RICC: Robust Collective Classification of Sybil Accounts

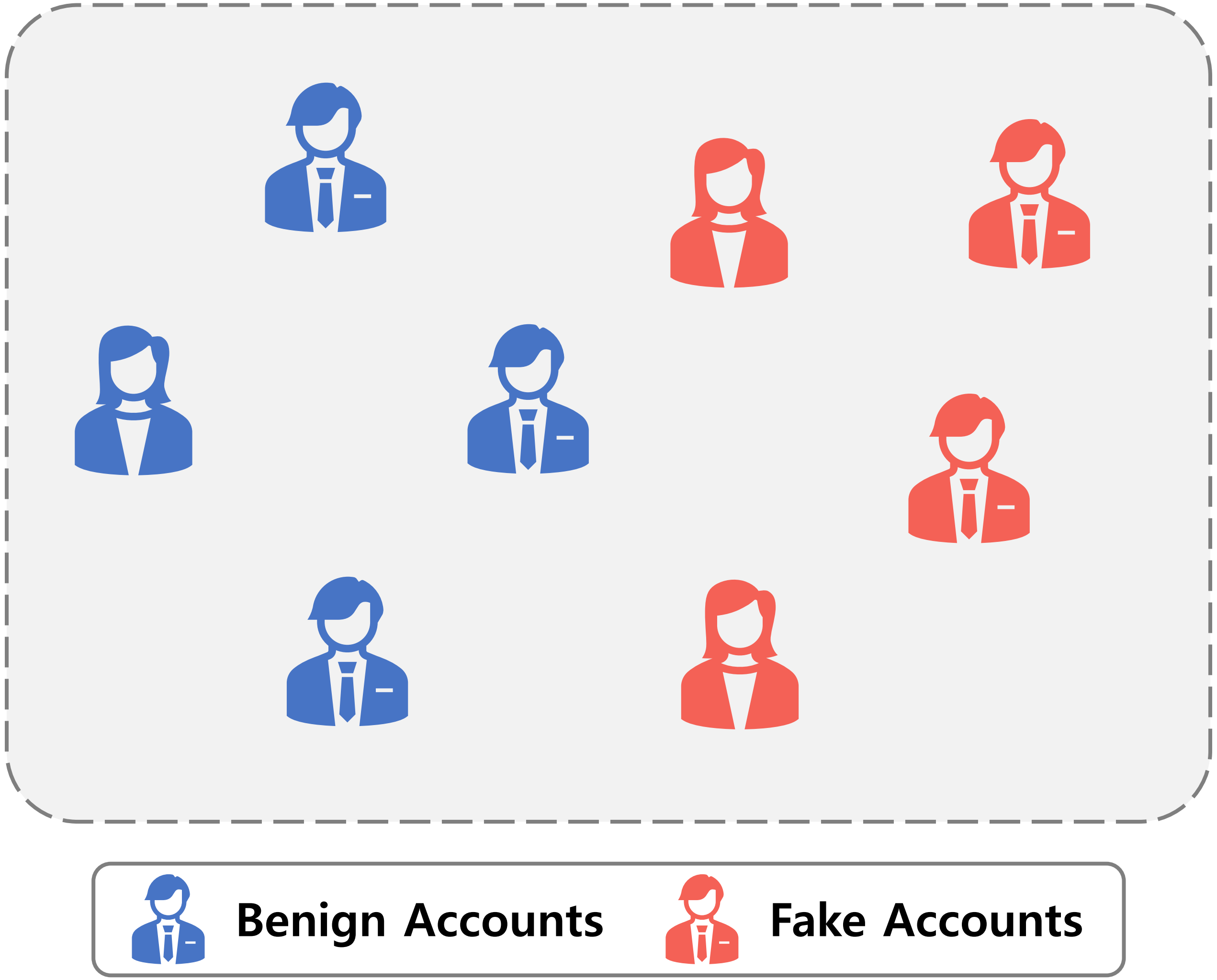
Dongwon Shin*, Suyoung Lee*, and Sooel Son

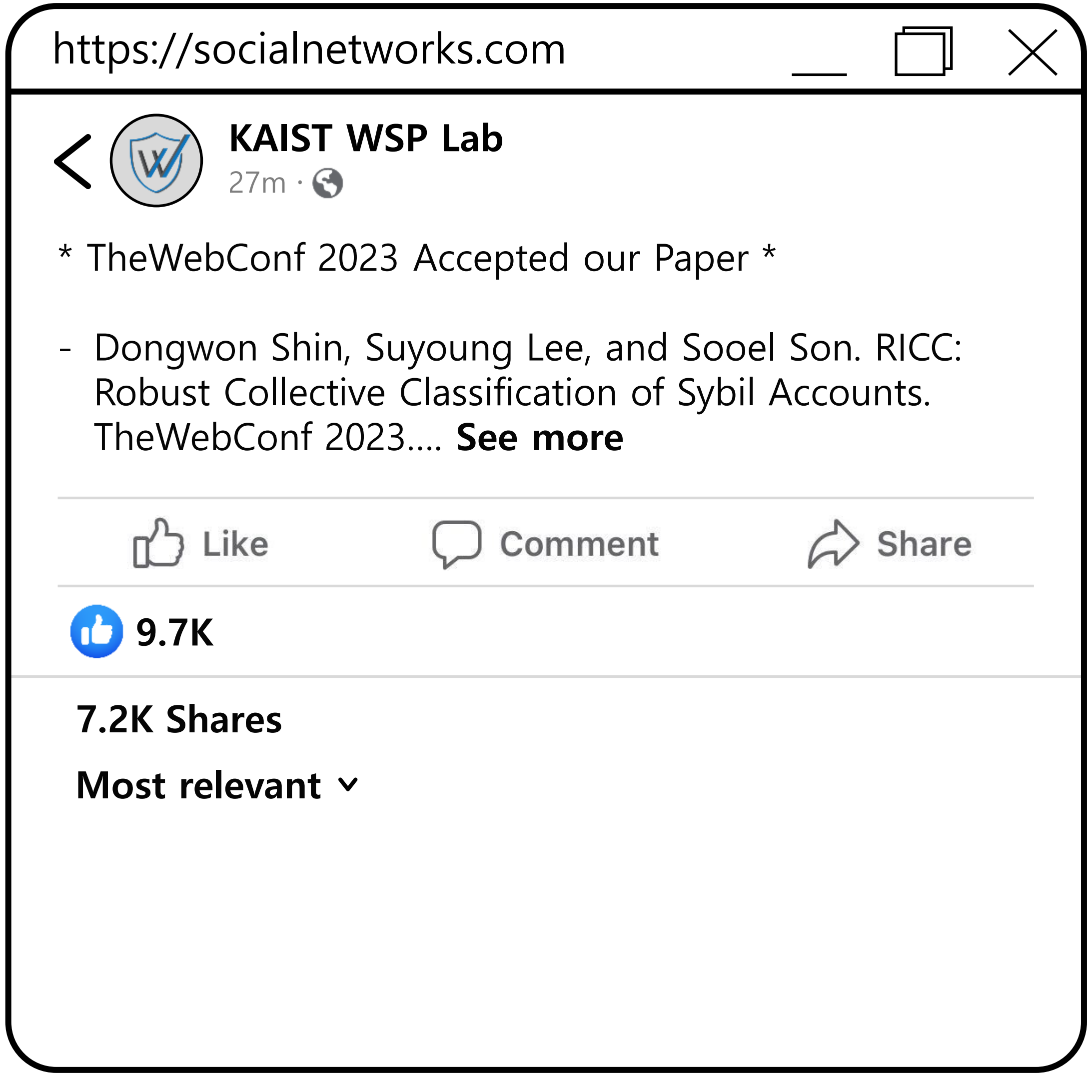
KAIST

TheWebConf 2023

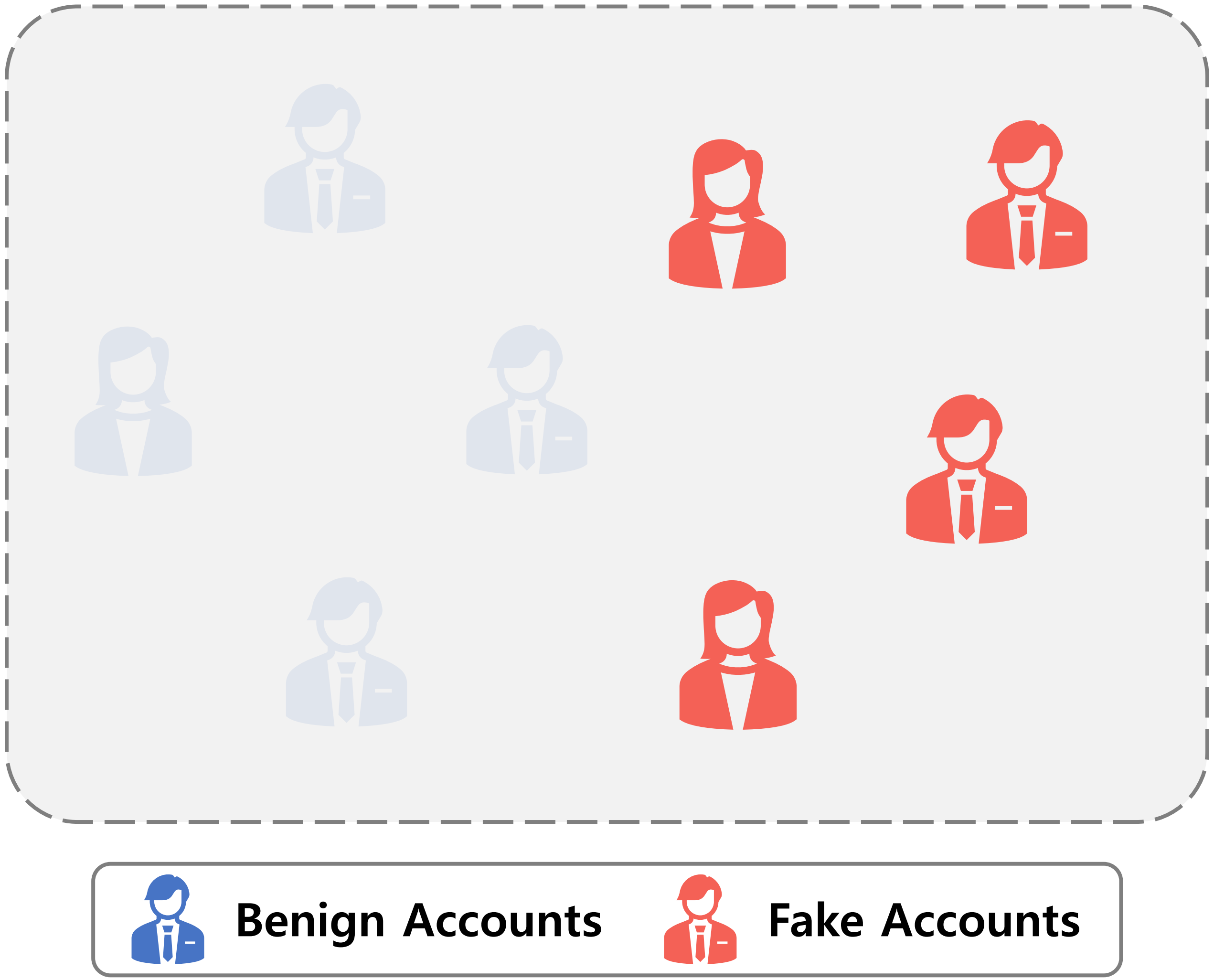
* Both authors contributed equally to the paper.

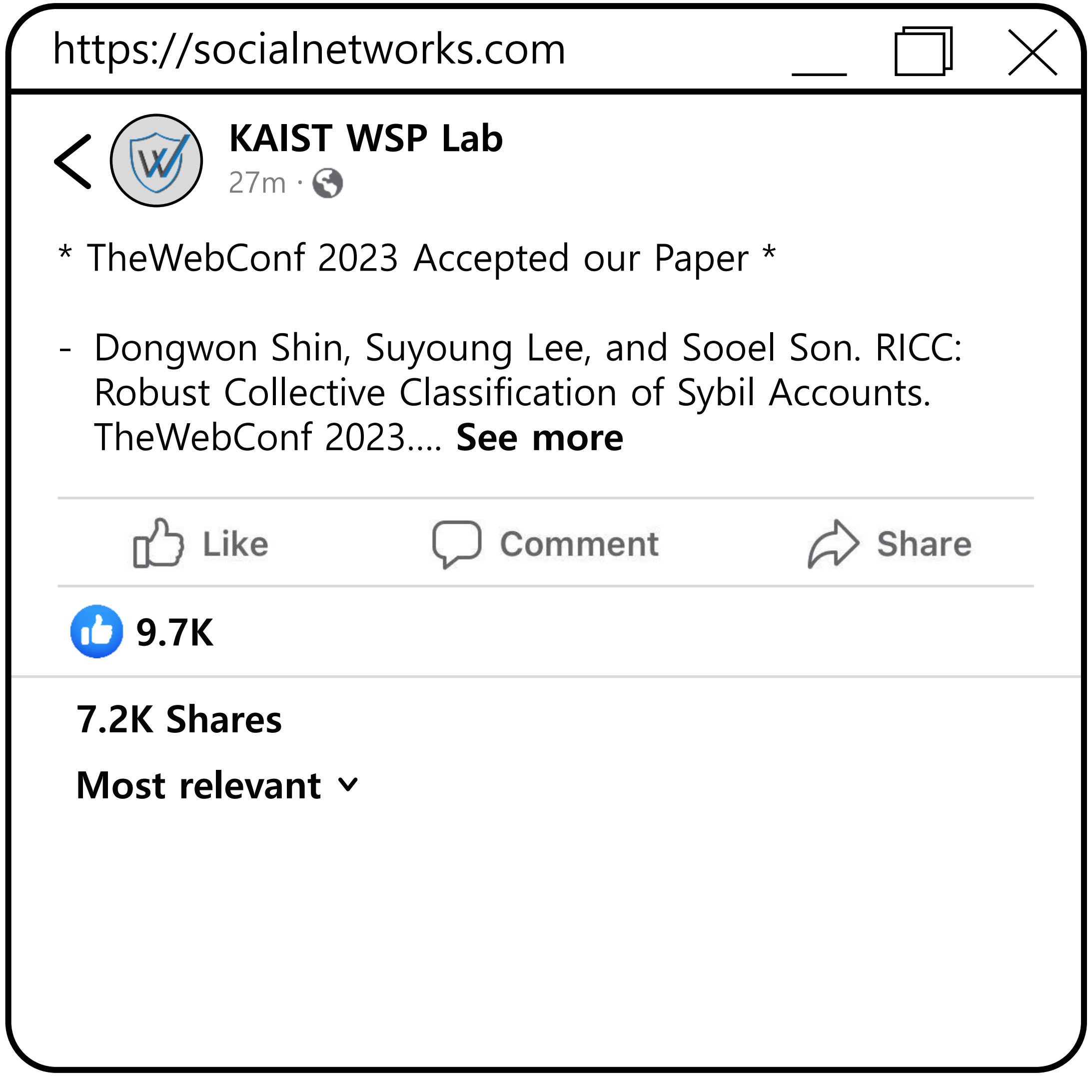
Fake User Accounts



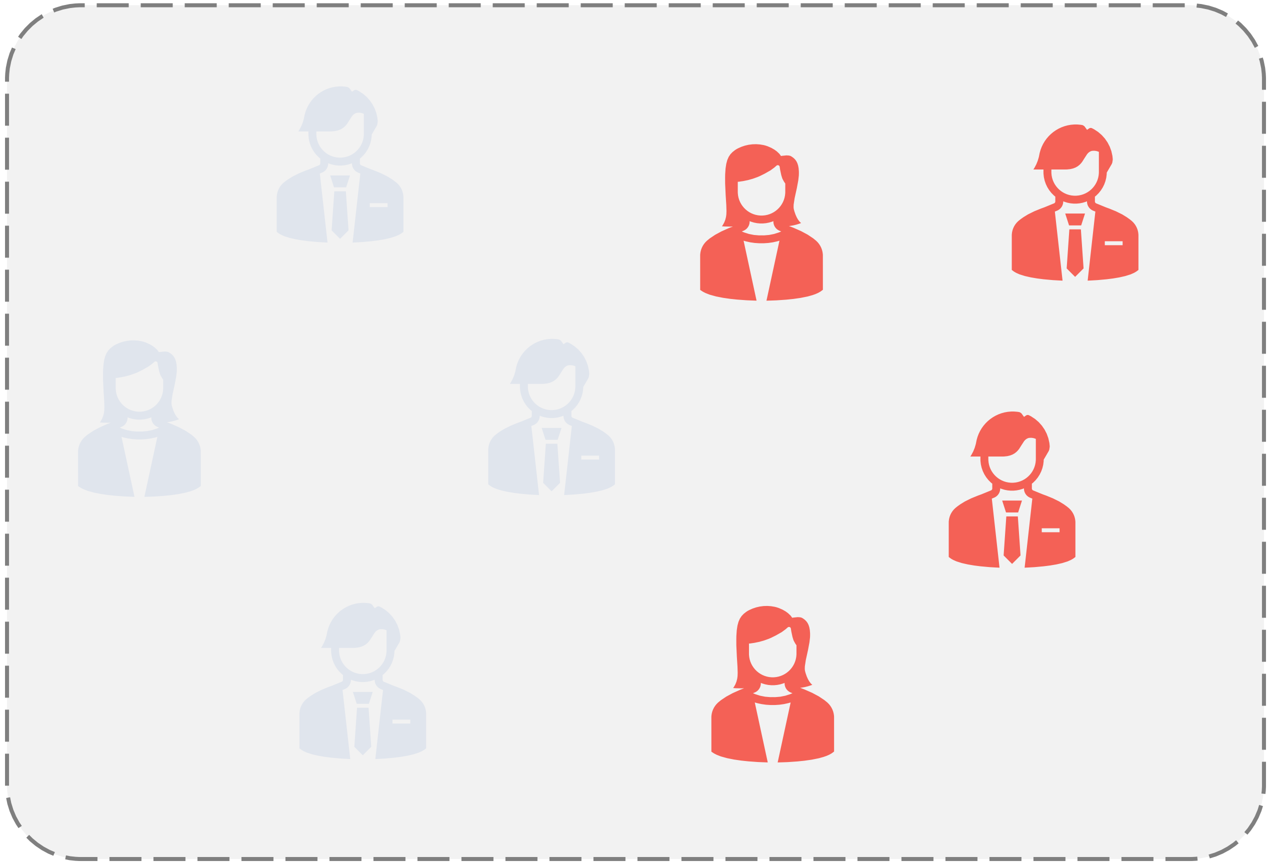


Fake User Accounts





Fake User Accounts



Benign Accounts Fake Accounts

Sybil accounts

https://socialnetworks.com

KAIST WSP Lab

27m ·

* TheWebConf 2023 Accepted our Paper *

- Dongwon Shin, Suyoung Lee, and Sooel Son. RICC: Robust Collective Classification of Sybil Accounts. TheWebConf 2023.... **See more**

