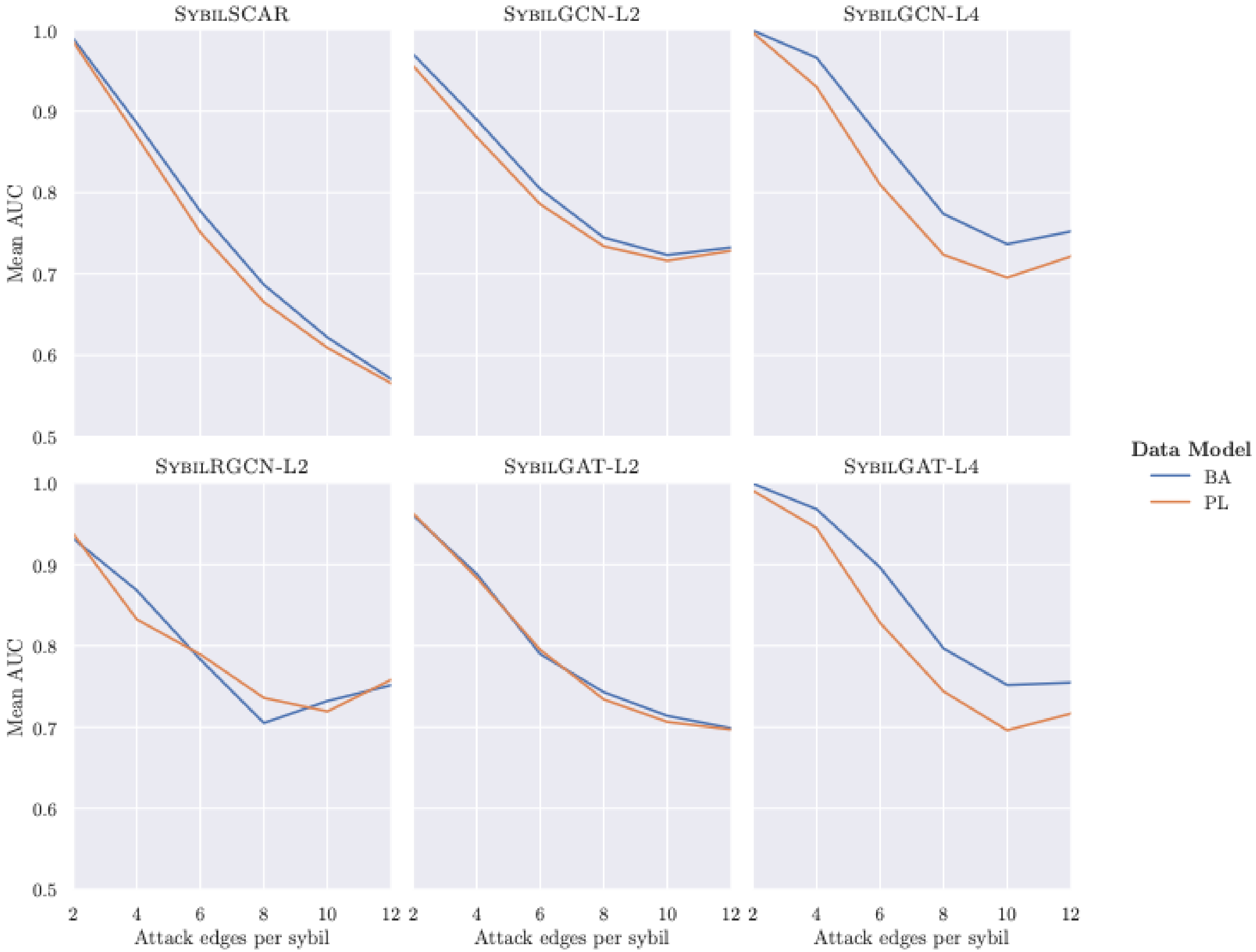


Results: Pre-training on Small Network



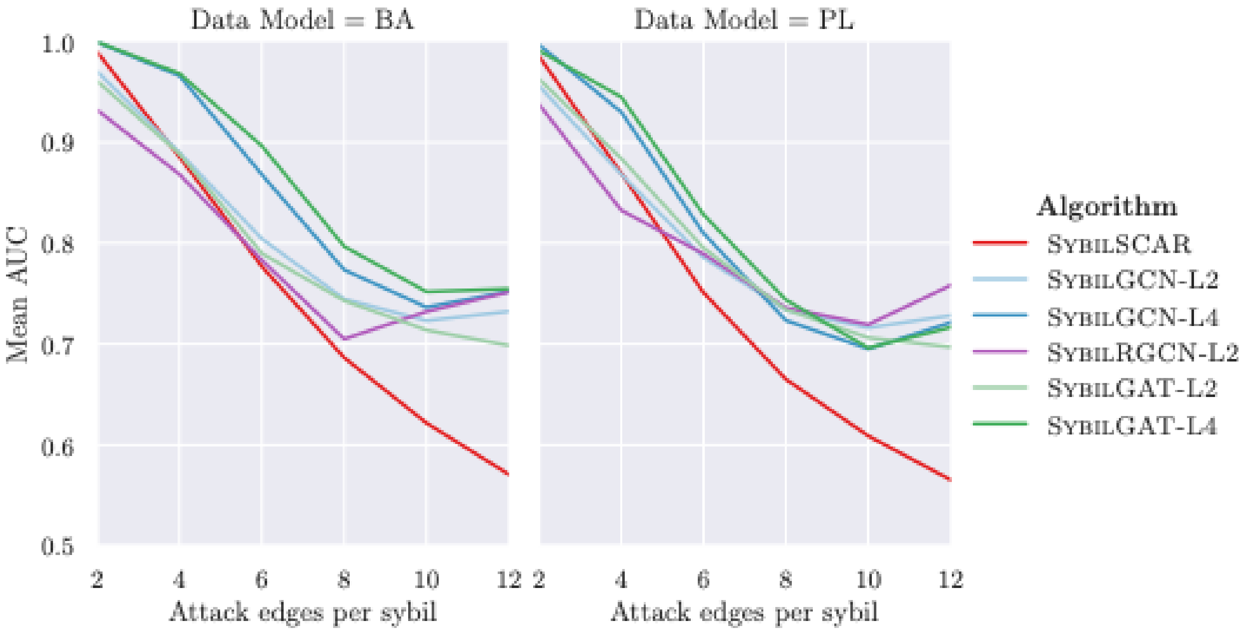
Results: Robustness

Data Model