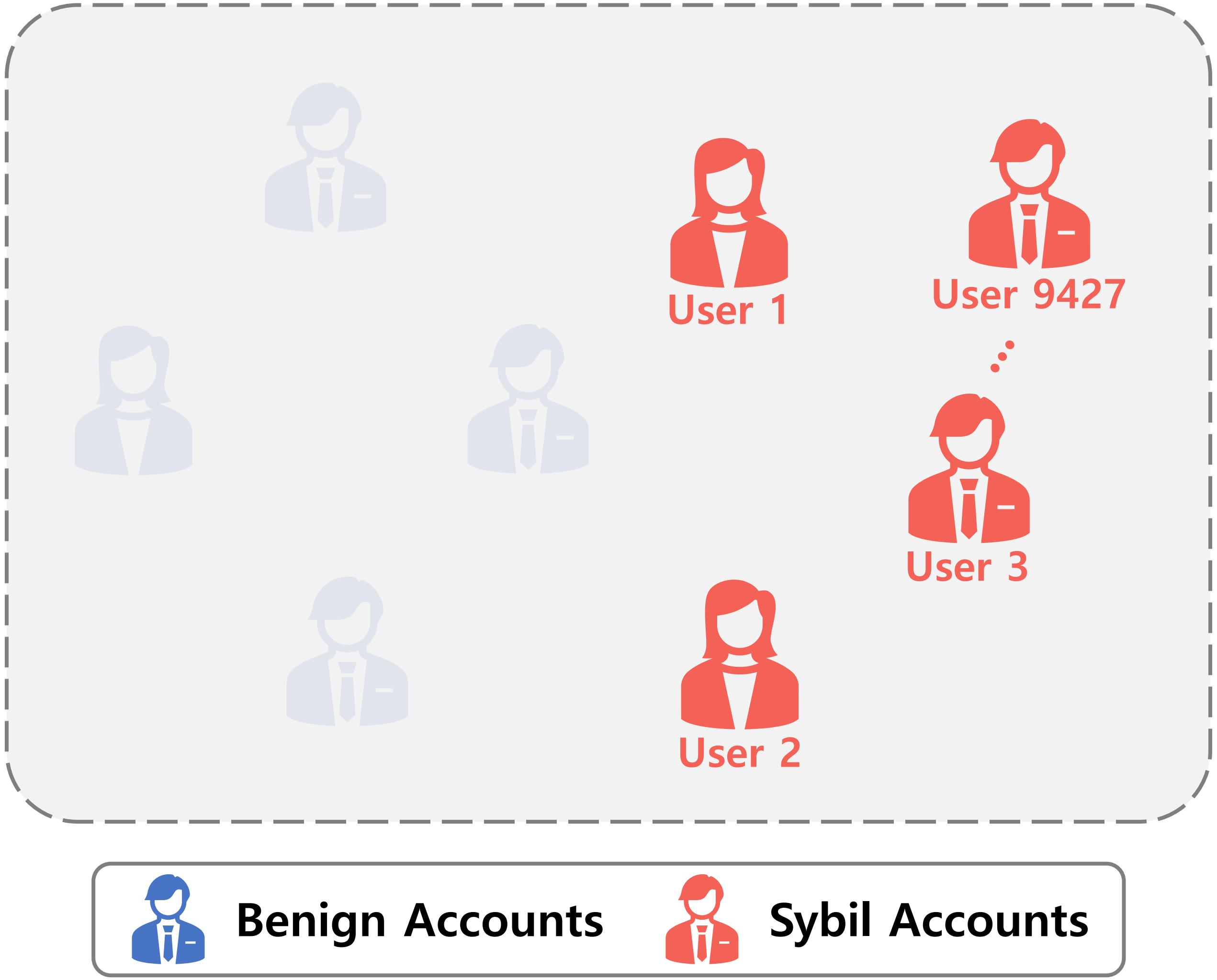
Like Comment Share

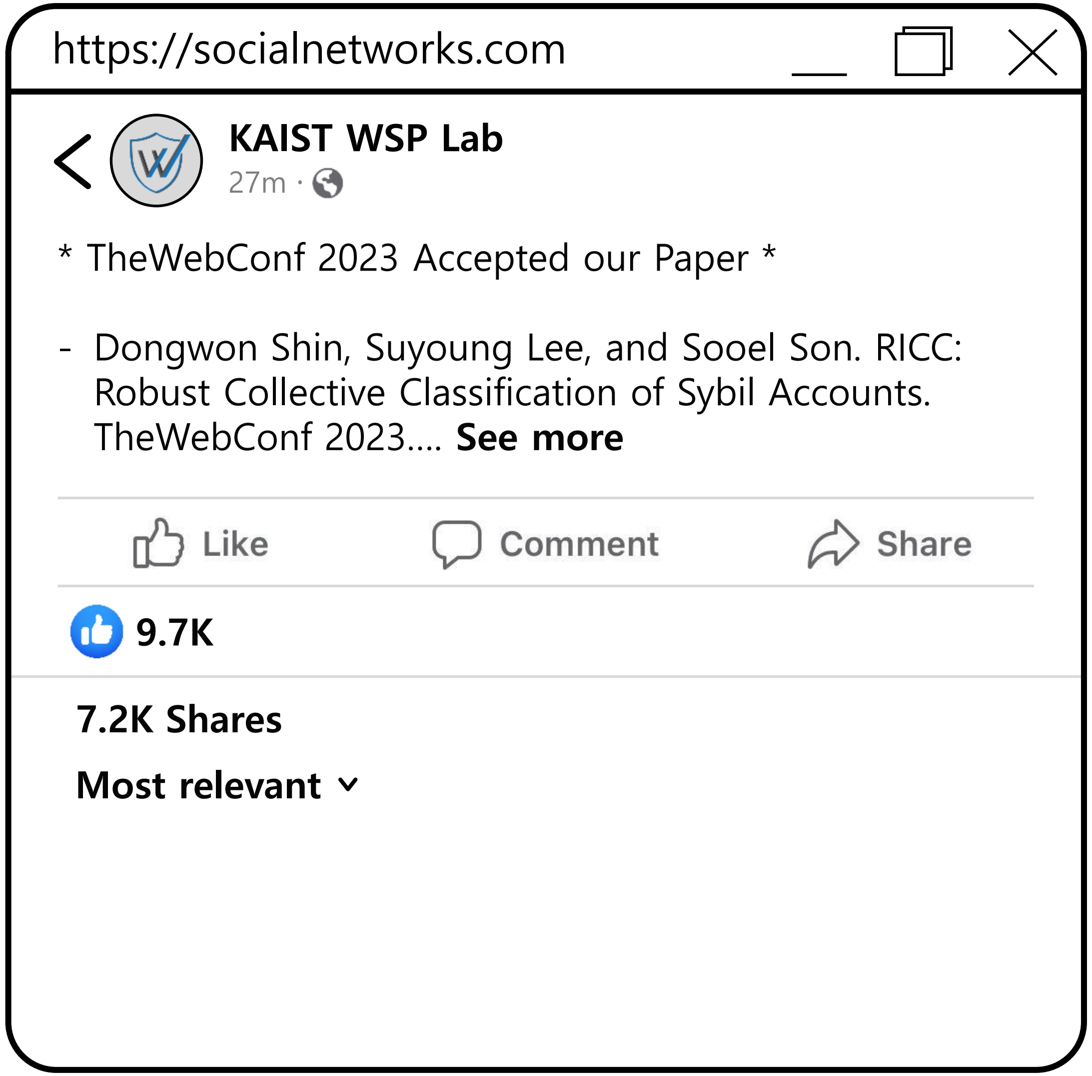
9.7K

7.2K Shares

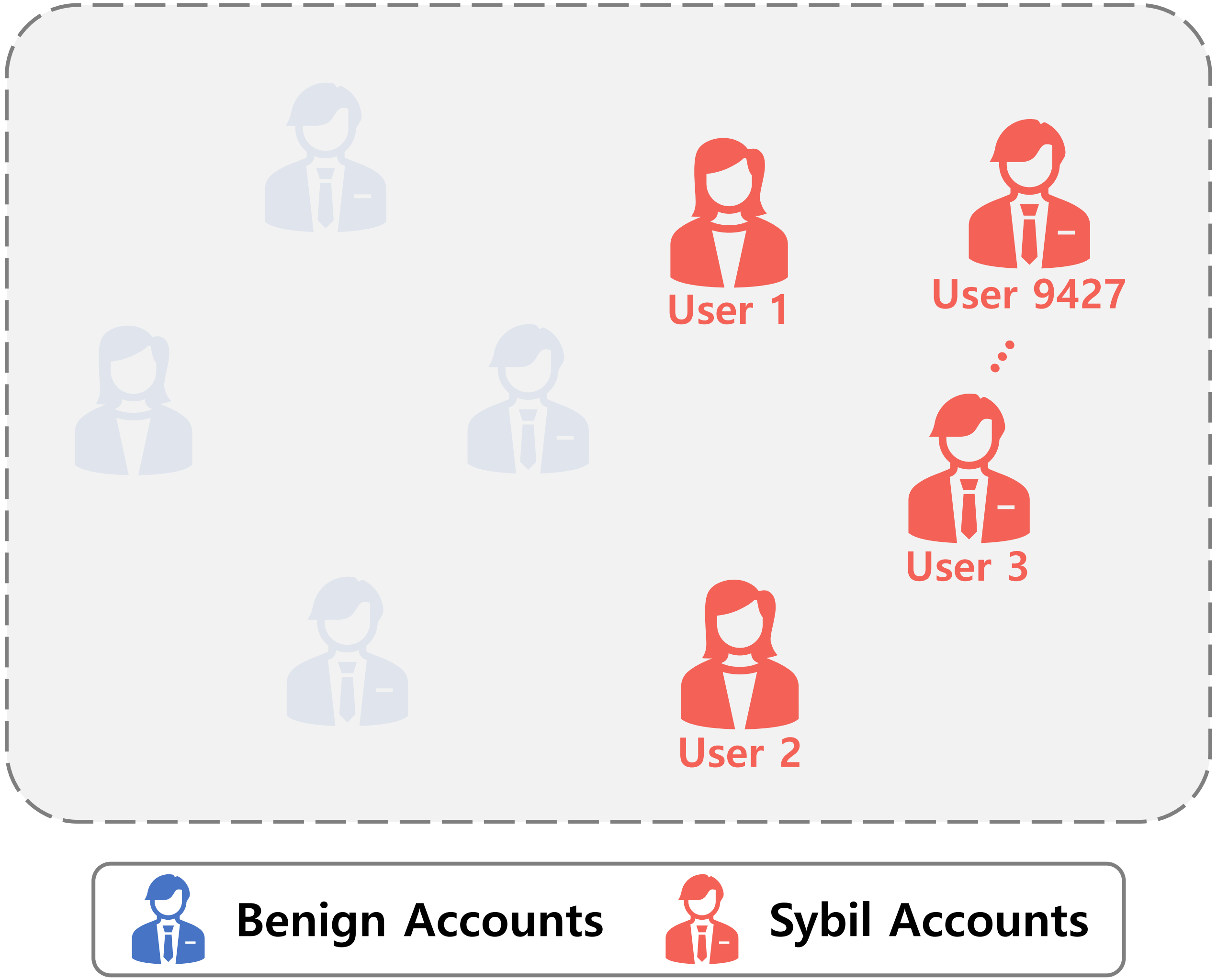
Most relevant

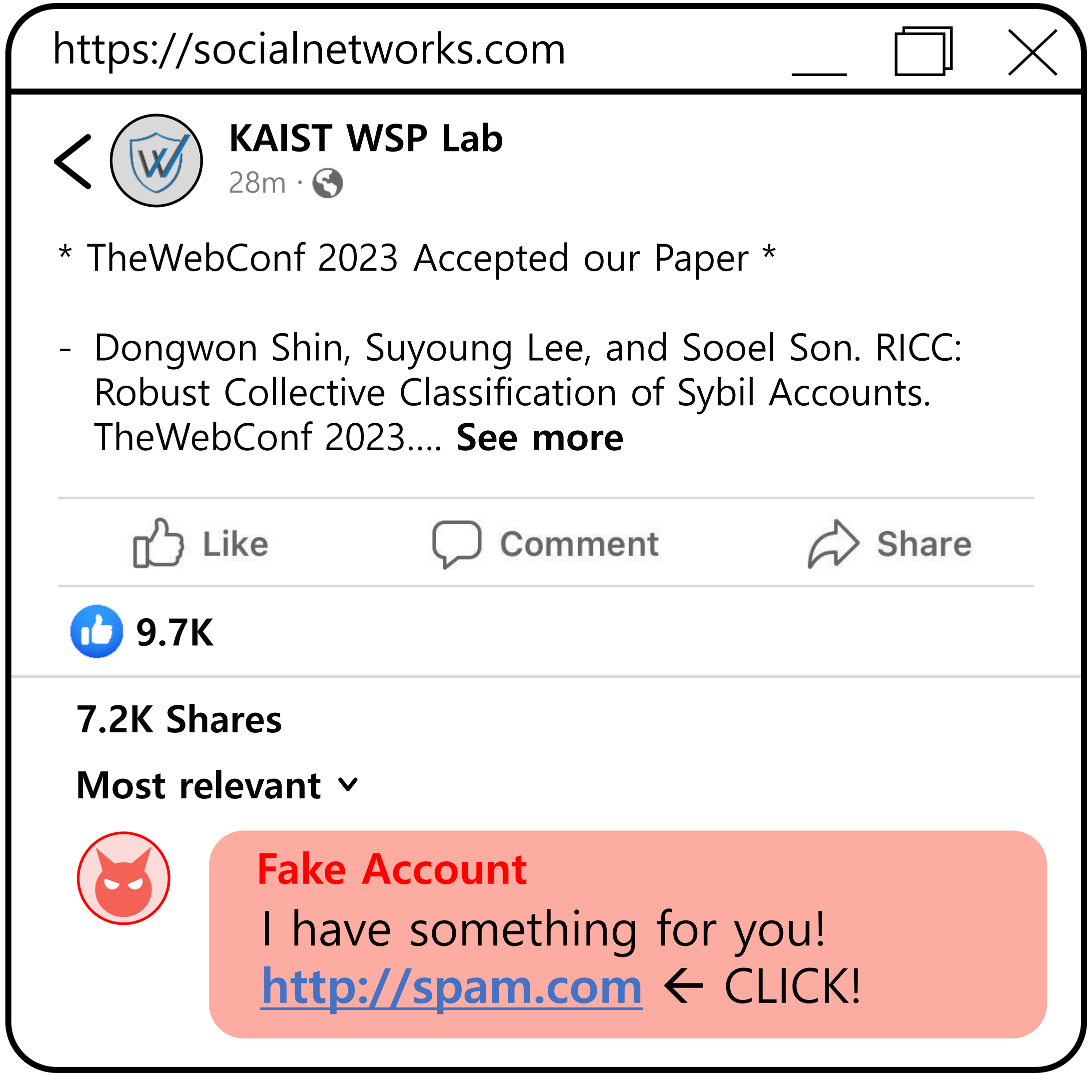
Sybil Accounts



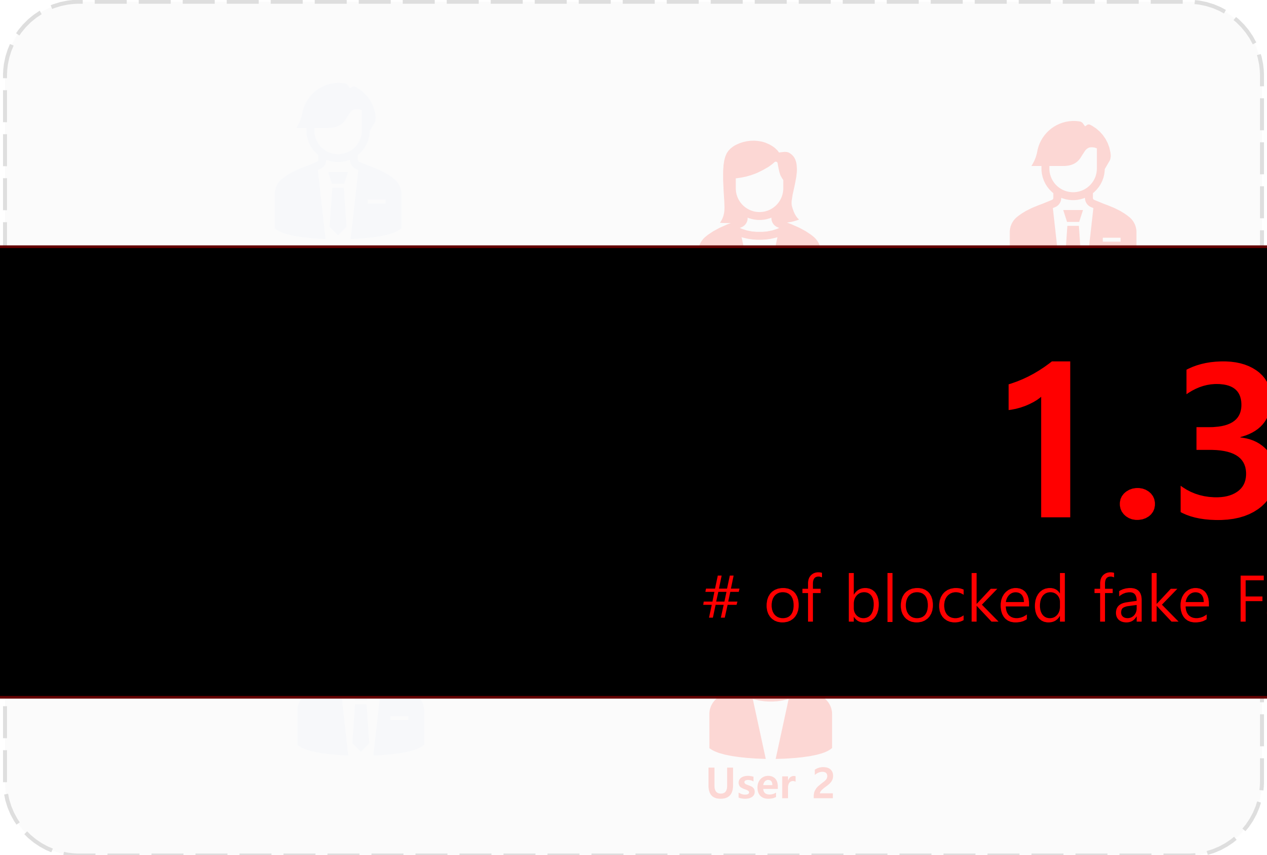


Sybil Accounts





Sybil Accounts



Benign Accounts   Sybil Accounts

https://socialnetworks.com

KAIST WSP Lab
28m ·

* TheWebConf 2023 Accepted our Paper *

Most relevant

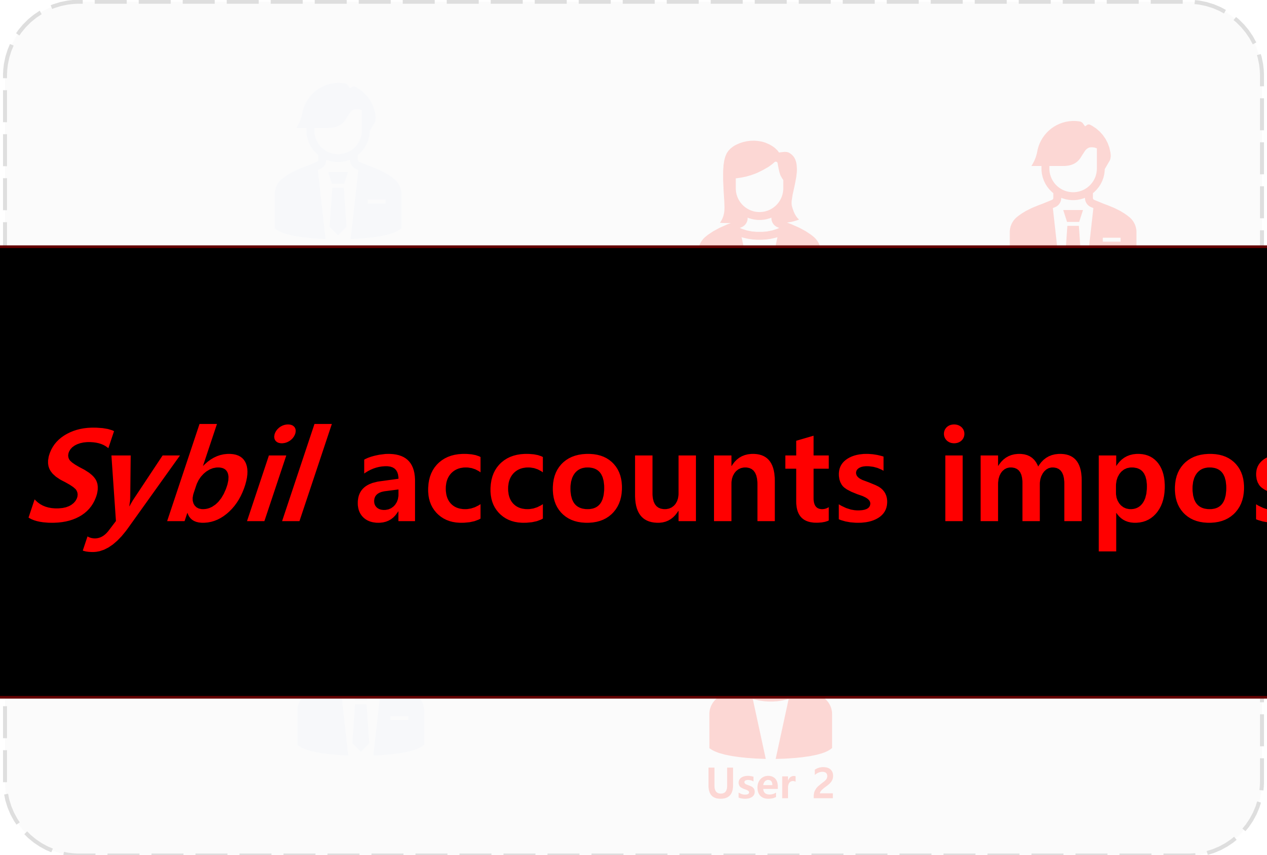
Fake Account
I have something for you!
http://spam.com ← CLICK!

1.3B

of blocked fake Facebook users*

* https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/

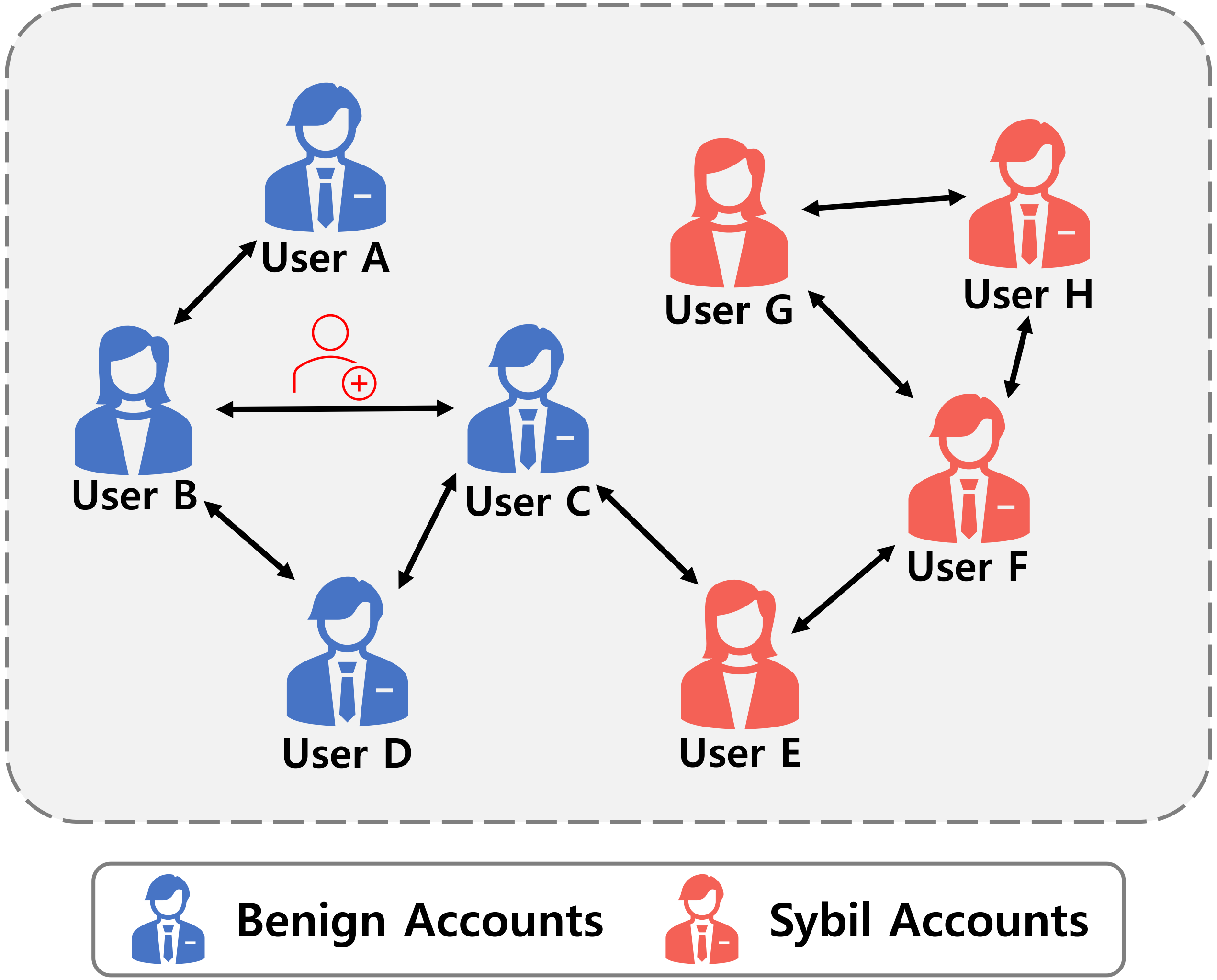
Sybil Accounts

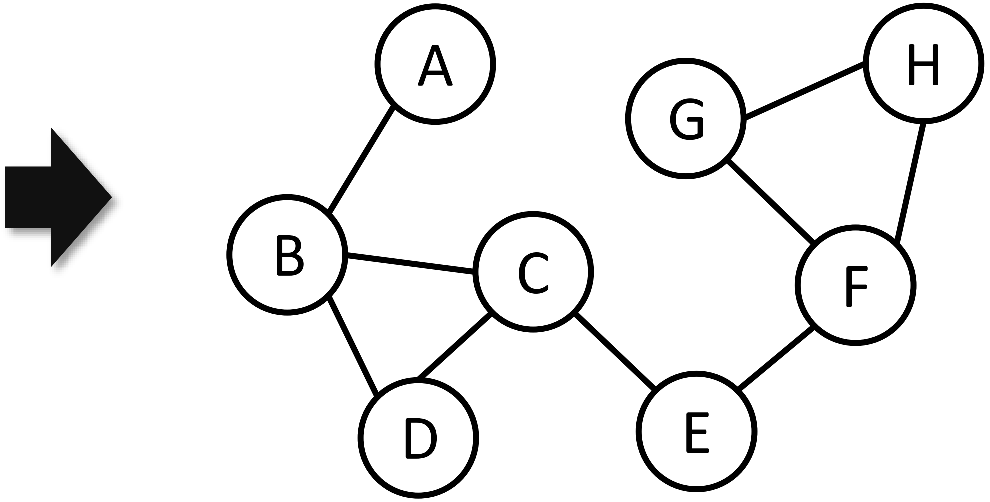


Benign Accounts   Sybil Accounts

***Sybil* accounts impose a critical threat!**

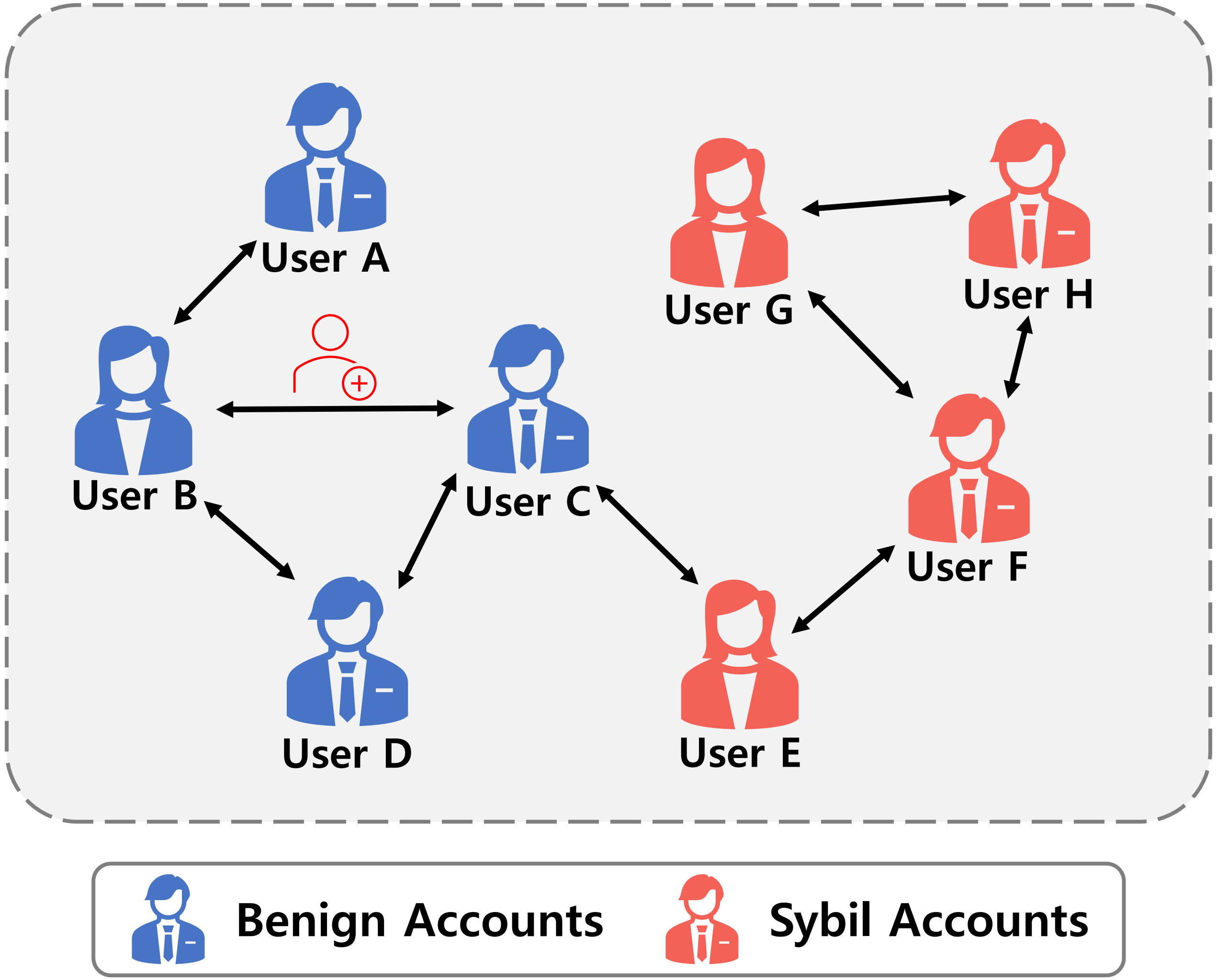
Graph-based Sybil Account Detection



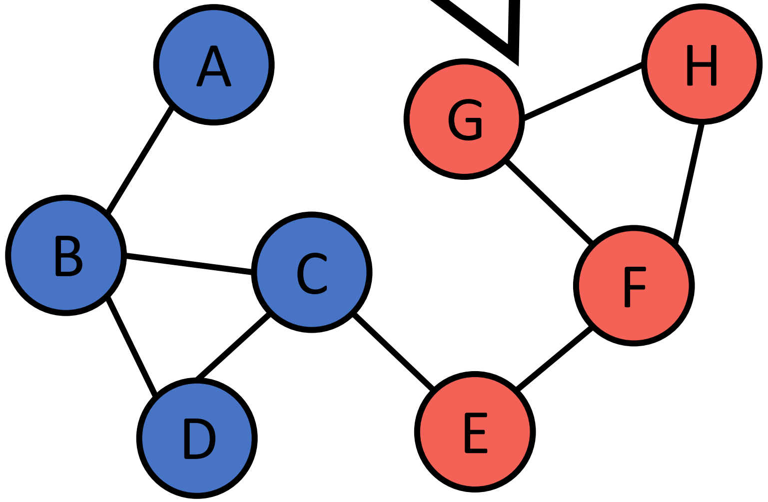


Graph-based Modeling

Graph-based Sybil Account Detection

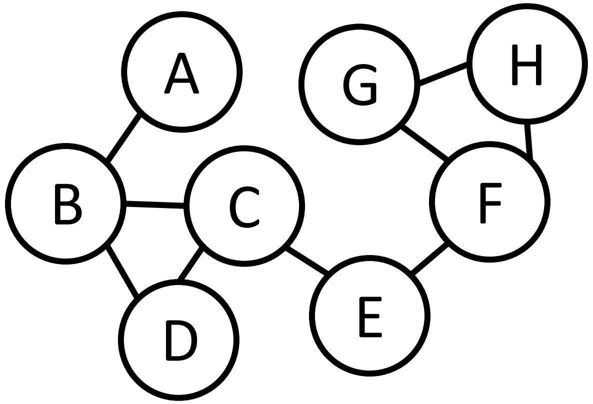


Node classification problem!



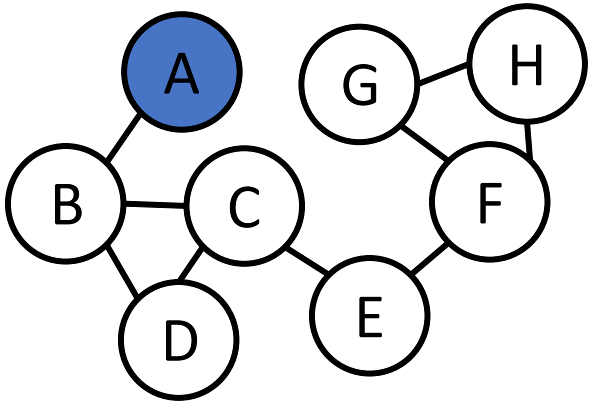
Graph-based Modeling

Collective Classification



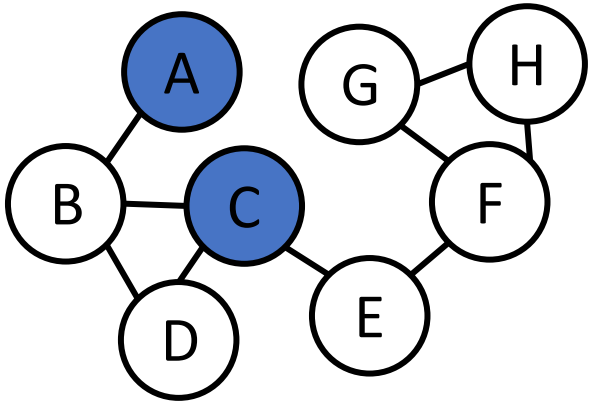
Graph structure

Collective Classification



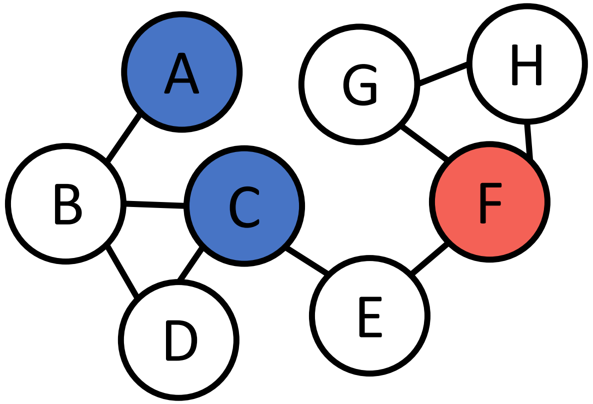
Graph structure

Collective Classification



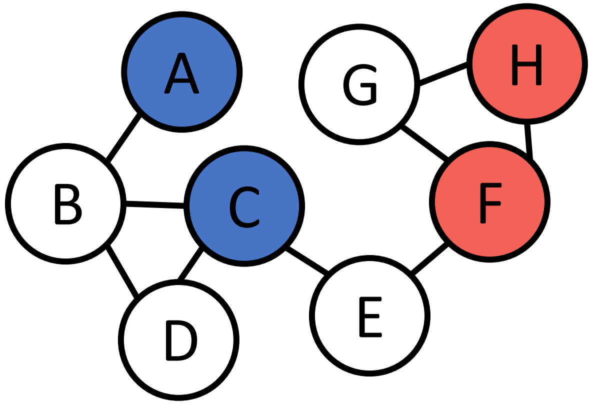
Graph structure

Collective Classification



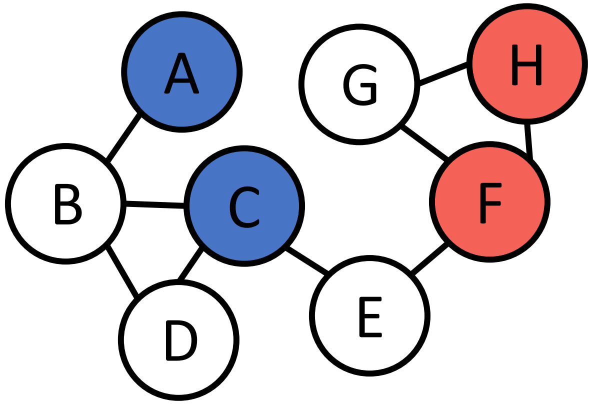
Graph structure

Collective Classification

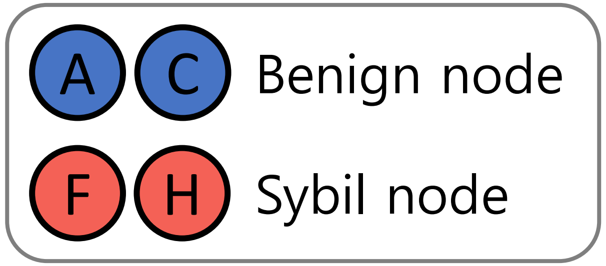


Graph structure

Collective Classification

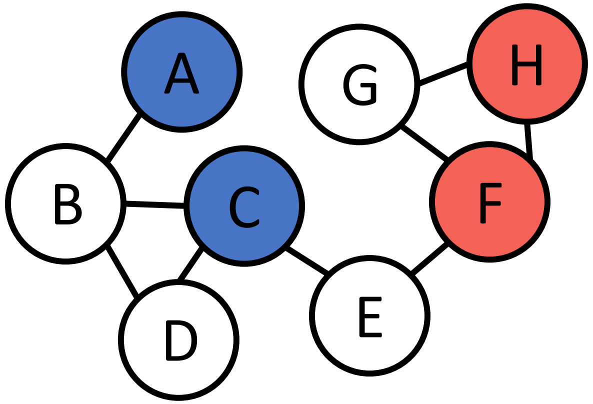


Graph structure

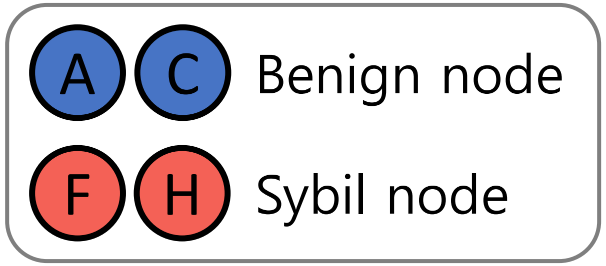


Training set

Collective Classification

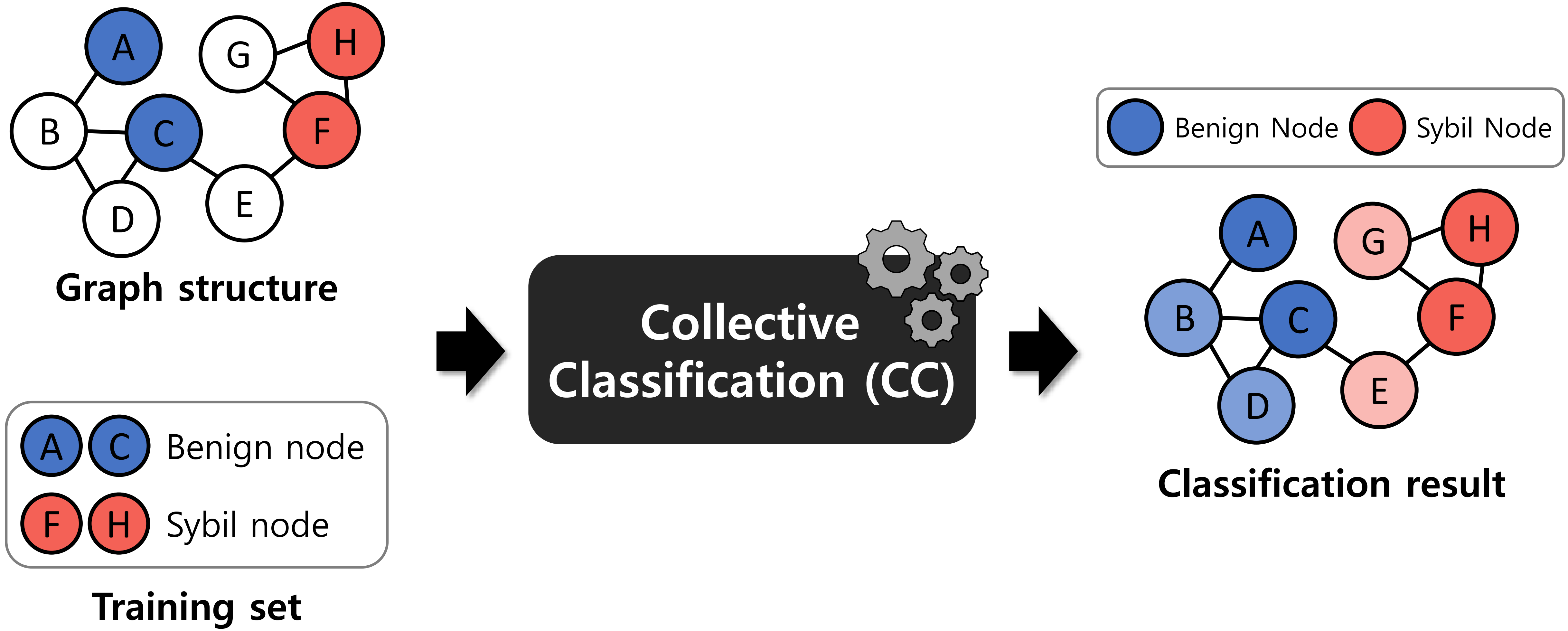


Graph structure

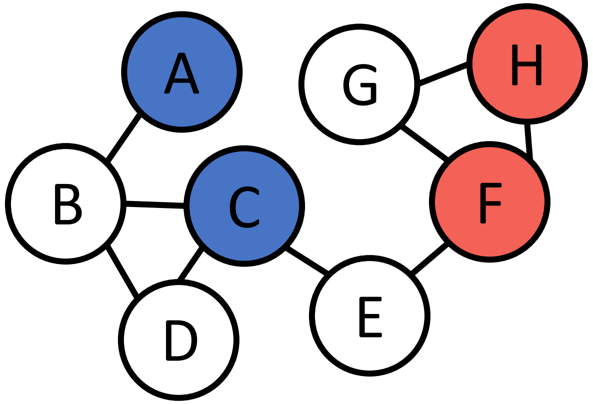


Training set

Collective Classification



Collective Classification



Graph structure

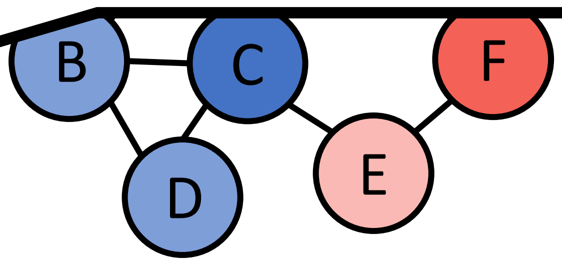
A C Benign node

F H Sybil node

Training set

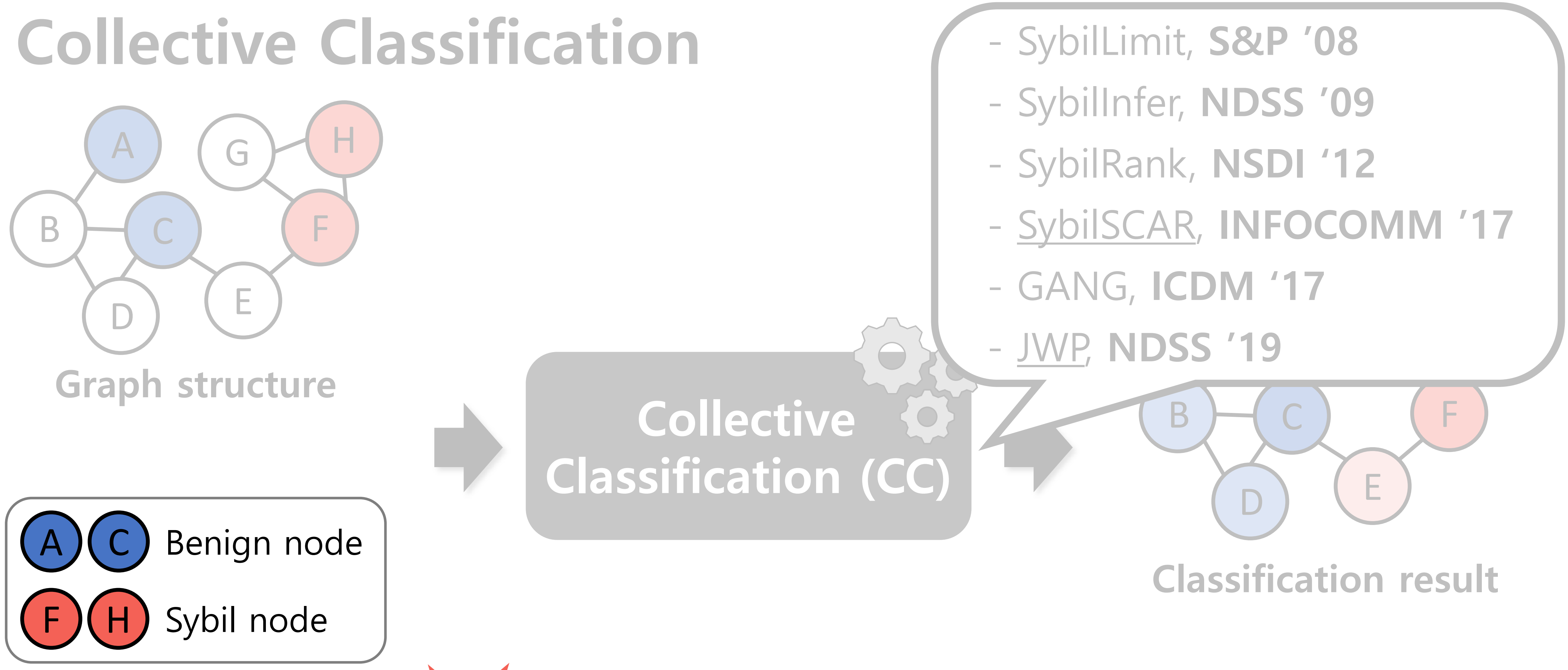
Collective Classification (CC)

- SybilLimit, **S&P '08**
- SybilInfer, **NDSS '09**
- SybilRank, **NSDI '12**
- SybilSCAR, **INFOCOMM '17**
- GANG, **ICDM '17**
- JWP, **NDSS '19**



Classification result

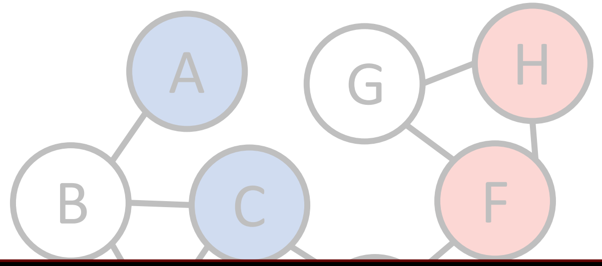
Collective Classification



Training set

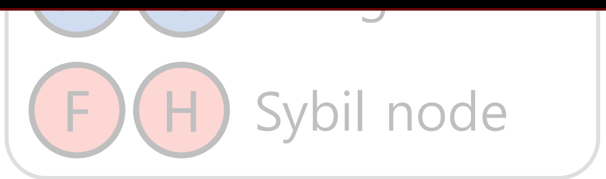
I know the training set!

Collective Classification



- SybilLimit, **S&P '08**
- SybilInfer, **NDSS '09**
- SybilRank, **NSDI '12**
- SybilSCAR, **INFOCOMM '17**

A strong adversary can bypass CC!