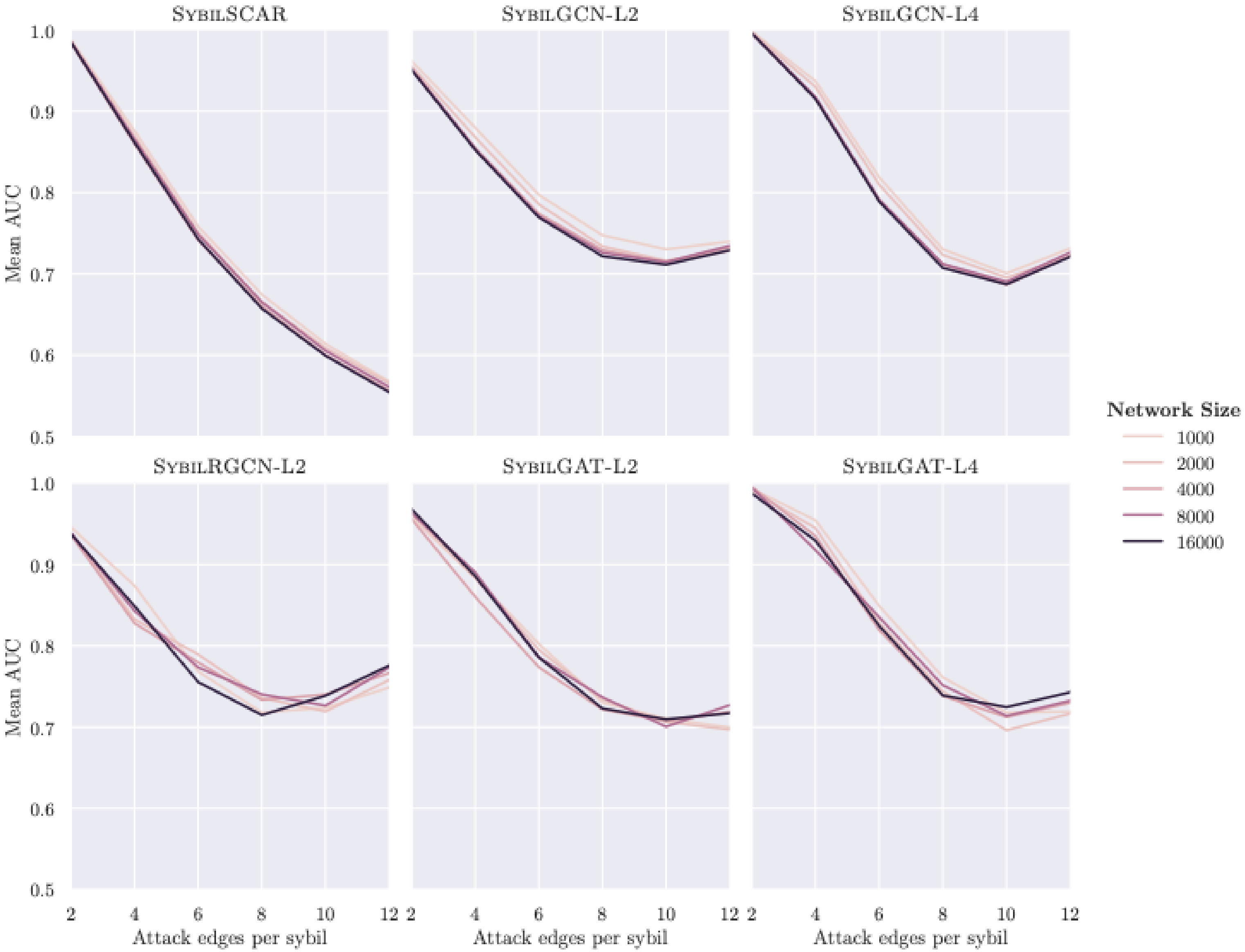
Results: Robustness

Data Model



Results: Robustness

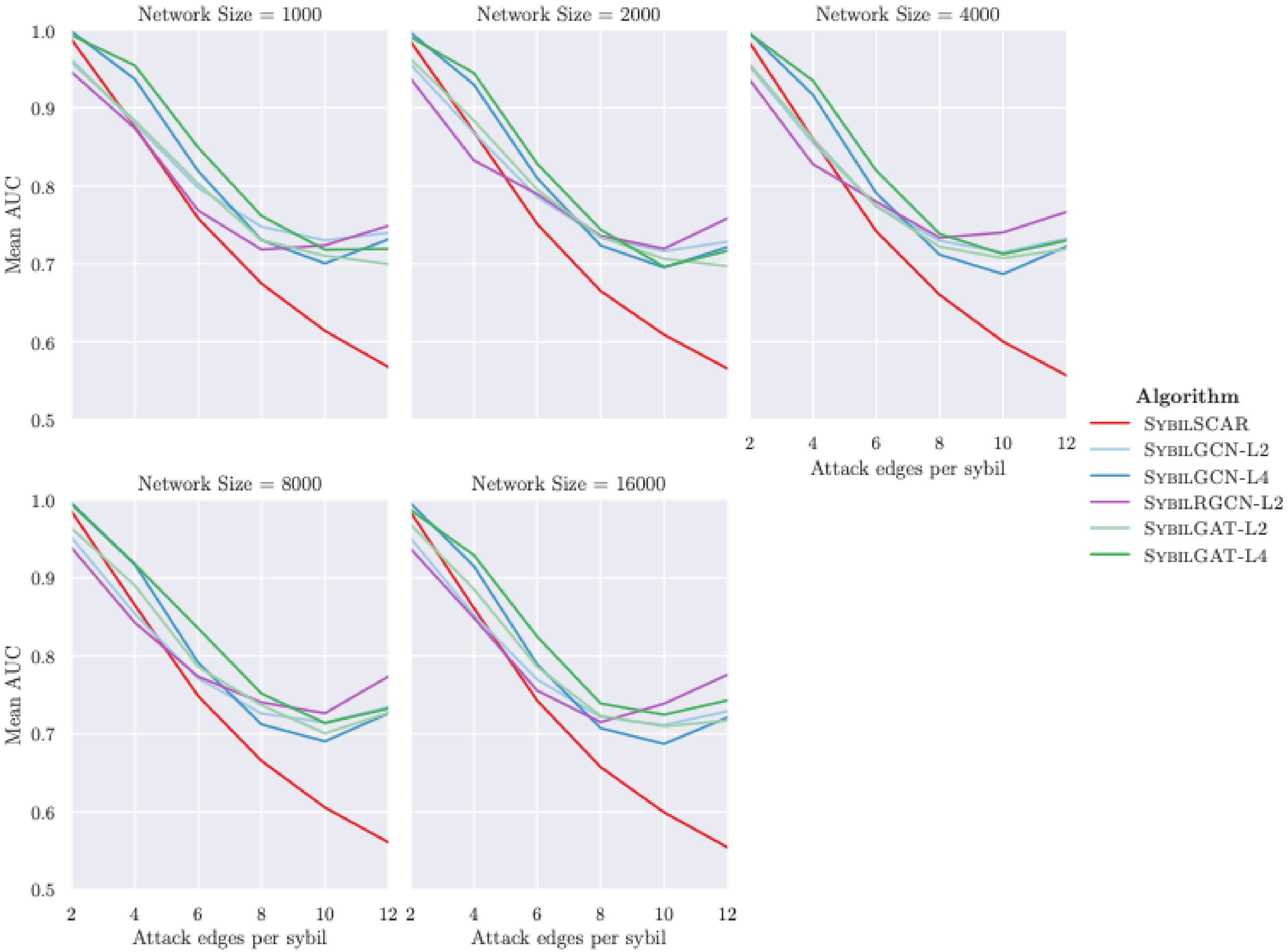
Network Size



ETH zürich

Results: Robustness