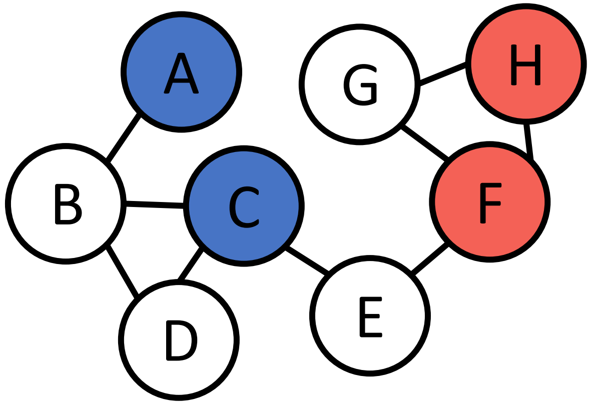


Classification result

Training set

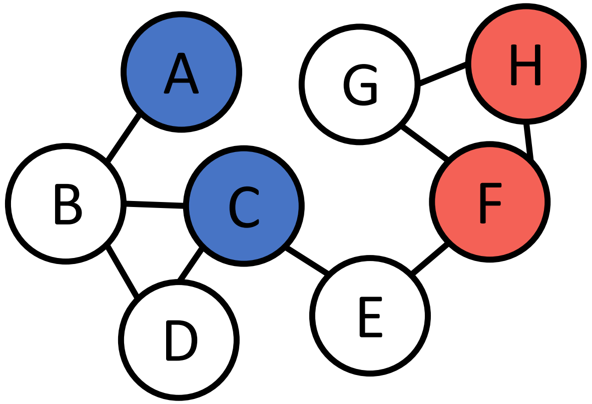
* Binghui Wang and Neil Zhenqiang Gong. Attacking Graph-based Classification via Manipulating the Graph Structure. CCS 2019
** Xu et al. Attacking Graph-Based Classification without Changing Existing Connections. ACSAC 2020

Adversarial Attacks against CC

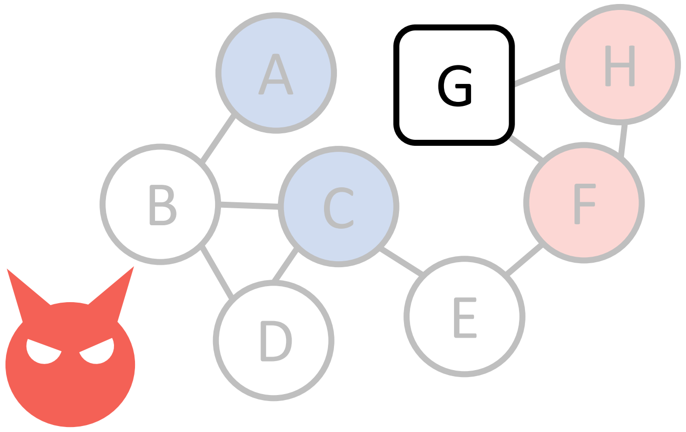


Original graph

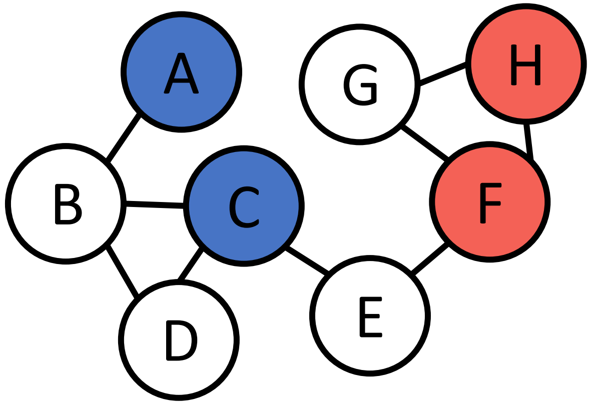
Adversarial Attacks against CC



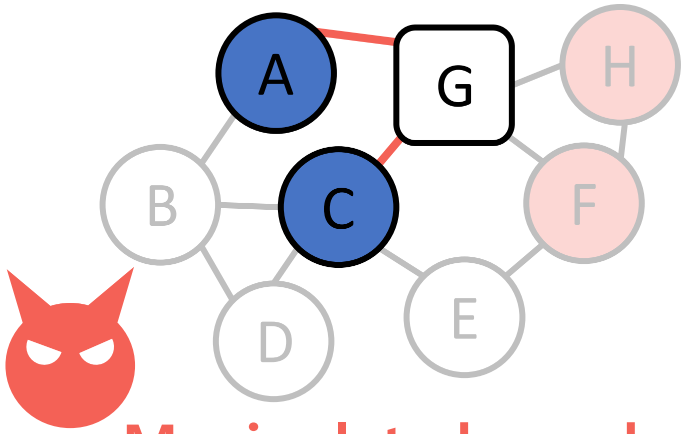
Original graph



Adversarial Attacks against CC

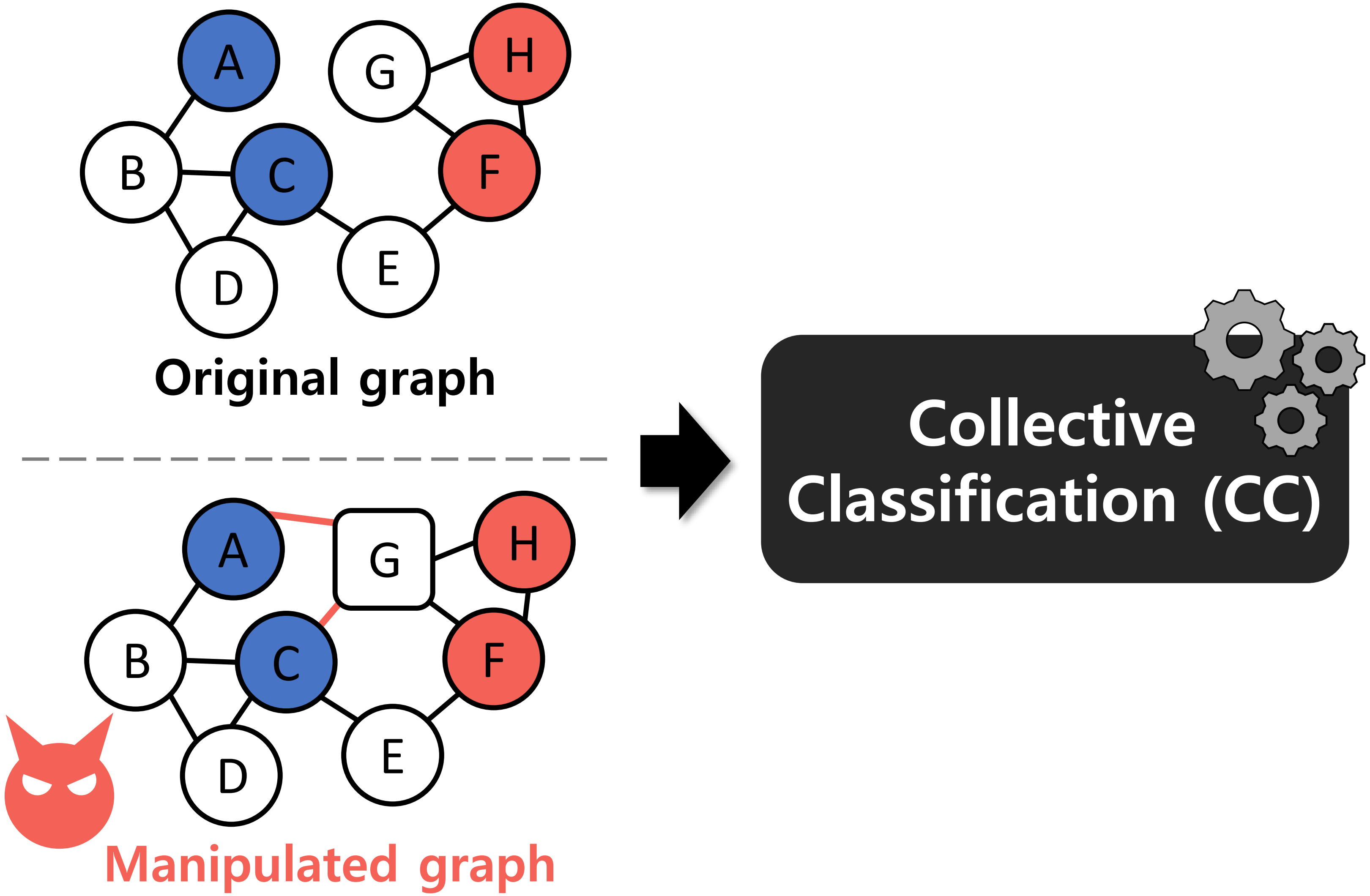


Original graph

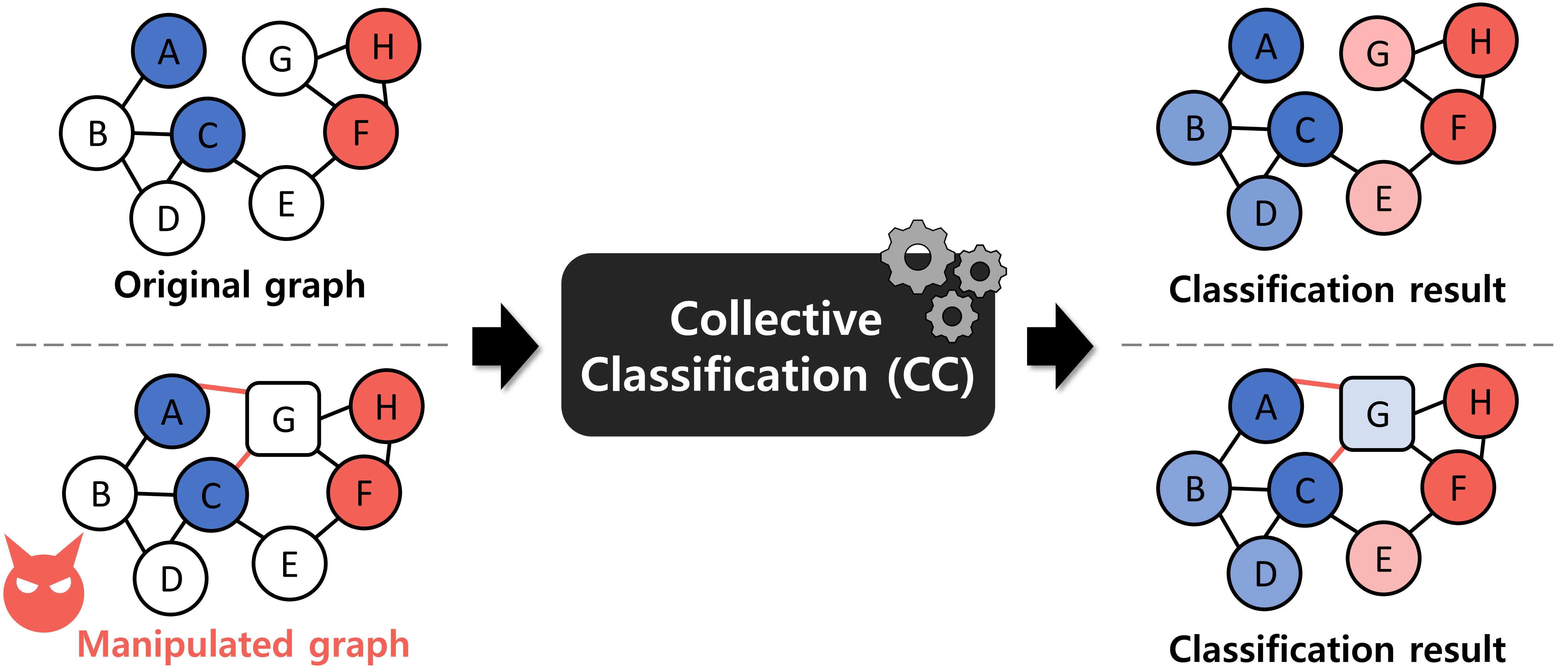


Manipulated graph

Adversarial Attacks against CC

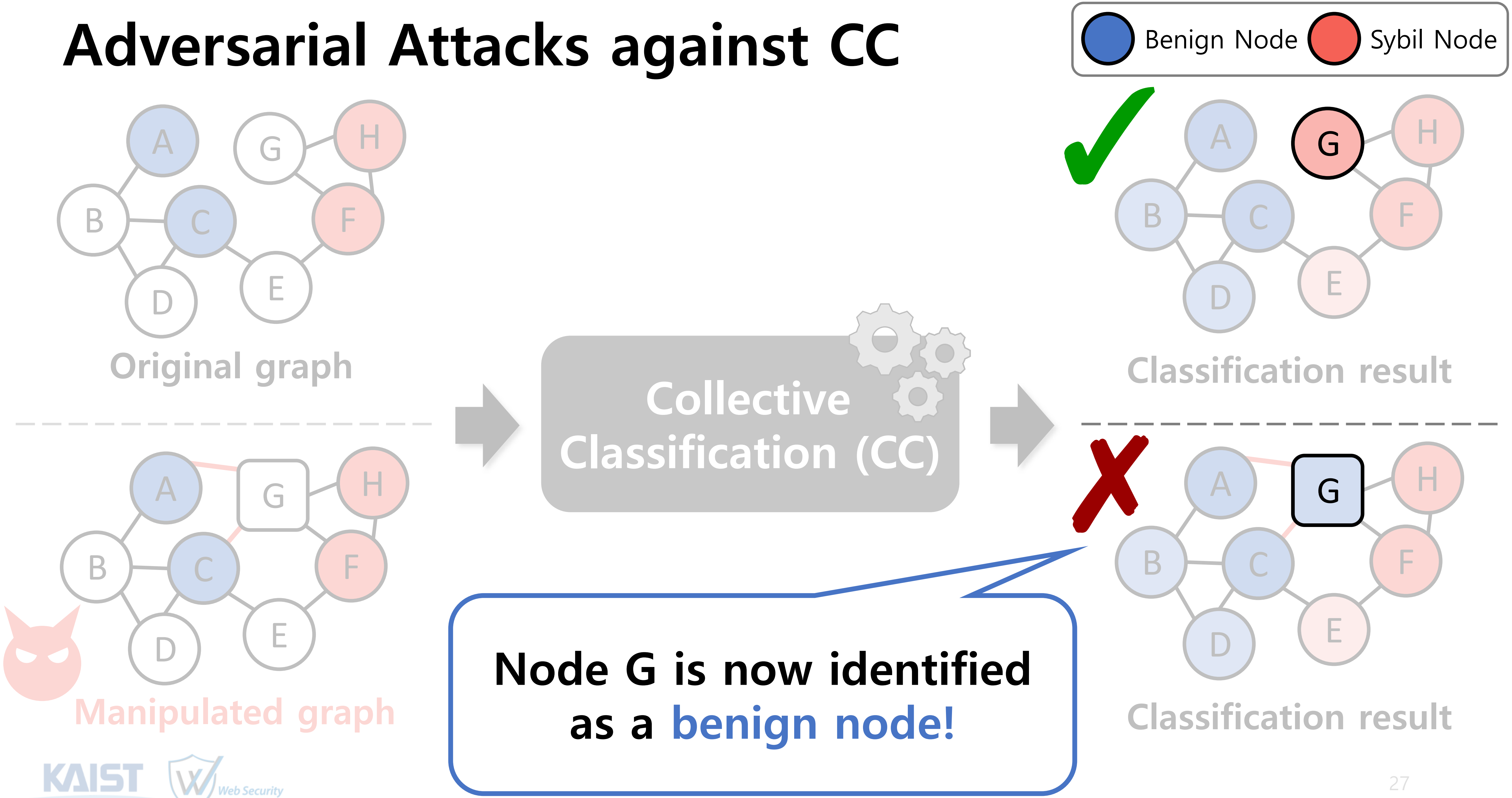


Adversarial Attacks against CC



KAIST Web Security & Privacy Lab

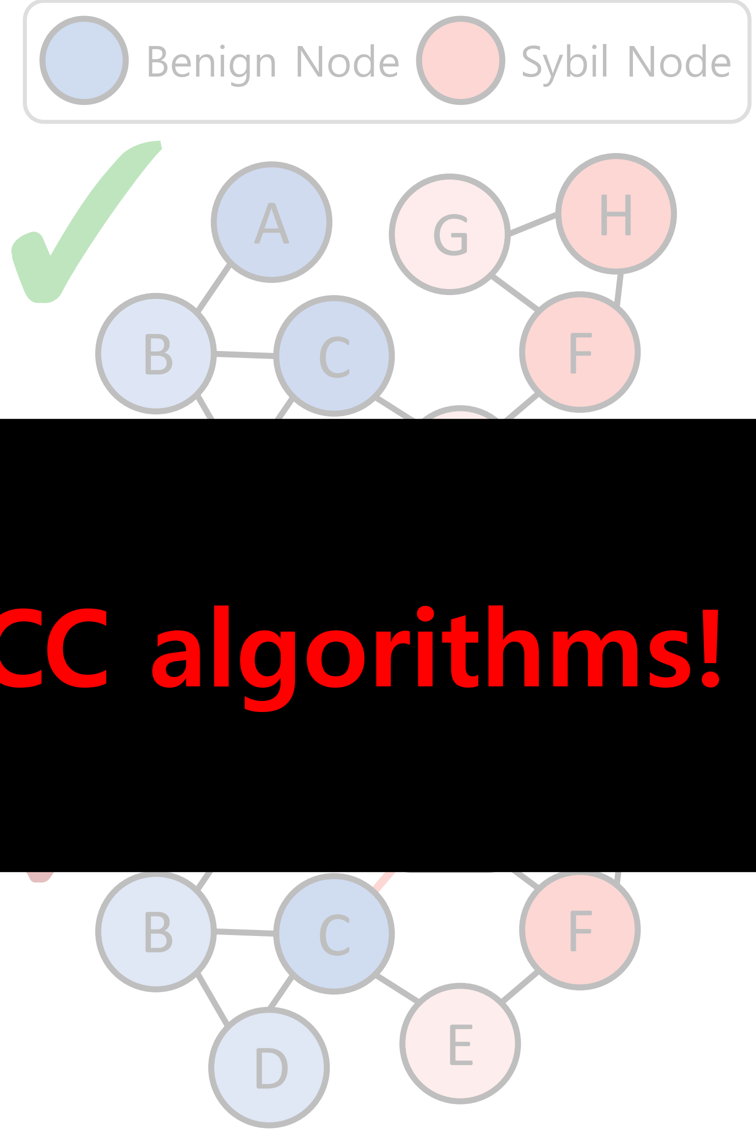
Adversarial Attacks against CC



KAIST Web Security & Privacy Lab

Adversarial Attacks against CC

Manipulated graph



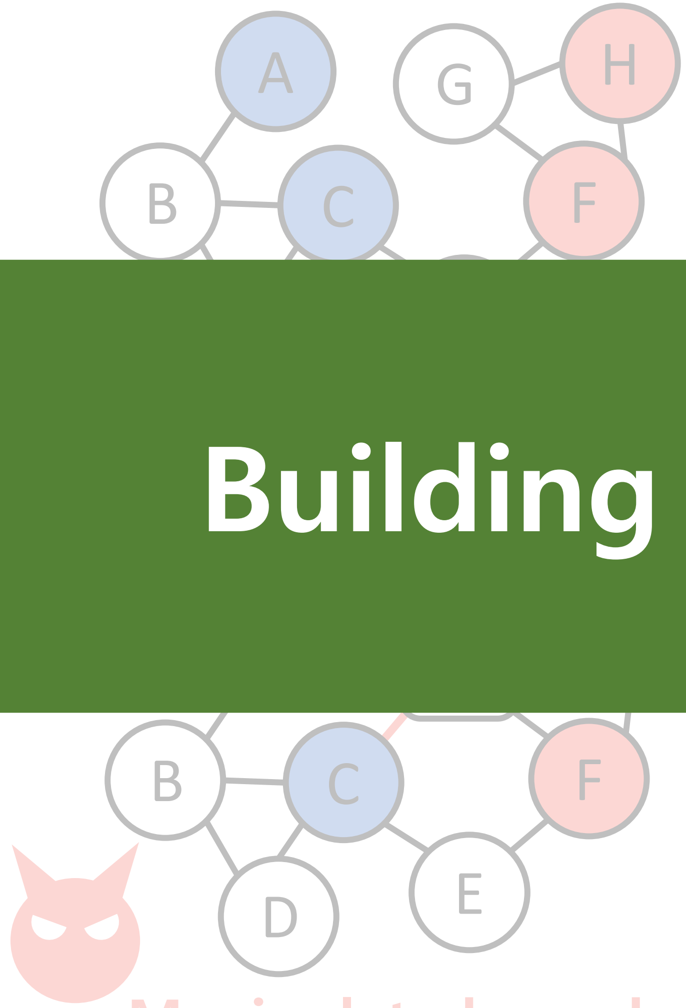
Classification result

These attacks destroyed existing CC algorithms!

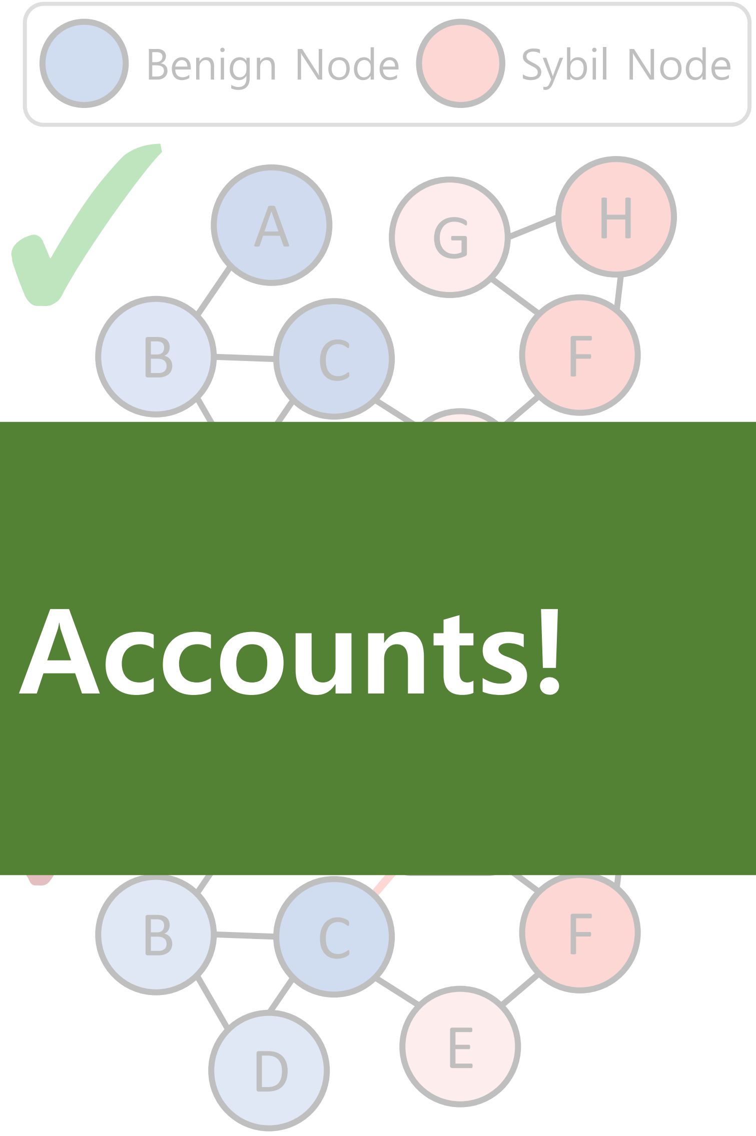
* Gong et al. SybilBelief: A Semi-supervised Learning Approach for Structure-based Sybil Detection. IEEE TIFS 2014
** Wang et al. SybilSCAR: Sybil Detection in Online Social Networks via Local Rule based Propagation. INFOCOMM 2017
*** Wang et al. Structure-based Sybil Detection in Social Networks via Local Rule-based Propagation. IEEE TNSE 2018.
**** Wang et al. Graph-based Security and Privacy Analytics via Collective Classification with Joint Weight Learning and Propagation. NDSS 2019

Our Goal

Building Robust CC of Sybil Accounts!



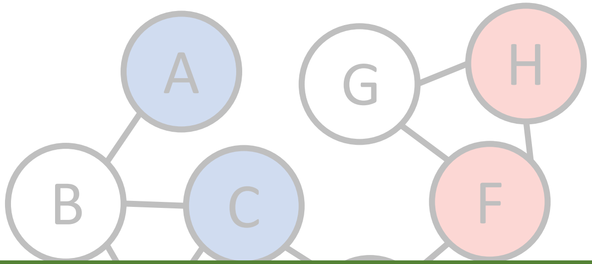
Manipulated graph

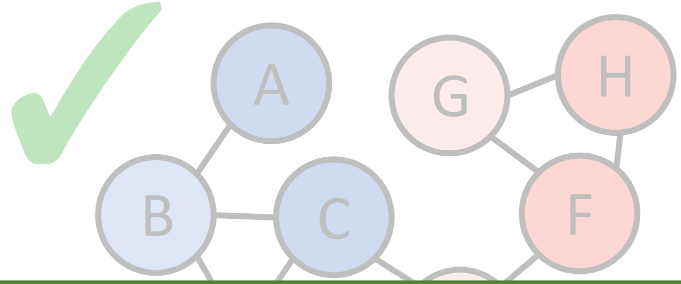


Classification result

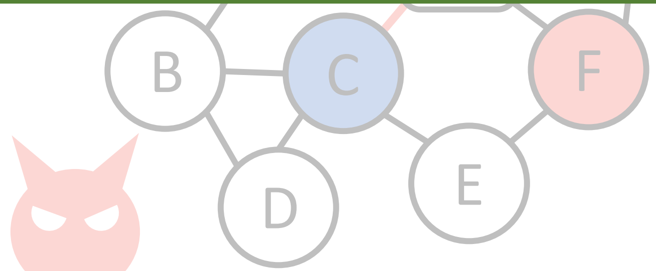
Our Goal

Benign Node Sybil Node

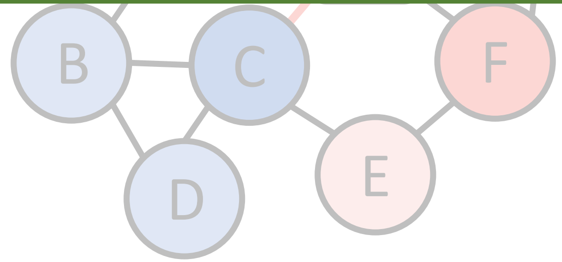




We propose **RICC**!



Manipulated graph



Classification result

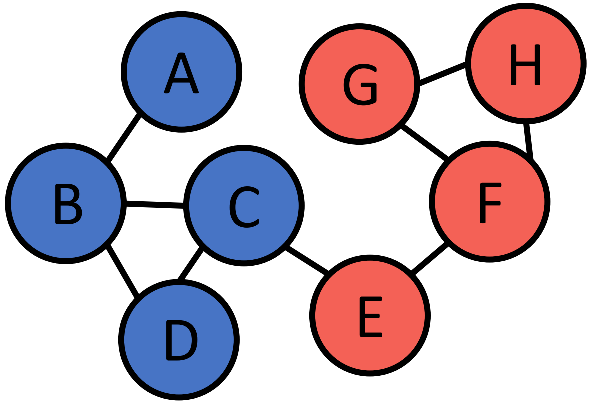
Our Observation on the Manipulated Graphs

Our Observation on the Manipulated Graphs

- **To which node** does the adversary connect adversarial edges?

Our Observation on the Manipulated Graphs

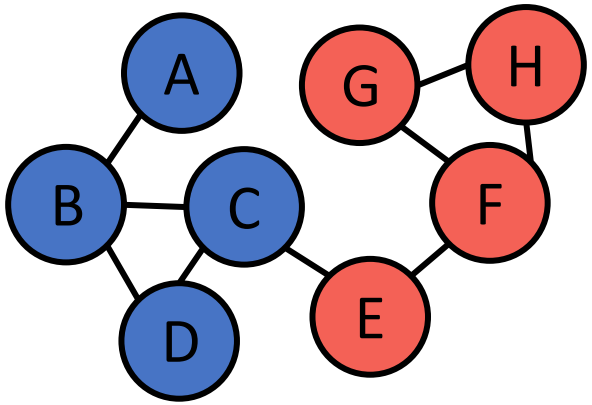
- **<u>To which node</u>** does the adversary connect adversarial edges?



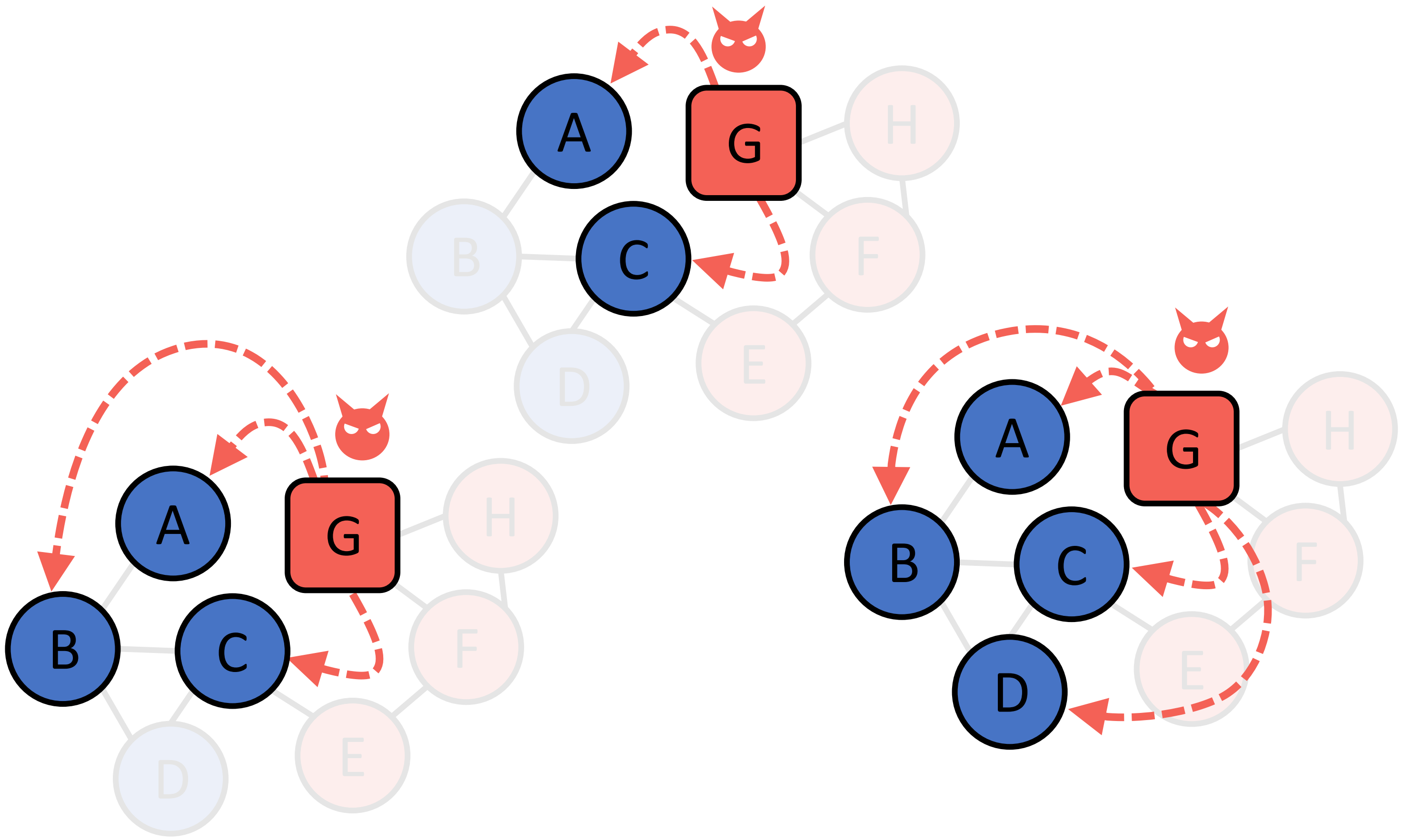
Original graph

Our Observation on the Manipulated Graphs

- **<u>To which node</u>** does the adversary connect adversarial edges?



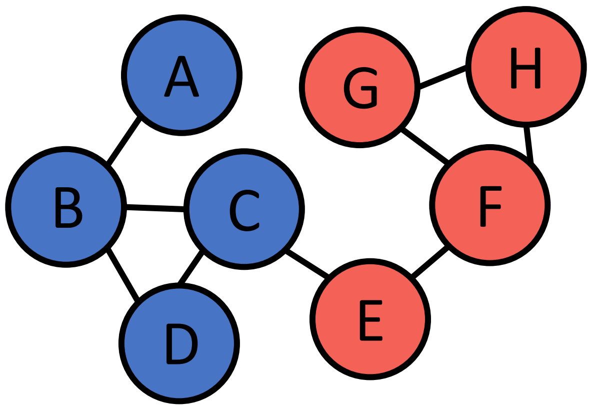
Original graph



Possible graph manipulations

Our Observation on the Manipulated Graphs

- **<u>To which node</u>** does the adversary connect adversarial edges?



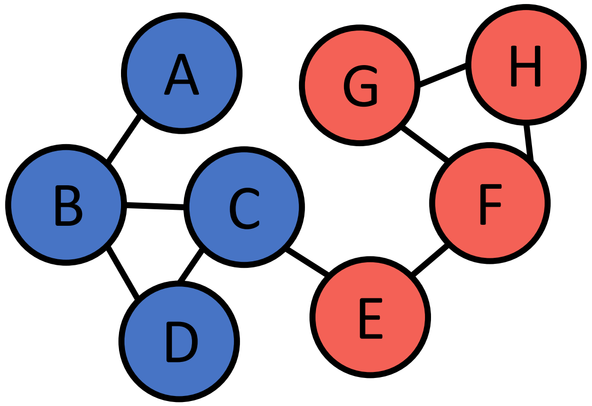
Original graph

Our observation

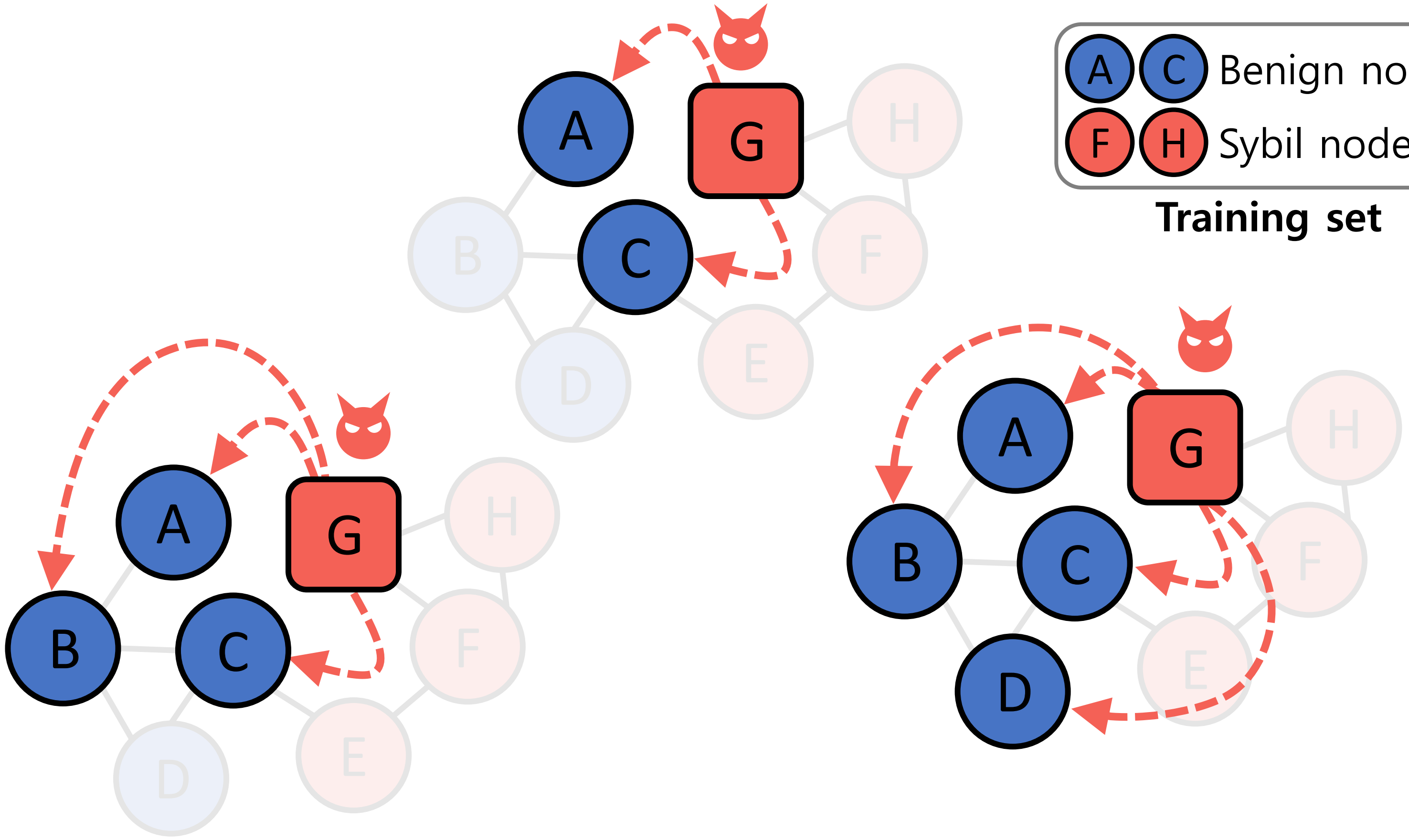
Adversarial edges are connected to **<u>benign</u>** nodes in a **<u>training set</u>**!

Our Observation on the Manipulated Graphs

- **<u>To which node</u>** does the adversary connect adversarial edges?



Original graph

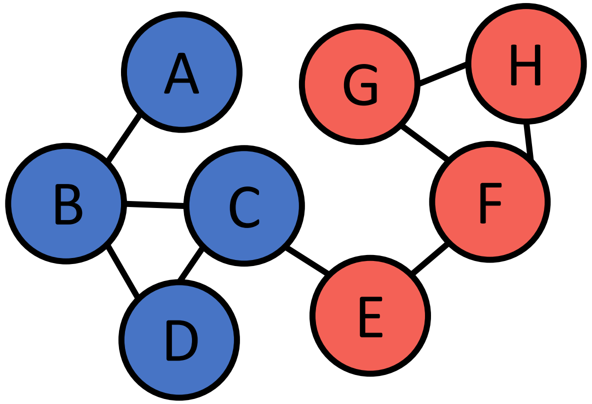


Training set

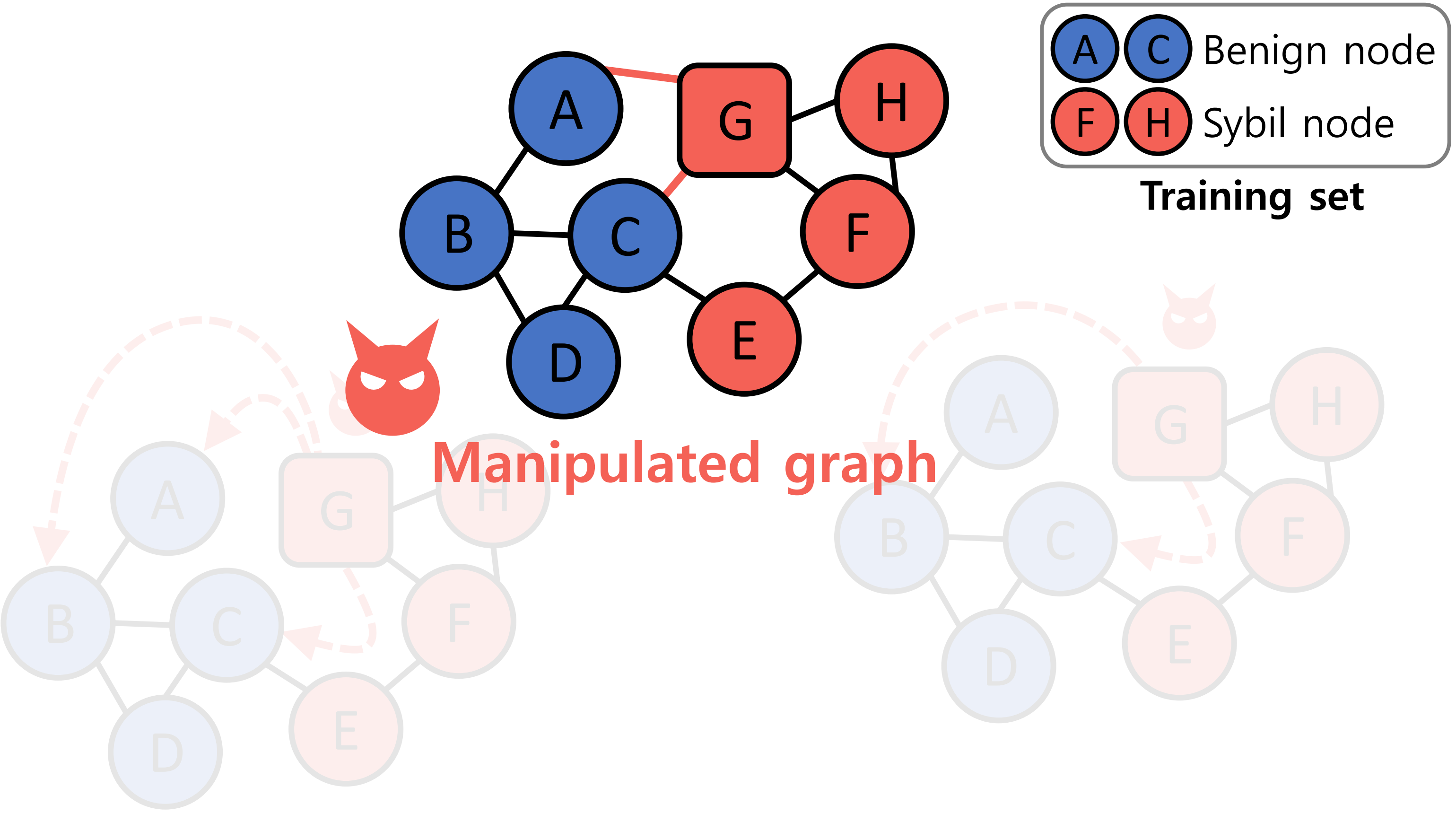
Possible graph manipulations

Our Observation on the Manipulated Graphs

- **To which node** does the adversary connect adversarial edges?



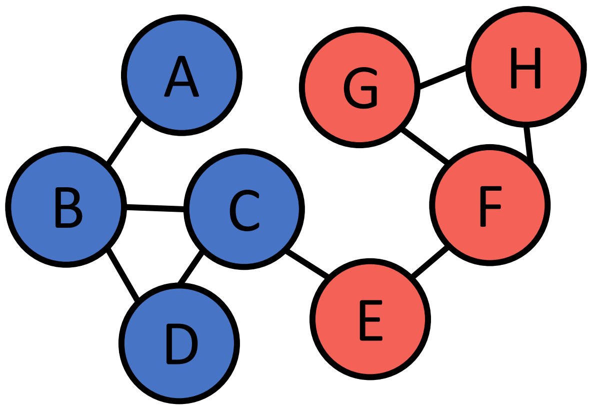
Original graph



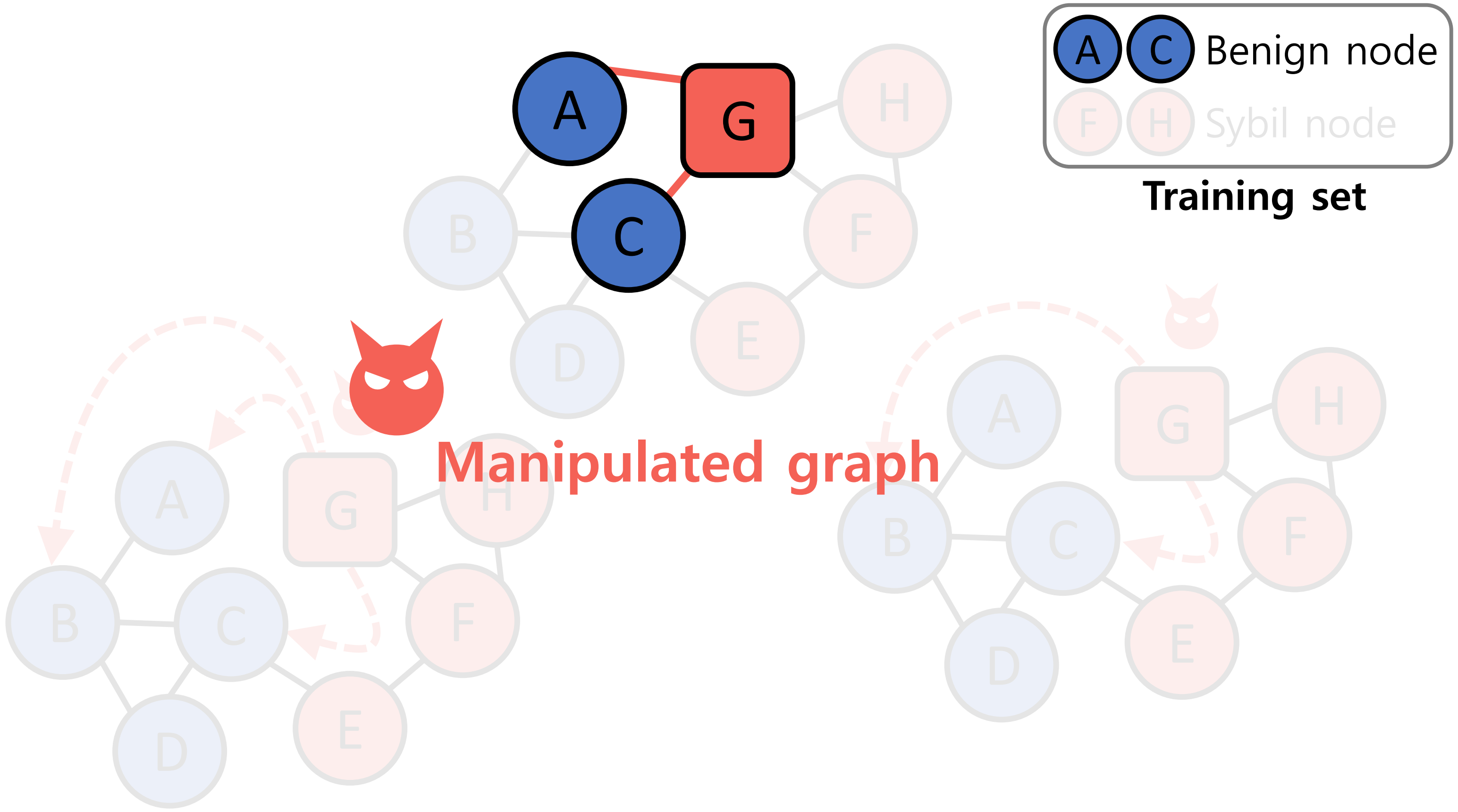
Possible graph manipulations

Our Observation on the Manipulated Graphs

- **<u>To which node</u>** does the adversary connect adversarial edges?