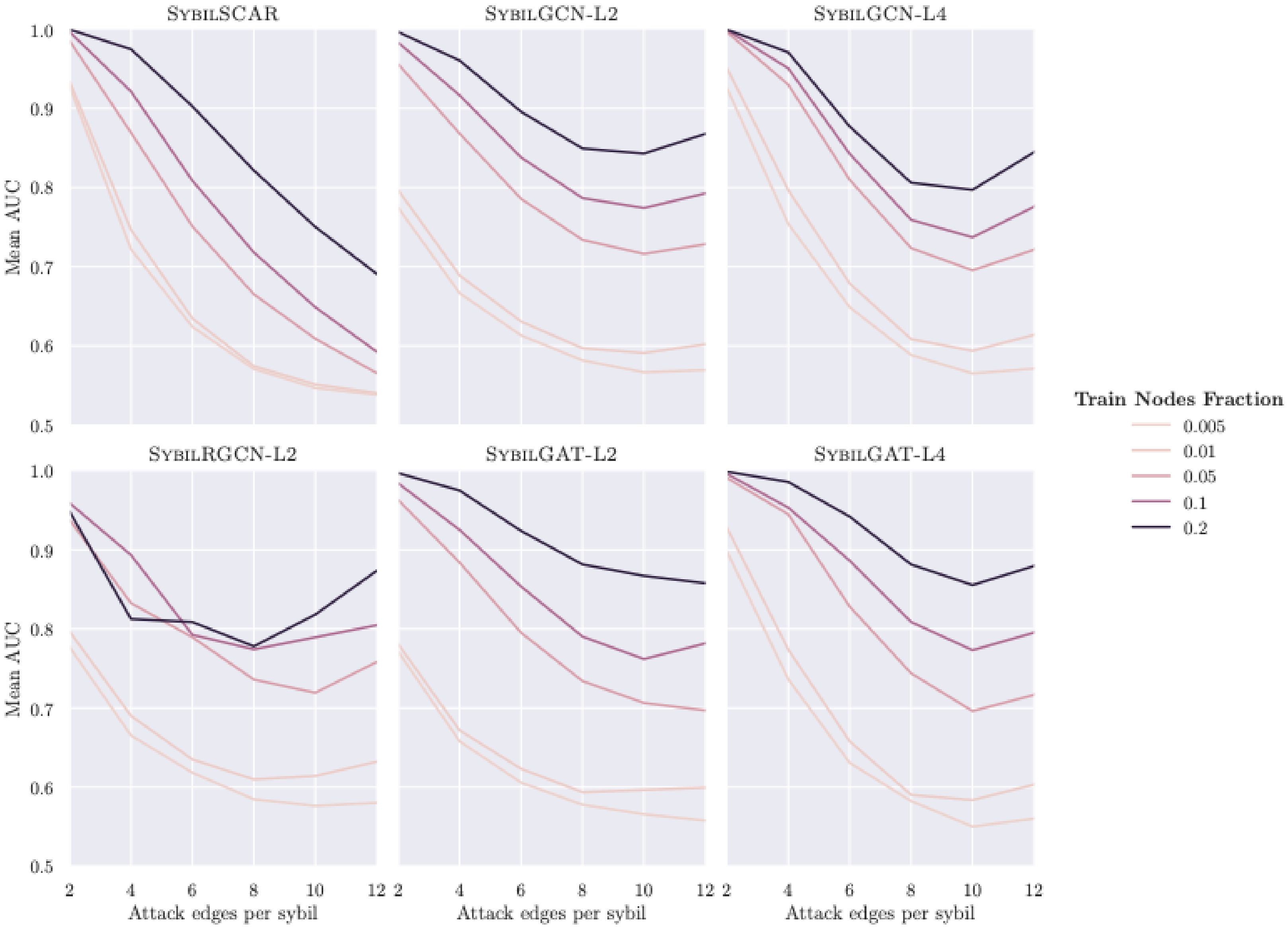
Network Size



ETH zürich

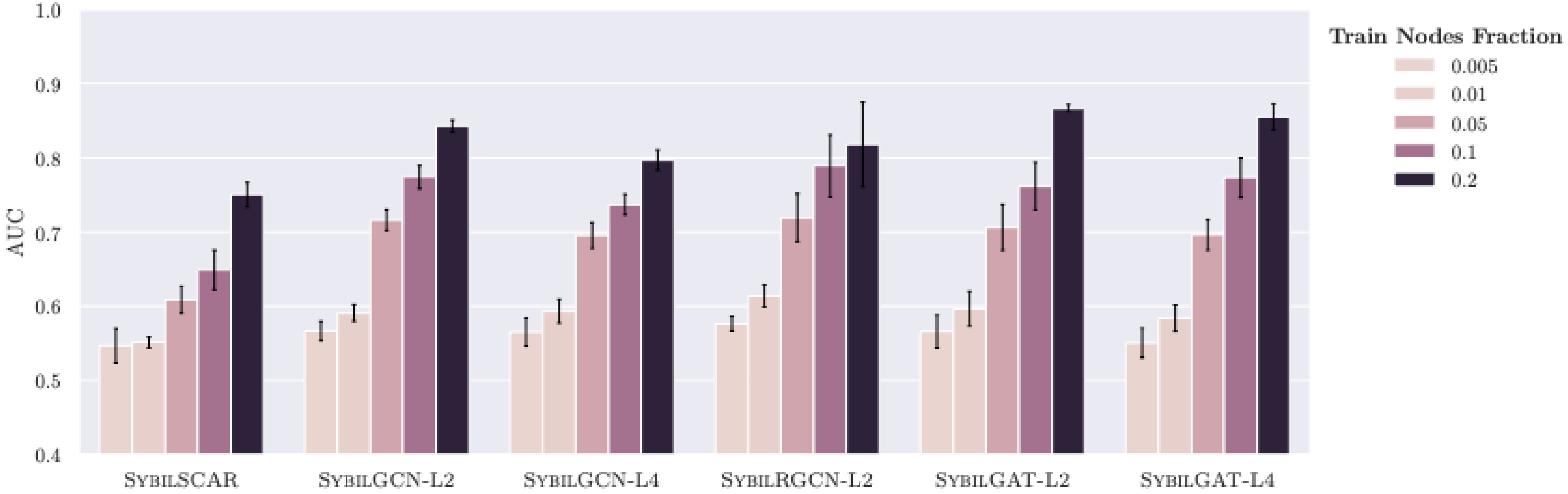
Results: Robustness

Training Set Size



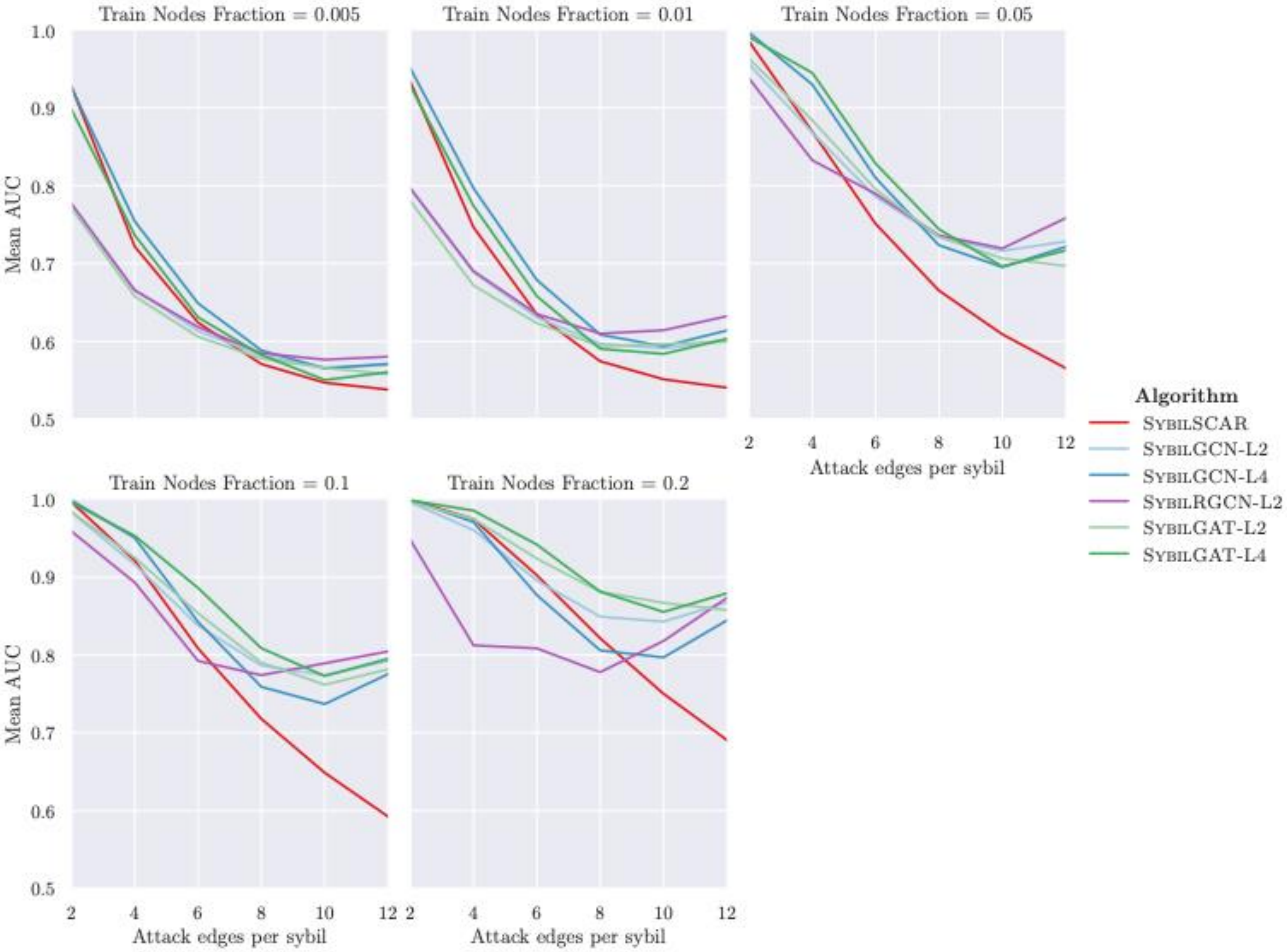