Original graph



Possible graph manipulations

Our Observation on the Manipulated Graphs

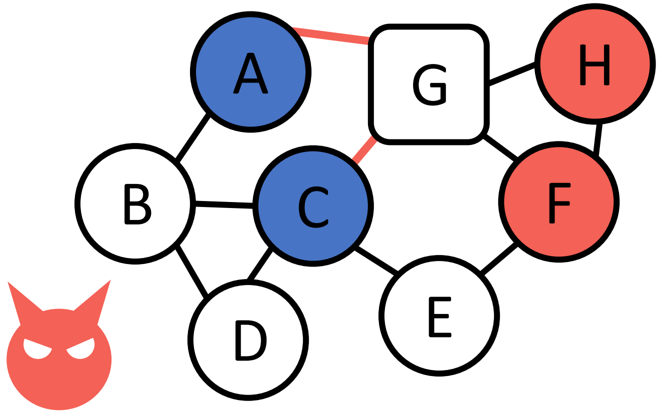
- **To which node** does the adversary connect adversarial edges?

Original graph

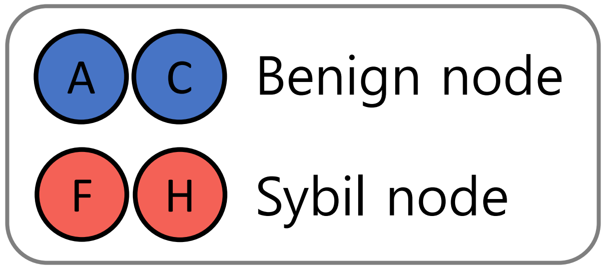
Possible graph manipulations

These attacks are ***tailored*** to the ***training set!***

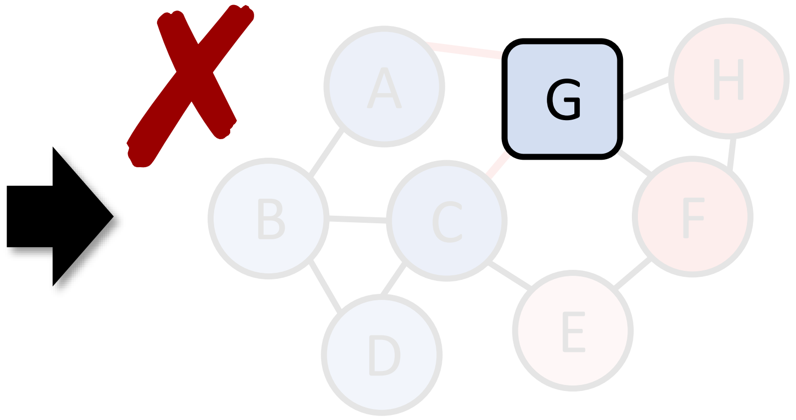
Our Motivation



Manipulated graph

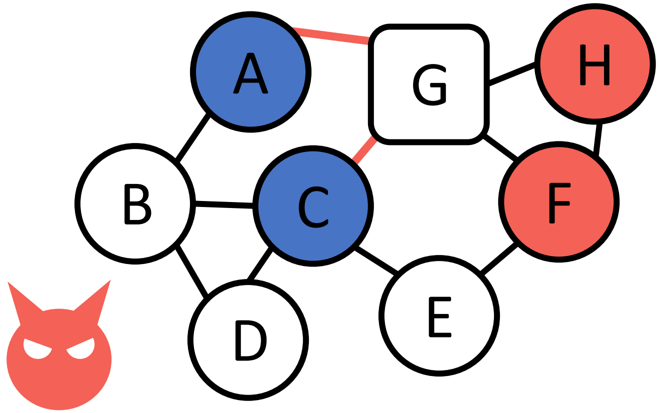


Original training set

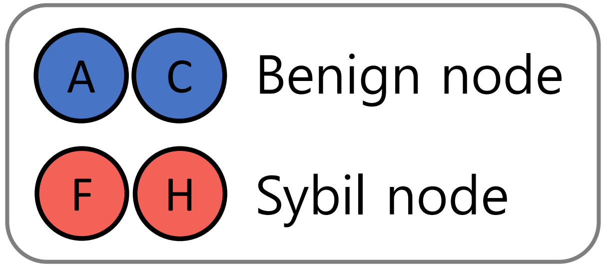


Classification result

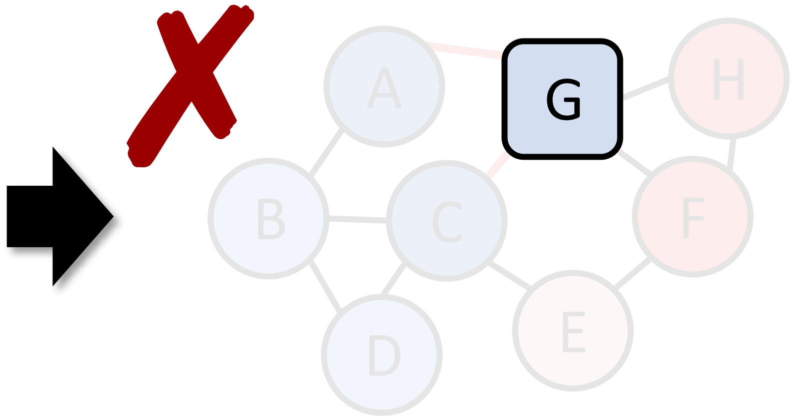
Our Motivation



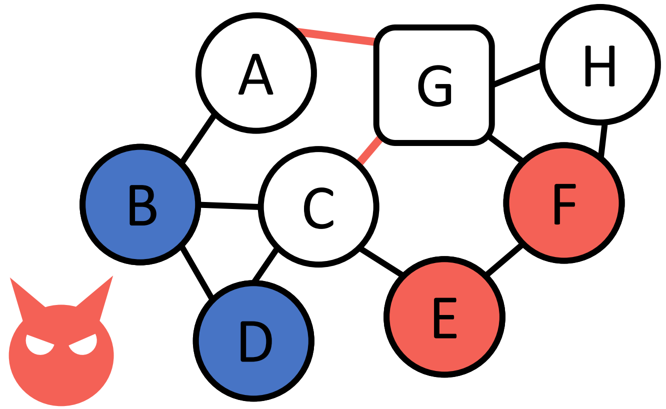
Manipulated graph



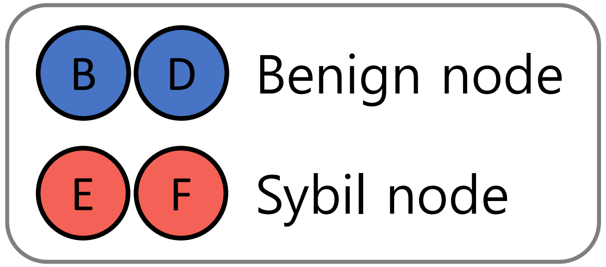
Original training set



Classification result



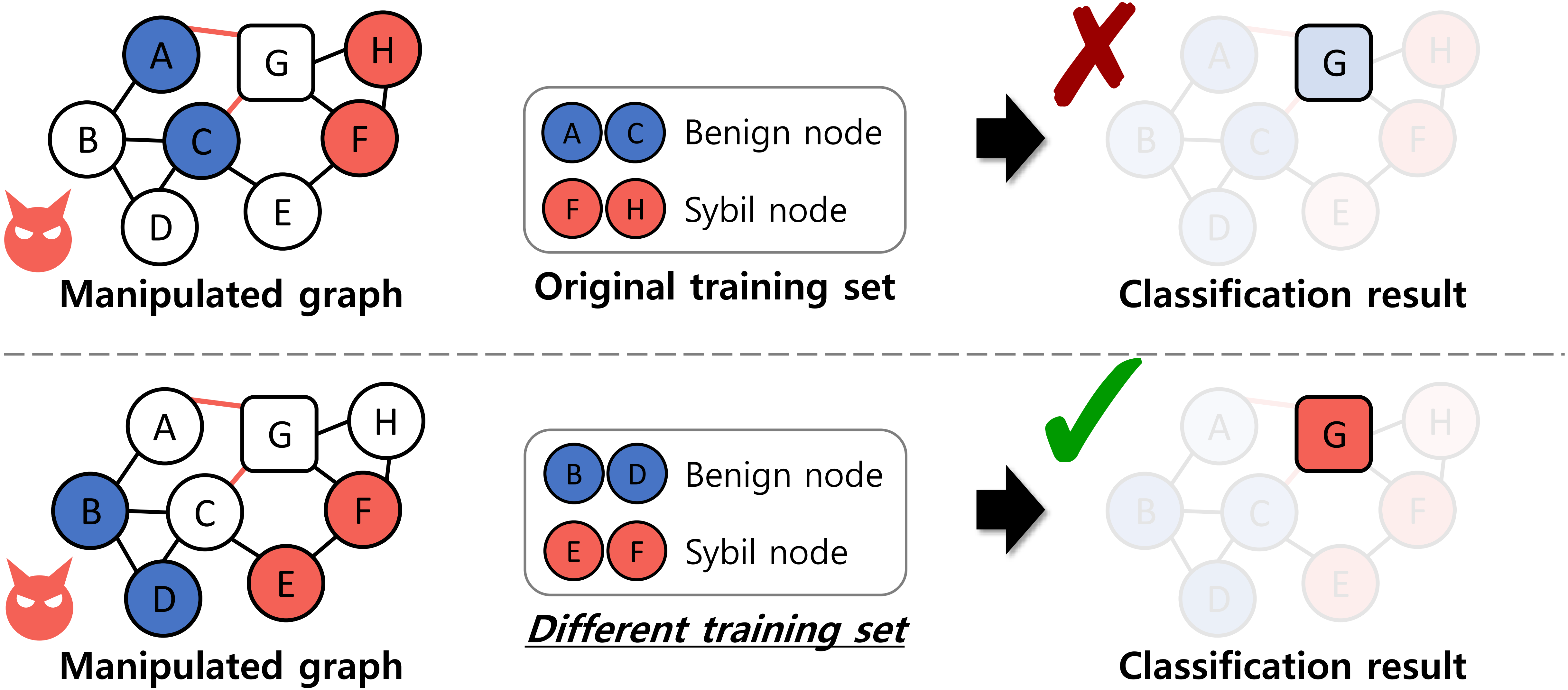
Manipulated graph



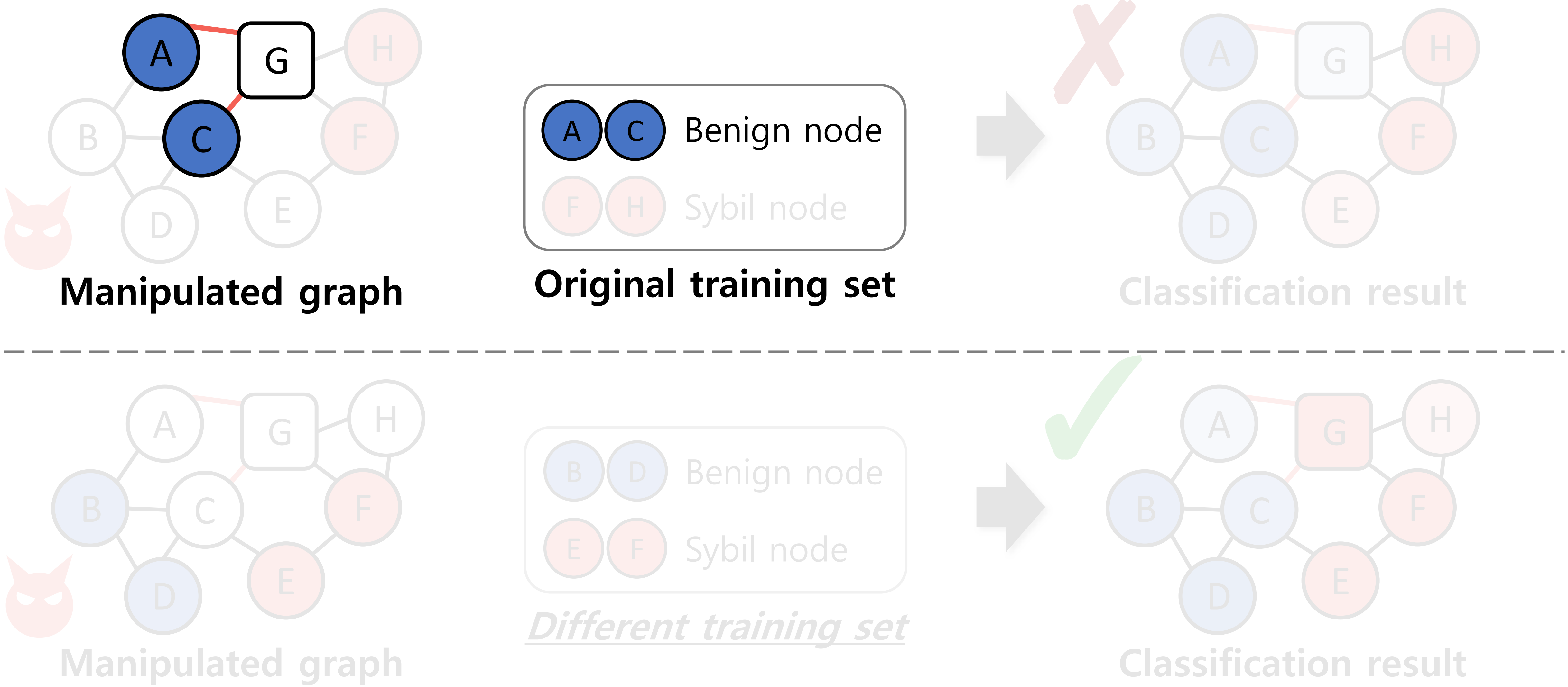
<u>Different training set</u>

KAIST Web Security & Privacy Lab

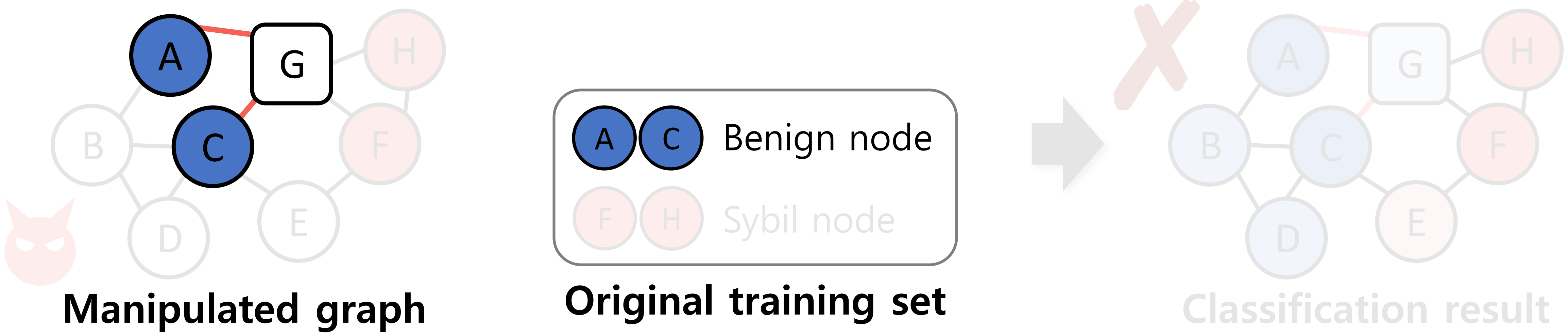
Our Motivation

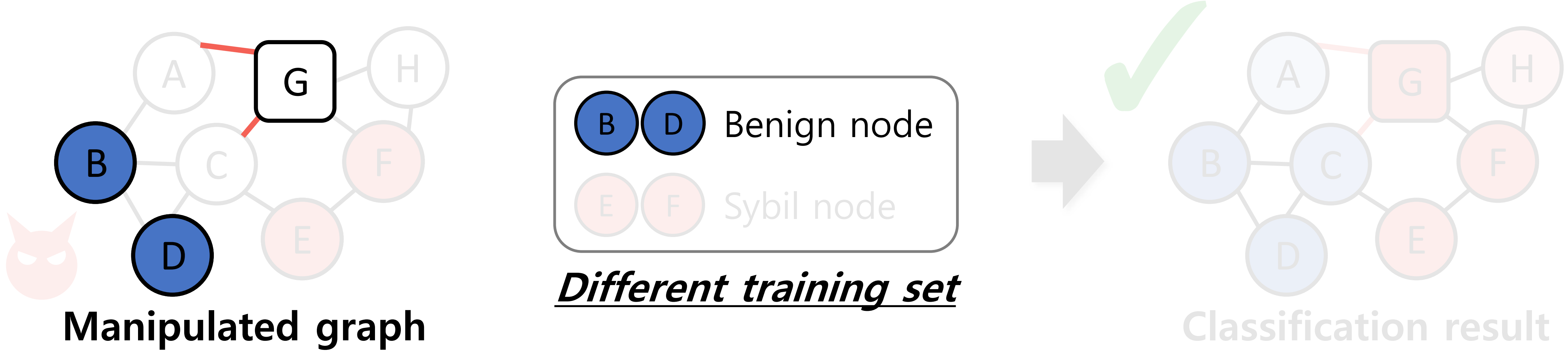


Our Motivation

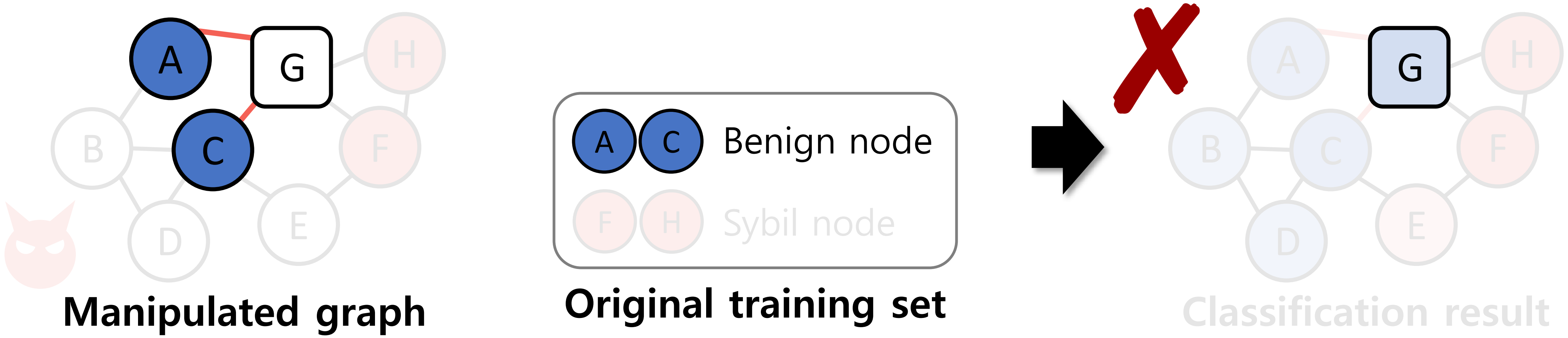


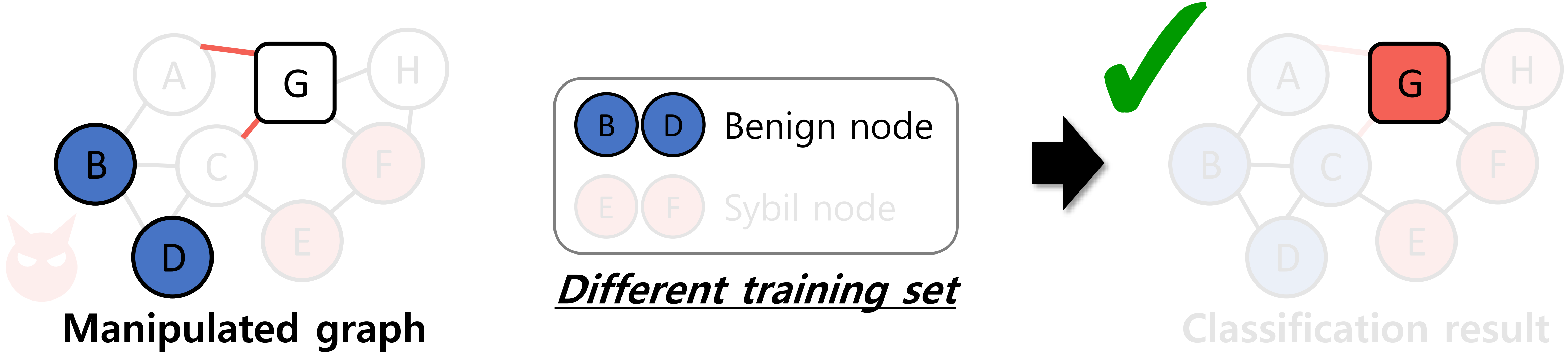
Our Motivation





Our Motivation





Our Motivation

Manipulated graph

A C Benign node

B D Benign node

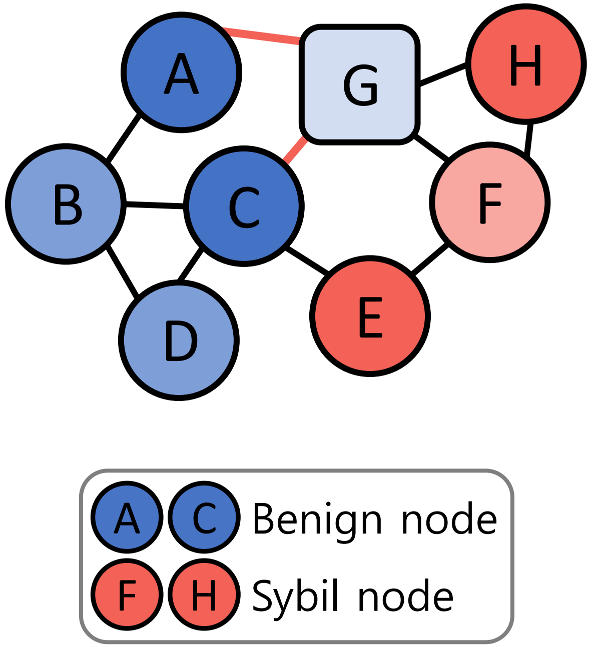
E F Sybil node

Different training set

Classification result

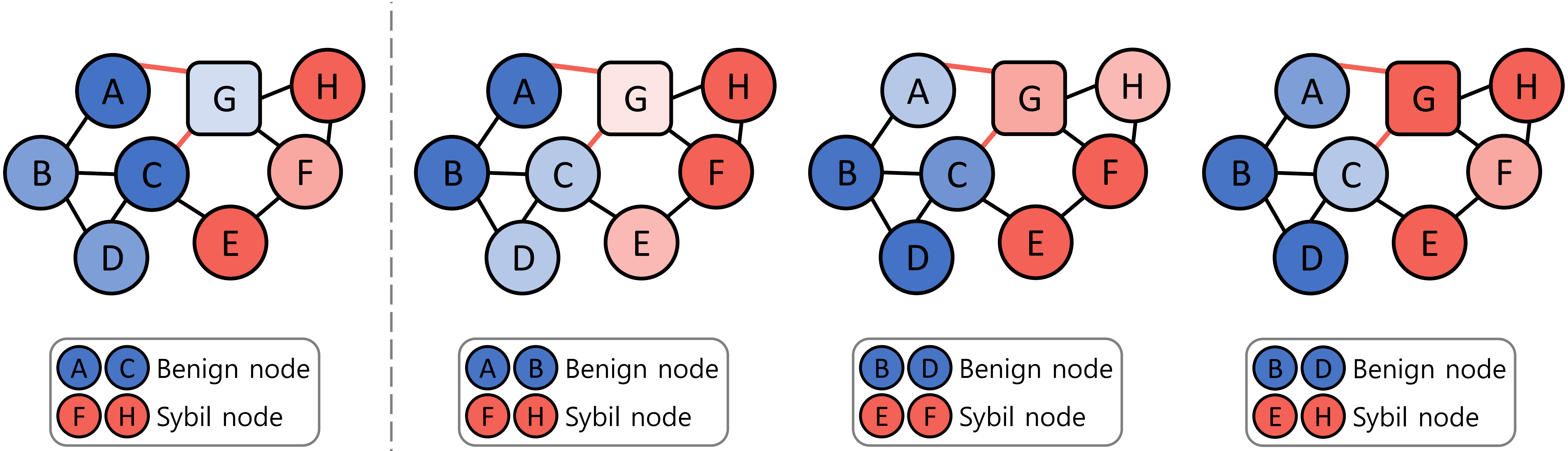
Different training set → **Reliable** classification!

Towards Reliable Classification Results



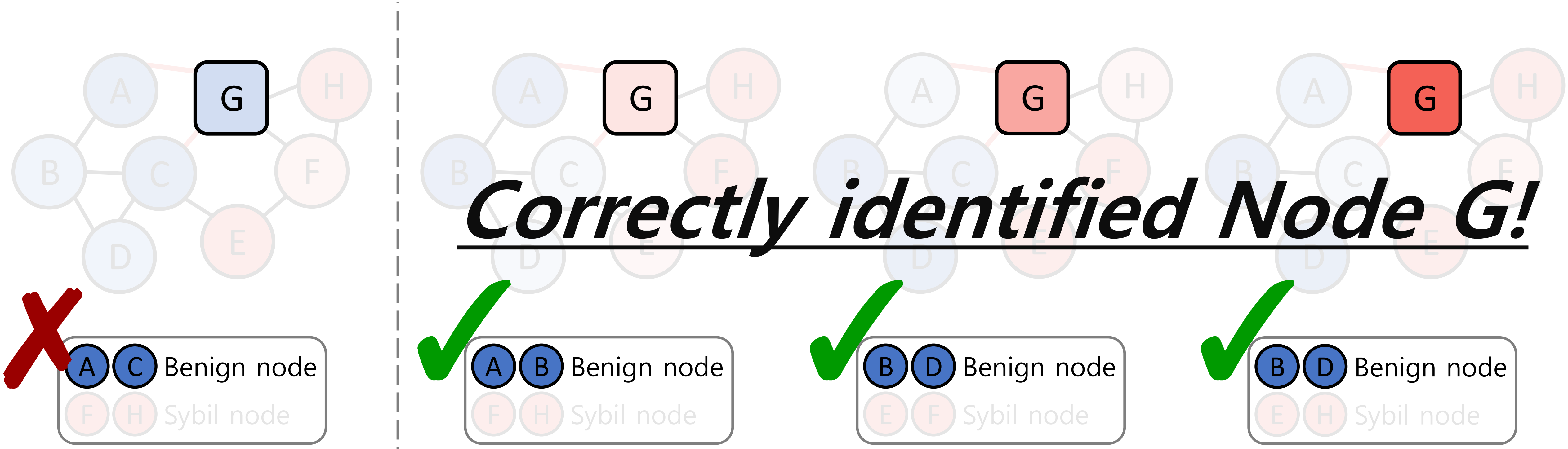
Original training set

Towards Reliable Classification Results



Original training set | **Randomly sampled training sets**

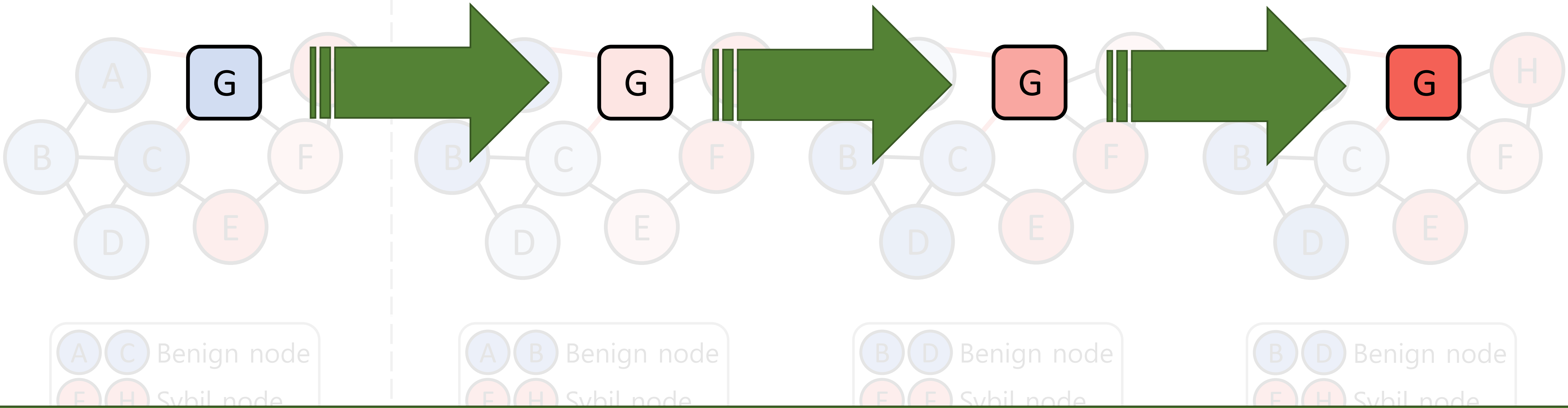
Towards Reliable Classification Results



Original training set

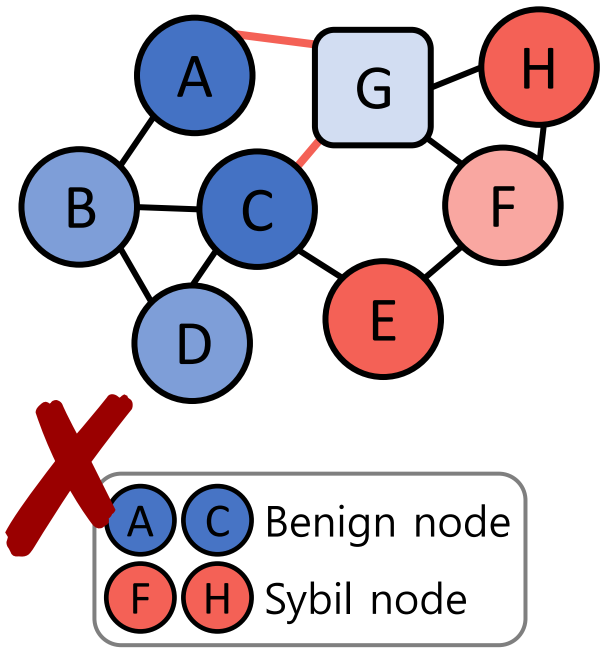
Randomly sampled training sets

Towards Reliable Classification Results



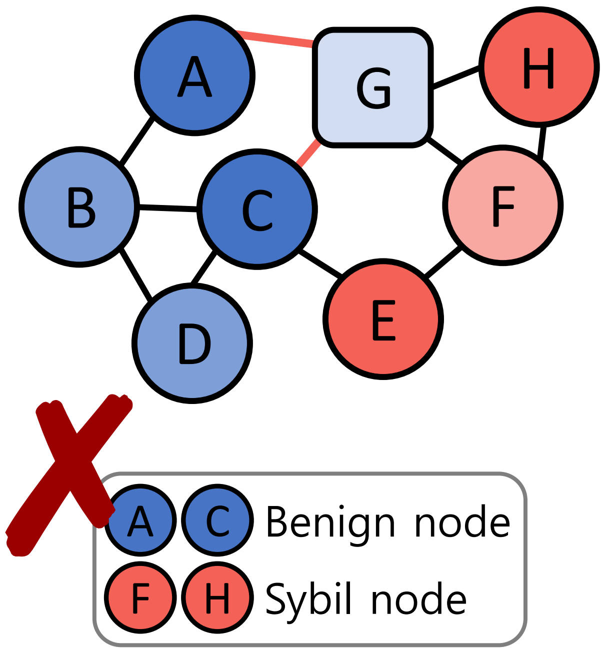
RICC ***gradually guides*** CC to output ***reliable results***!

Random Sampling-based Collective Classification

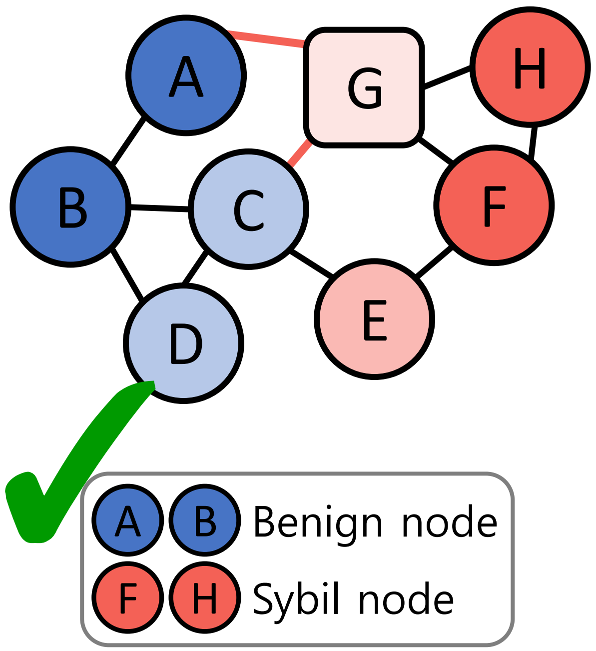


Original training set

Random Sampling-based Collective Classification

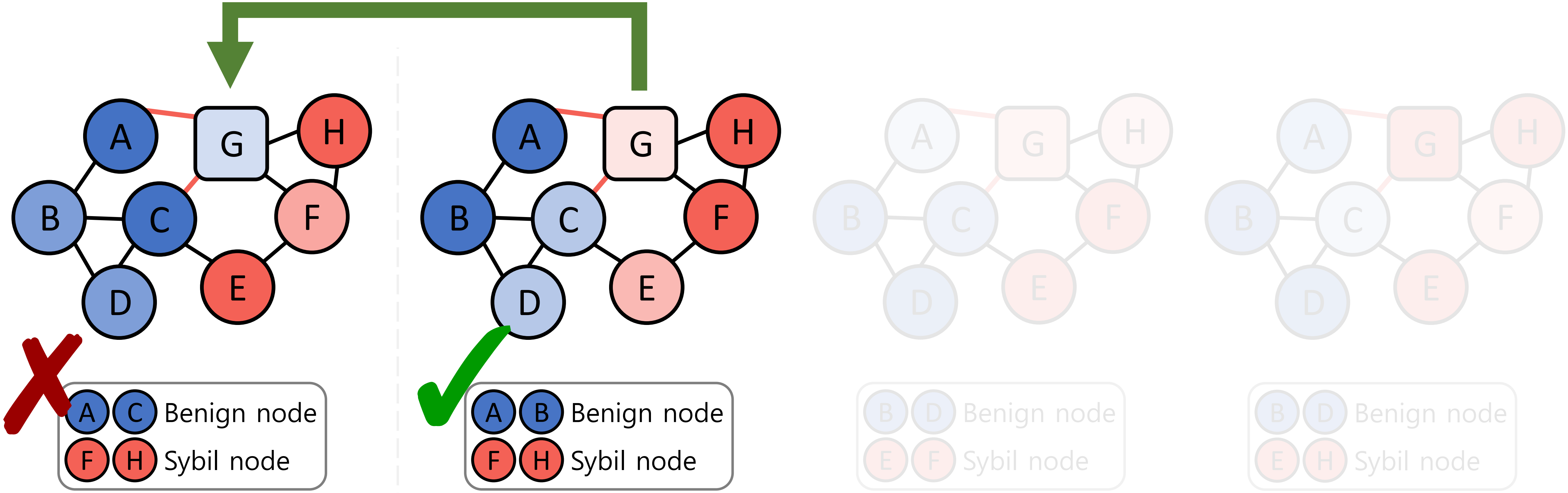


Original training set



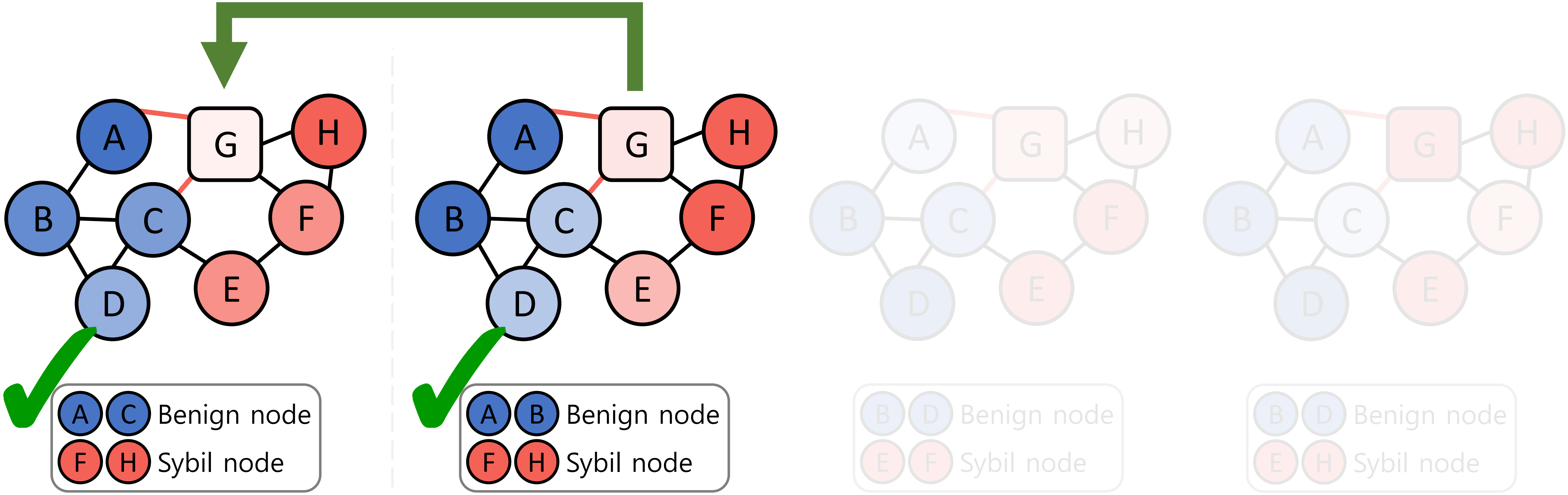
Sampled set I

Random Sampling-based Collective Classification



Original training set | **Sampled set I**

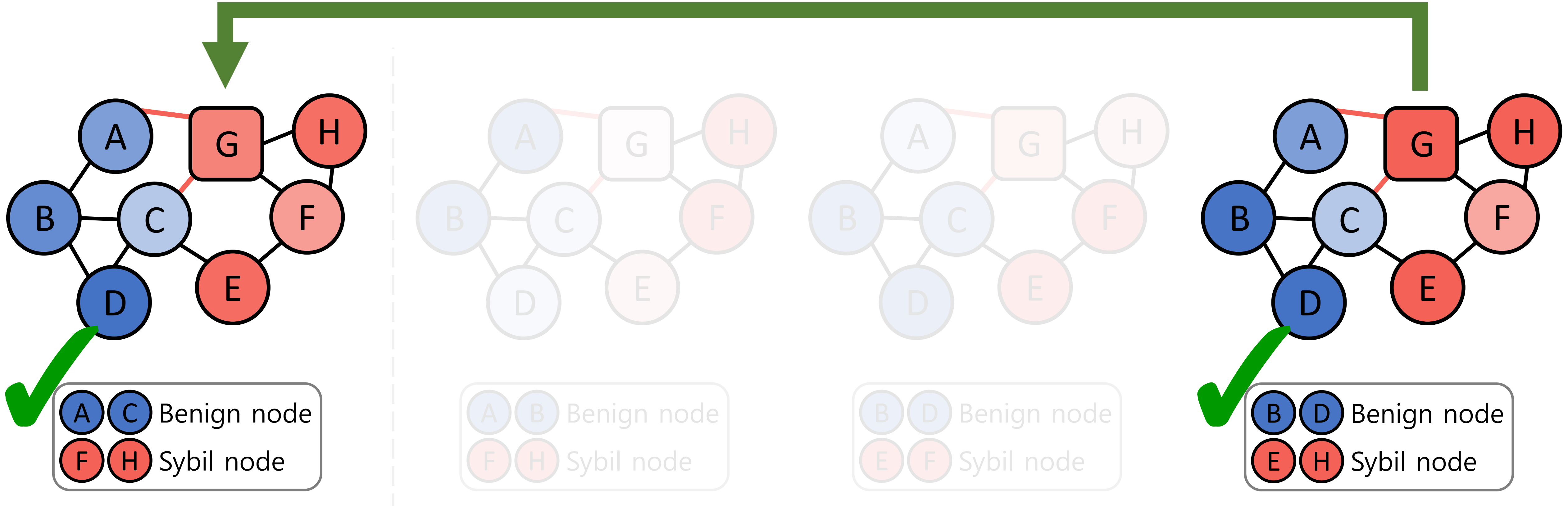
Random Sampling-based Collective Classification



Original training set

Sampled set I

Random Sampling-based Collective Classification



Original training set **Sampled set I** **•••** **Sampled set *N***

Evaluation

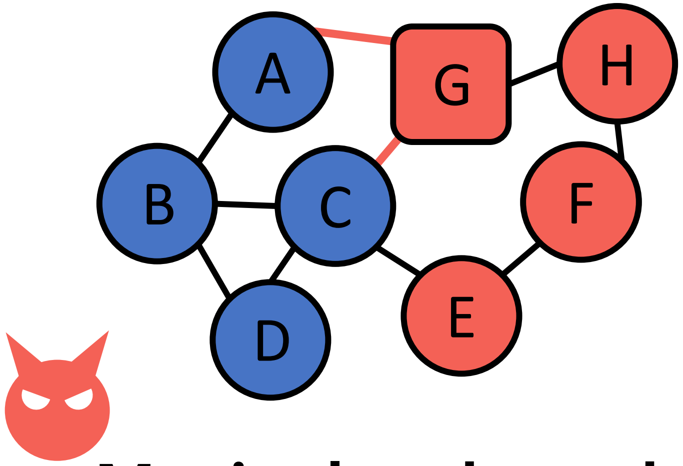
Datasets

- Four datasets: Enron, Facebook, Twitter_S, and Twitter_L

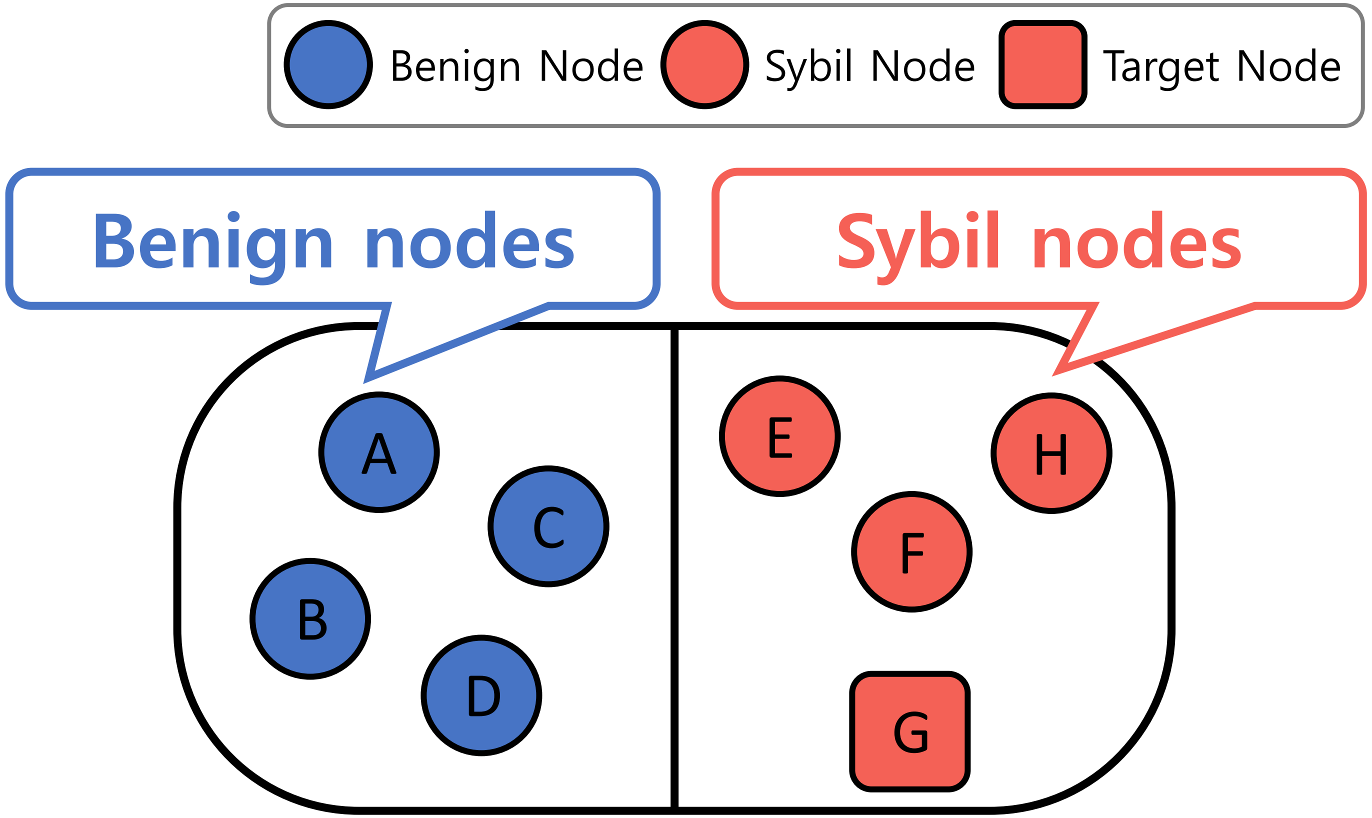
Dataset	Enron	Facebook	Twitter_S	Twitter_L
# of nodes	67K+	8K+	8K+	21M+
# of edges	371K+	176K+	54K+	265M+
Node degree	11	44	13	25

These graphs cover diverse scenarios!