Results: Robustness

Training Set Size



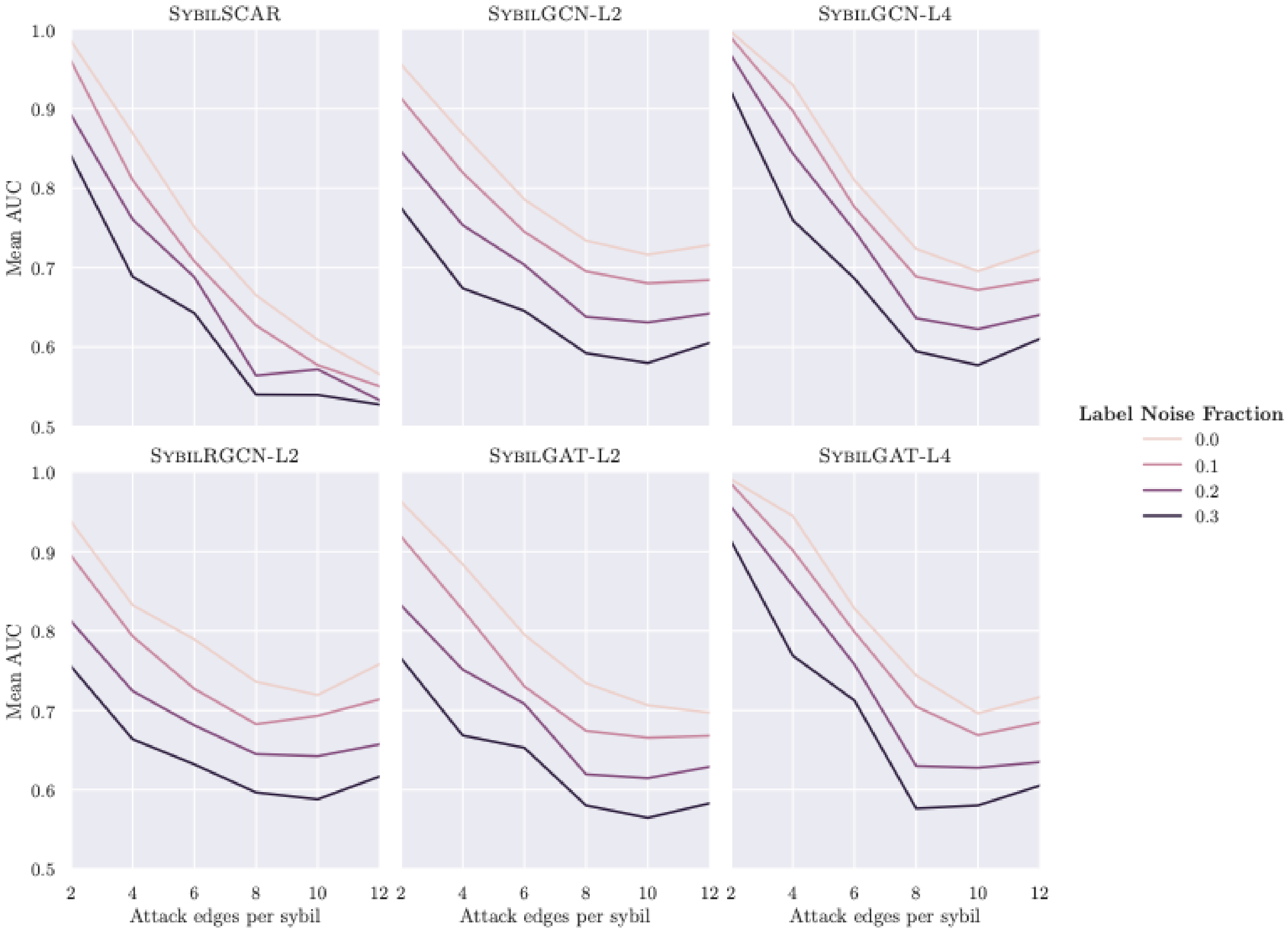
Results: Robustness

Training Set Size



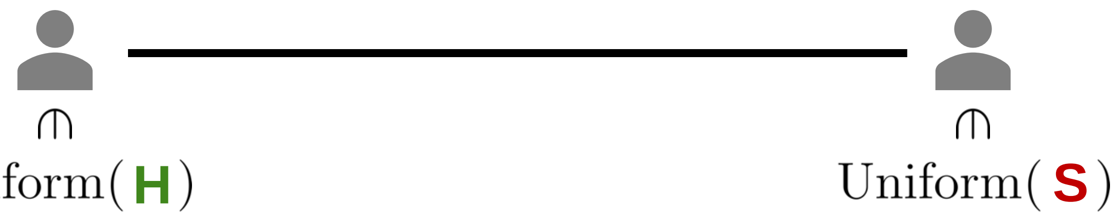
Results: Robustness

Label Noise Level

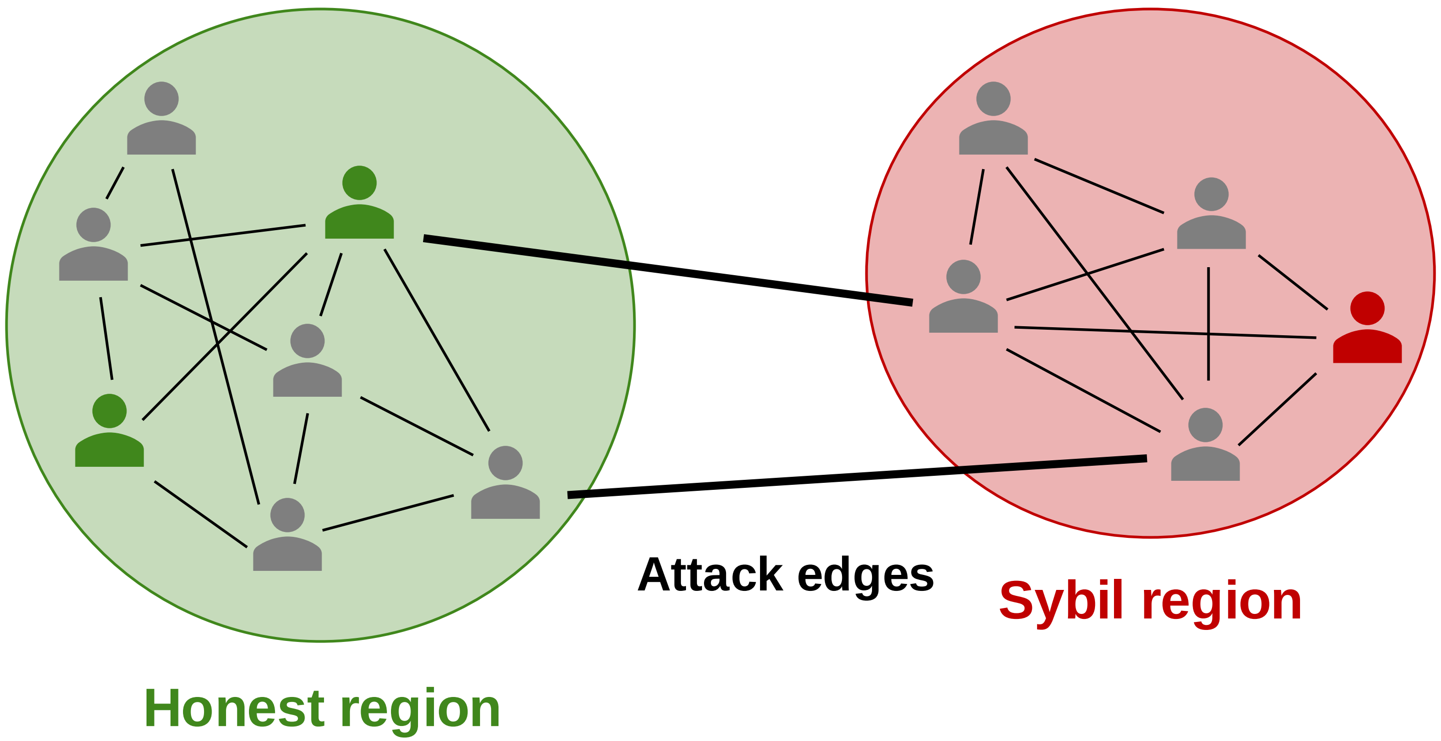


Adversarial Attack

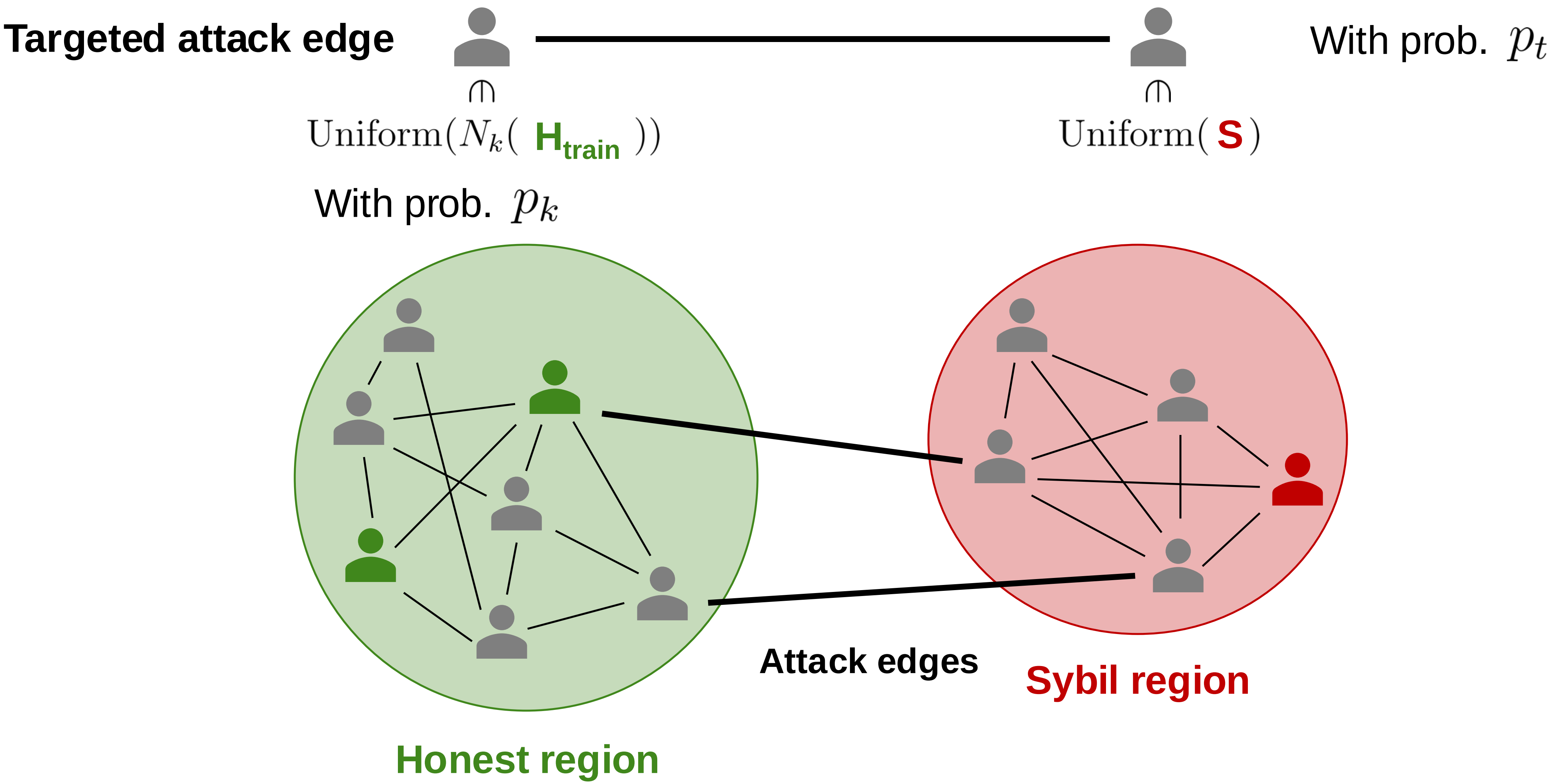
Random attack edge