Attack Scenario



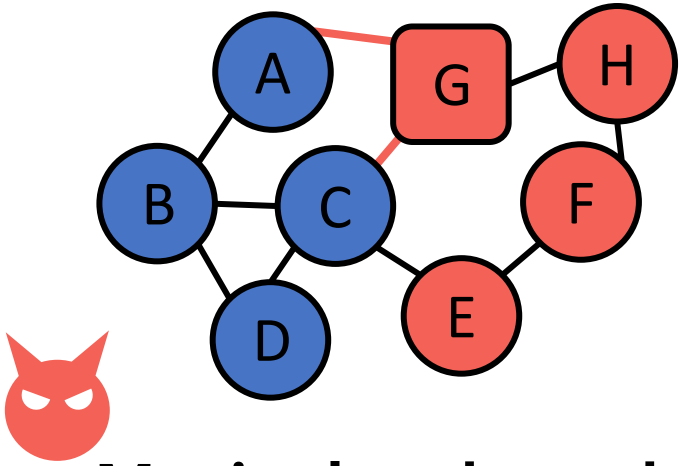
Manipulated graph



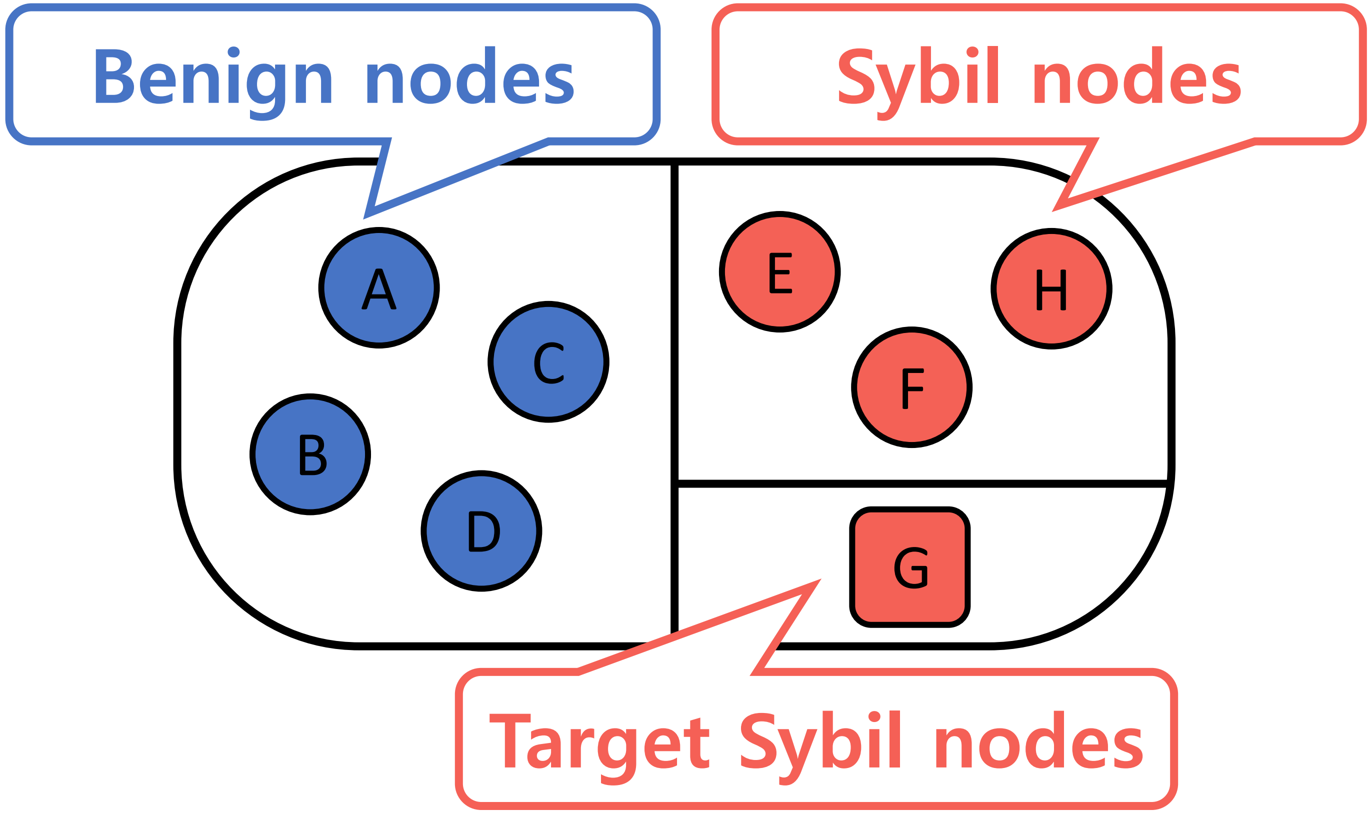
Node set V

Attack Scenario

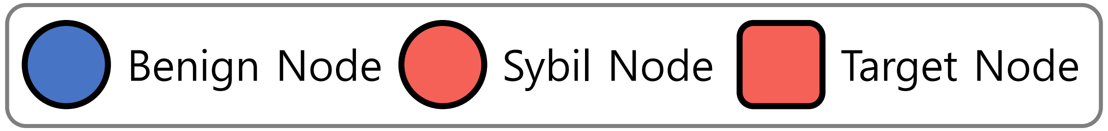
Benign Node Sybil Node Target Node

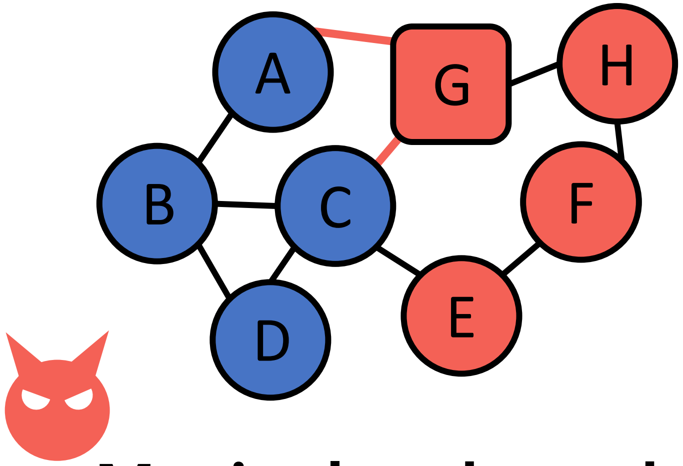


Manipulated graph

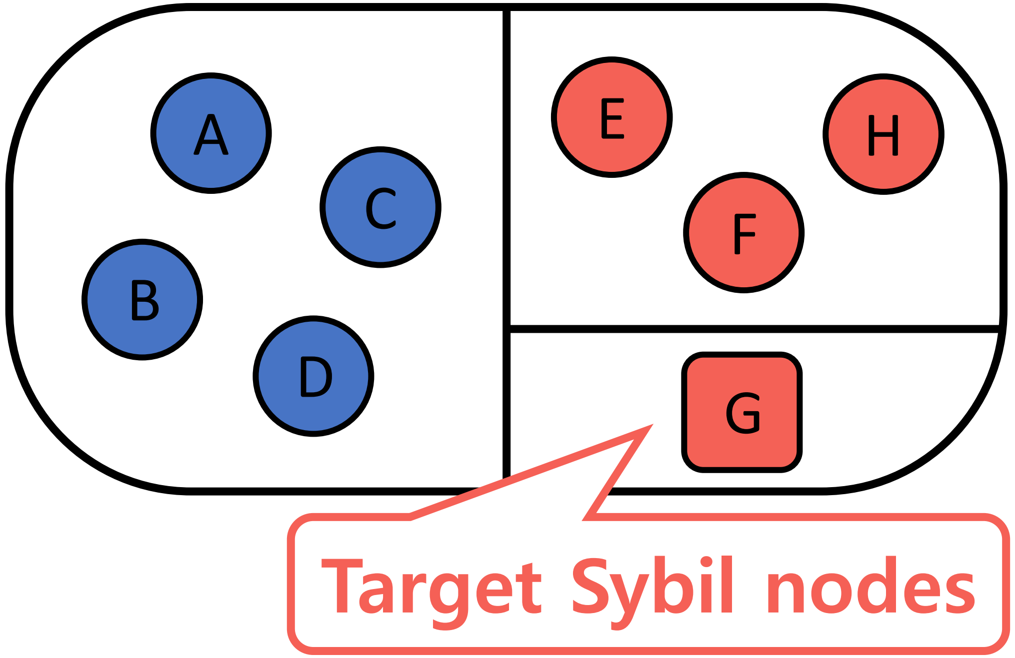


Evaluation Metrics



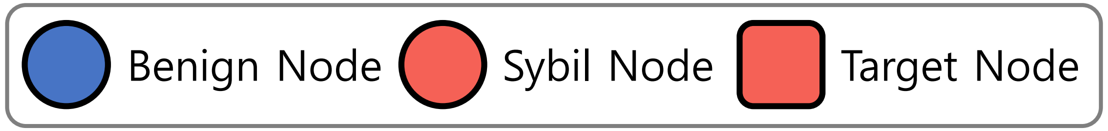


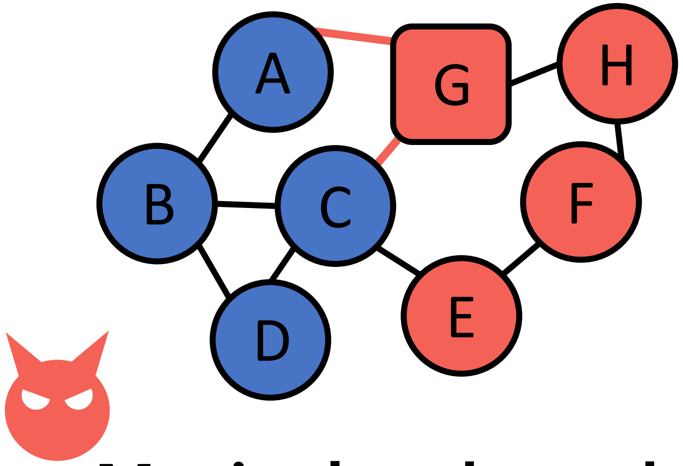
Manipulated graph



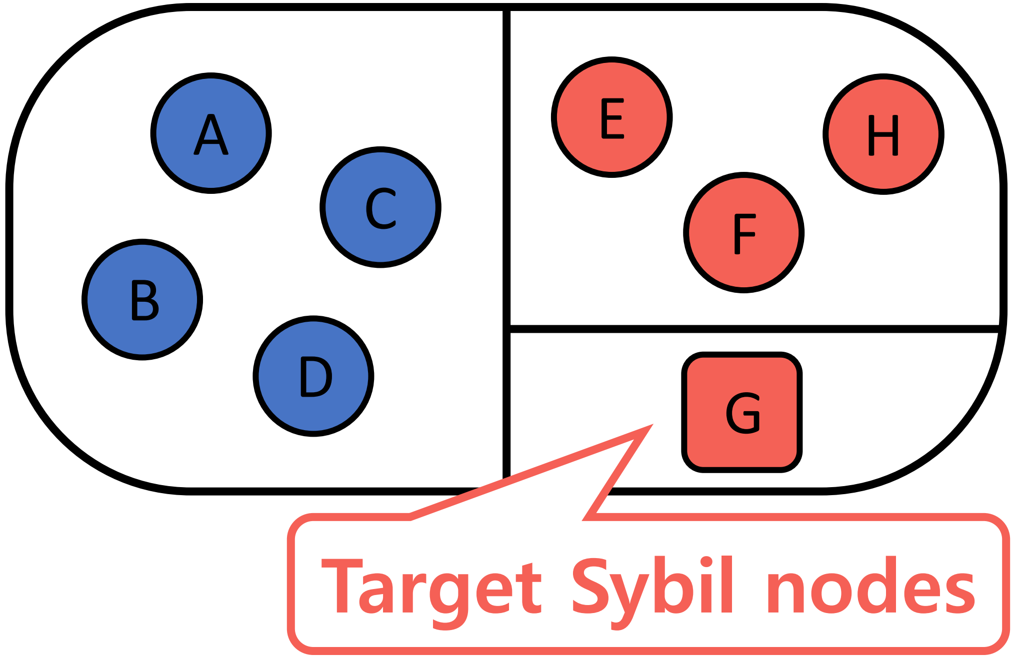
Goal 1. Identifying all ***target nodes*!**

Evaluation Metrics



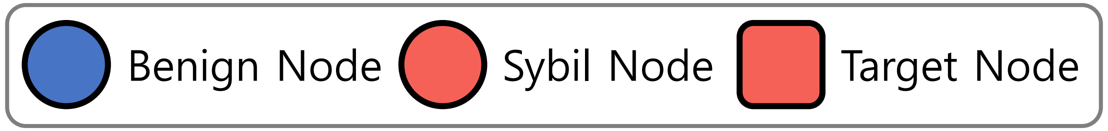


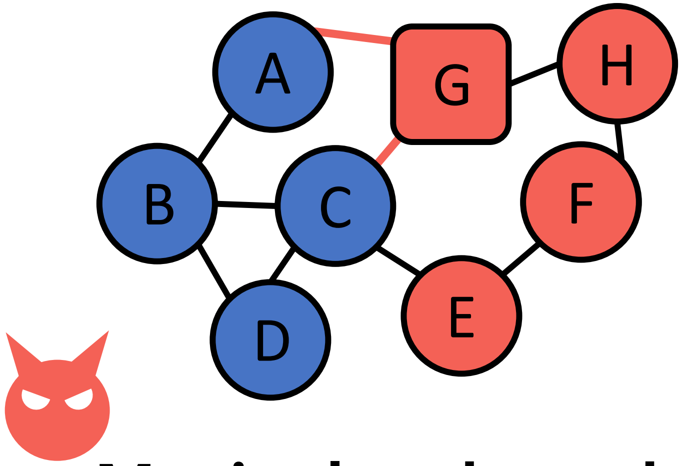
Manipulated graph



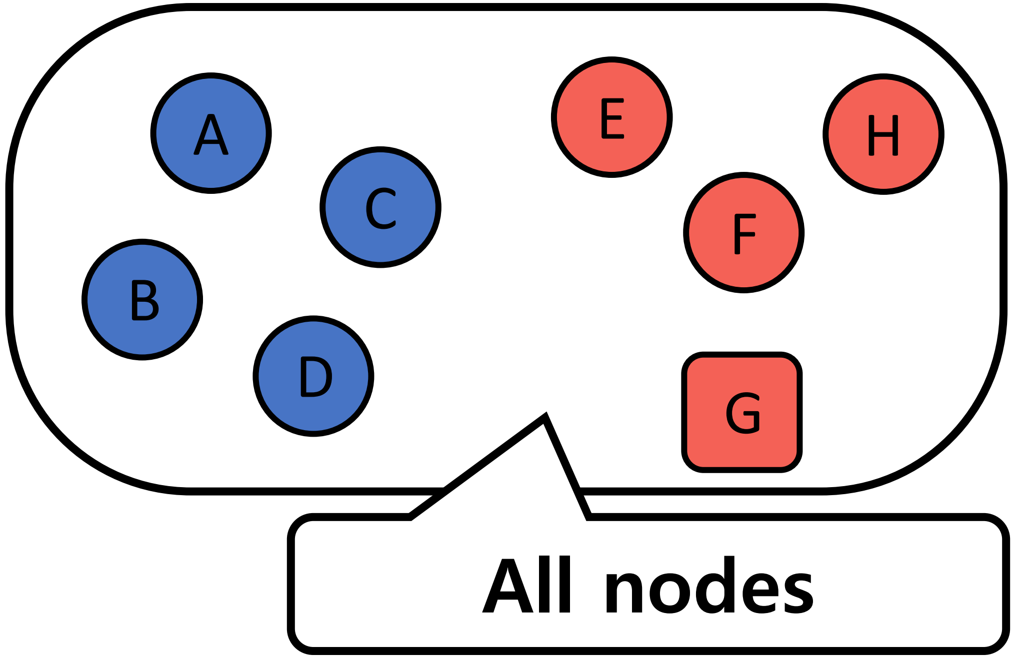
Goal 1. Low ***<u>false negative rate</u>*** of target nodes!

Evaluation Metrics



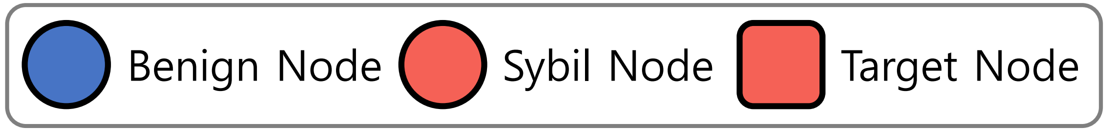


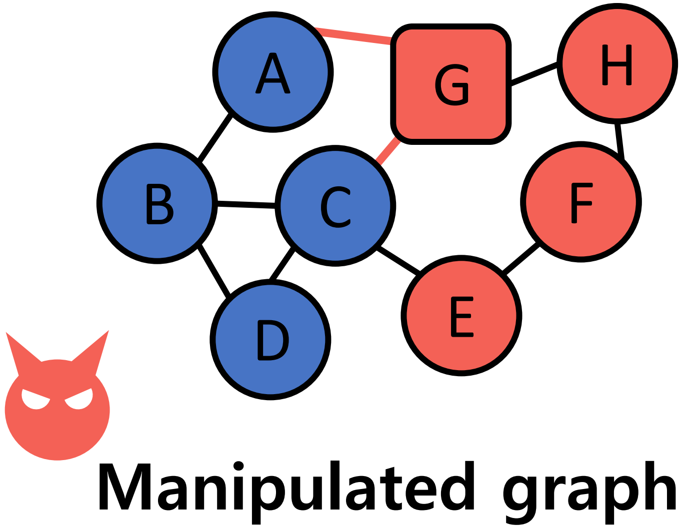
Manipulated graph

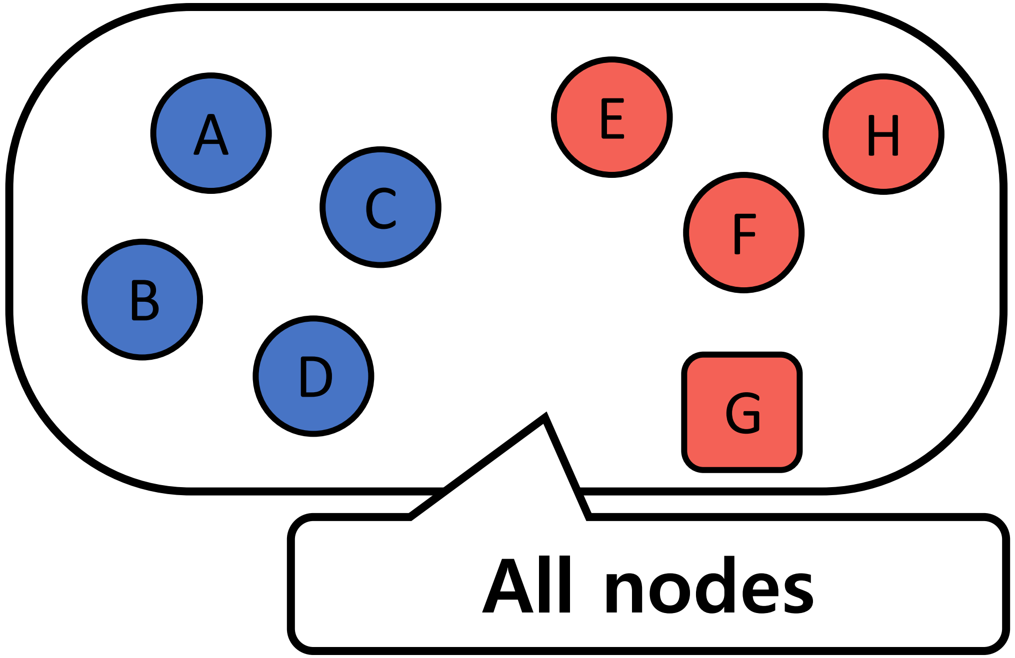


Goal 2. Correctly classifying ***all nodes!***

Evaluation Metrics







Goal 2. High ***area under the curve!***

vs. State-of-the-art Collective Classification

Dataset	FNR (↓)			AUC (↑)		
	RICC	SybilSCAR*	JWP**	RICC	SybilSCAR*	JWP**
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

* Wang et al. SybilSCAR: Sybil Detection in Online Social Networks via Local Rule based Propagation. INFOCOMM 2017
** Wang et al. Graph-based Security and Privacy Analytics via Collective Classification with Joint Weight Learning and Propagation. NDSS 2019

vs. State-of-the-art Collective Classification

- False negative rate (FNR) of target nodes

Dataset	FNR (↓)			AUC (↑)		
	RICC	SybilSCAR*	JWP**	RICC	SybilSCAR*	JWP**
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

The attack destroyed SybilSCAR and JWP!

KAIST Web Security & Privacy Lab

vs. State-of-the-art Collective Classification

- False negative rate (FNR) of target nodes

Use the exposed training set!

Dataset				(↑)		
	RICC	**SybilSCAR***	**JWP****		**SybilSCAR***	**JWP****
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

The attack destroyed SybilSCAR and JWP!

vs. State-of-the-art Collective Classification

- False negative rate (FNR) of target nodes

Dataset	FNR (↓)			AUC (↑)		
	RICC	SybilSCAR*	JWP**	RICC	SybilSCAR*	JWP**
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

RICC correctly identified target nodes!

KAIST Web Security & Privacy Lab

vs. State-of-the-art Collective Classification

- False negative rate (FNR) of target nodes

Use randomly sampled training sets!

Dataset				↑)		
	RICC		JWP**	RICC	SybilSCAR*	JWP**
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

RICC correctly identified target nodes!

vs. State-of-the-art Collective Classification

- Area under the curve (AUC)

Dataset	FNR (↓)			AUC (↑)		
	RICC	SybilSCAR*	JWP**	RICC	SybilSCAR*	JWP**
Enron	**0.01**	1.00	1.00	**0.9912**	0.9884	0.9875
Facebook	**0.11**	0.95	0.97	**0.9995**	0.9372	0.9551
Twitter_S	**0.00**	1.00	0.99	**0.8911**	0.7117	0.6921
Twitter_L	**0.01**	1.00	1.00	**0.7388**	0.7371	0.7375

RICC correctly classified other nodes!

For More Details

- Rationale behind our observations
- Random sampling-based collective classification algorithms
- Effect of the attacker’s budget
- Effect of the attacker’s strategy
- Effect of the hyperparameters
- RICC vs. GNN
- https://github.com/WSP-LAB/RICC

Conclusion

- We made a **novel observation** that adversarial attacks are highly tailored to the training set.

Conclusion

- We made a **novel observation** that adversarial attacks are highly tailored to the training set.

- Leveraging this observation, **we propose RICC, a novel CC framework** for the robust identification of Sybil accounts.

Conclusion

- We made a **novel observation** that adversarial attacks are highly tailored to the training set.
- Leveraging this observation, **we propose RICC, a novel CC framework** for the robust identification of Sybil accounts.
- **RICC significantly outperformed existing CC** in terms of identifying adversarial Sybil accounts.

Conclusion

- We made a **novel observation** that adversarial attacks are highly tailored to the training set.
- Leveraging this observation, **we propose RICC, a novel CC framework** for the robust identification of Sybil accounts.
- **RICC significantly outperformed existing CC** in terms of identifying adversarial Sybil accounts.

Question?