With prob. $1 - p_t$



Adversarial Attack



Targeted attack edge

With prob. p_t

$\mathrm{Uniform}(N_k(\,\mathbf{H}_{\text{train}}\,))$

$\mathrm{Uniform}(\,\mathbf{S}\,)$

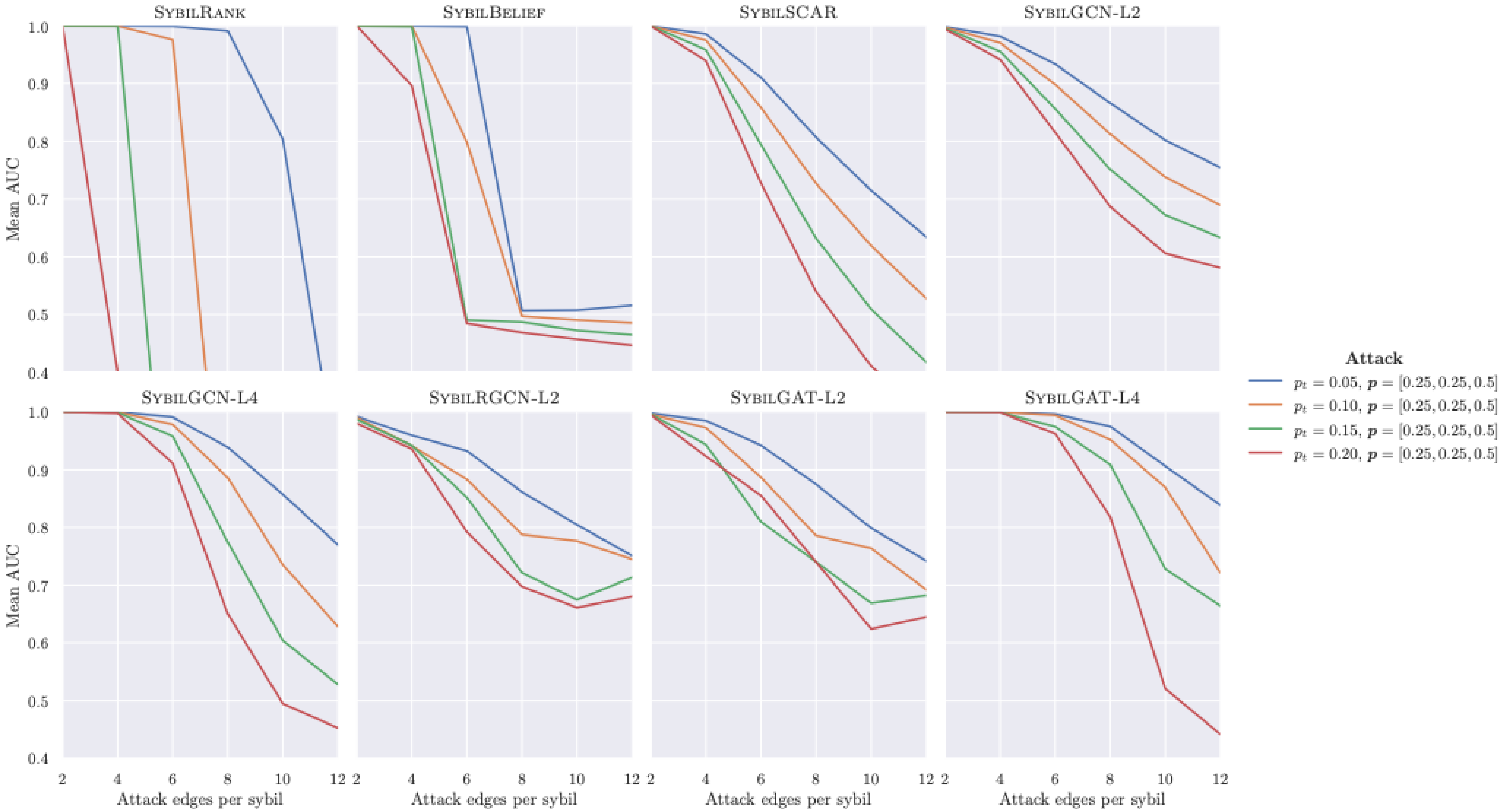
With prob. p_k

Attack edges

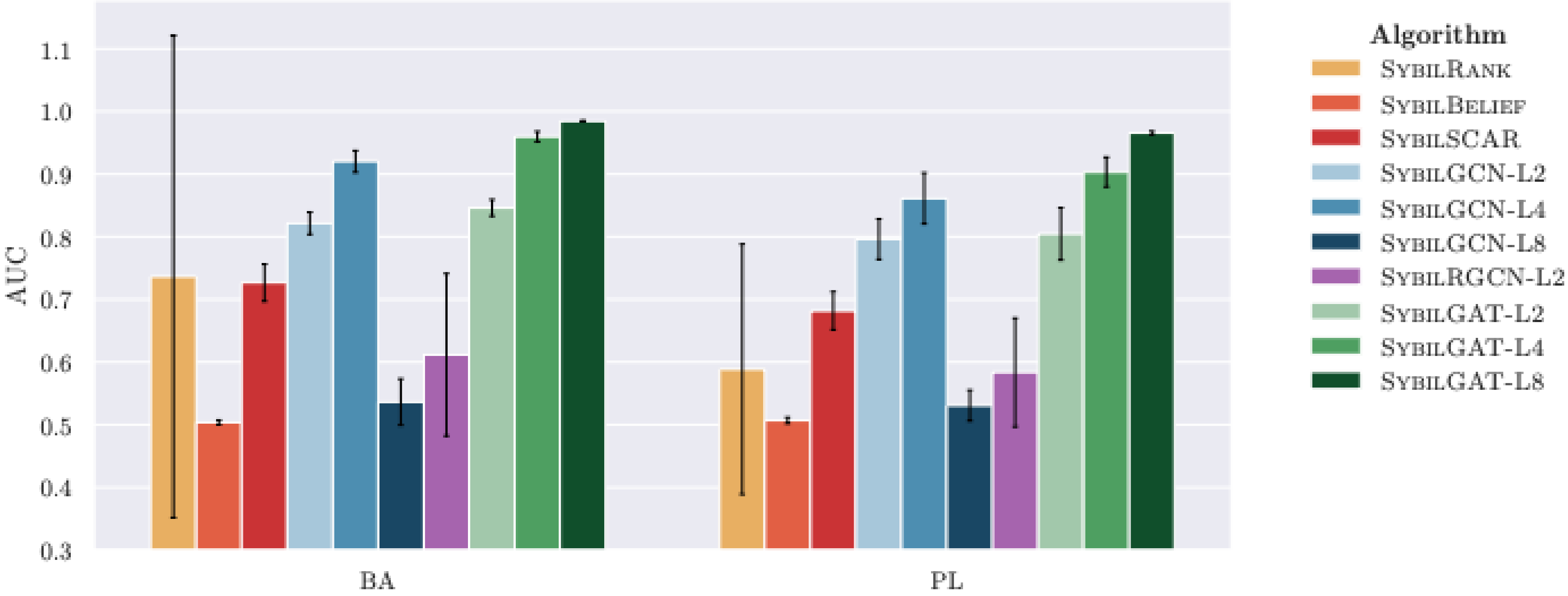
Sybil region

Honest region

Results: Adversarial Attacks



Results: Pre-training Before Attack


AUC
BA
PL
Algorithm
SybilRank
SybilBelief
SybilSCAR
SybilGCN-L2
SybilGCN-L4
SybilGCN-L8
SybilRGCN-L2
SybilGAT-L2
SybilGAT-L4
SybilGAT-L8