蚂蚁集团
ANT GROUP

Robust Fraud Tansaction Detection: A Two-Player Game Approach

Qi Tan, Yi Zhao, Laizhong Cui, Qi Li, Min Zhu, Xing Fu, Weiqiang Wang, Xiaotong Lin, and Ke Xu

February, 2026

Outline


Background


Existing Techniques


Causal Models of Feature Falsification

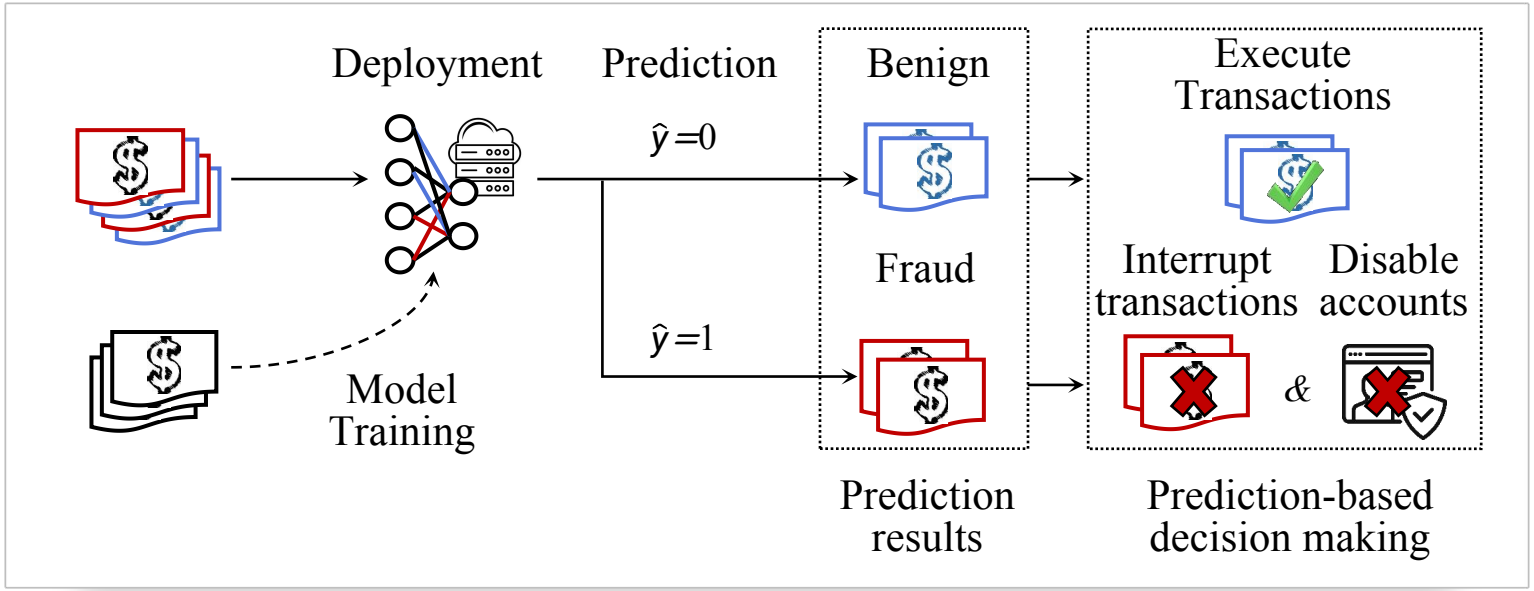

Two-Player Games in Fraud Detection


Design of GAMER


Evaluation


Conclusion

Background

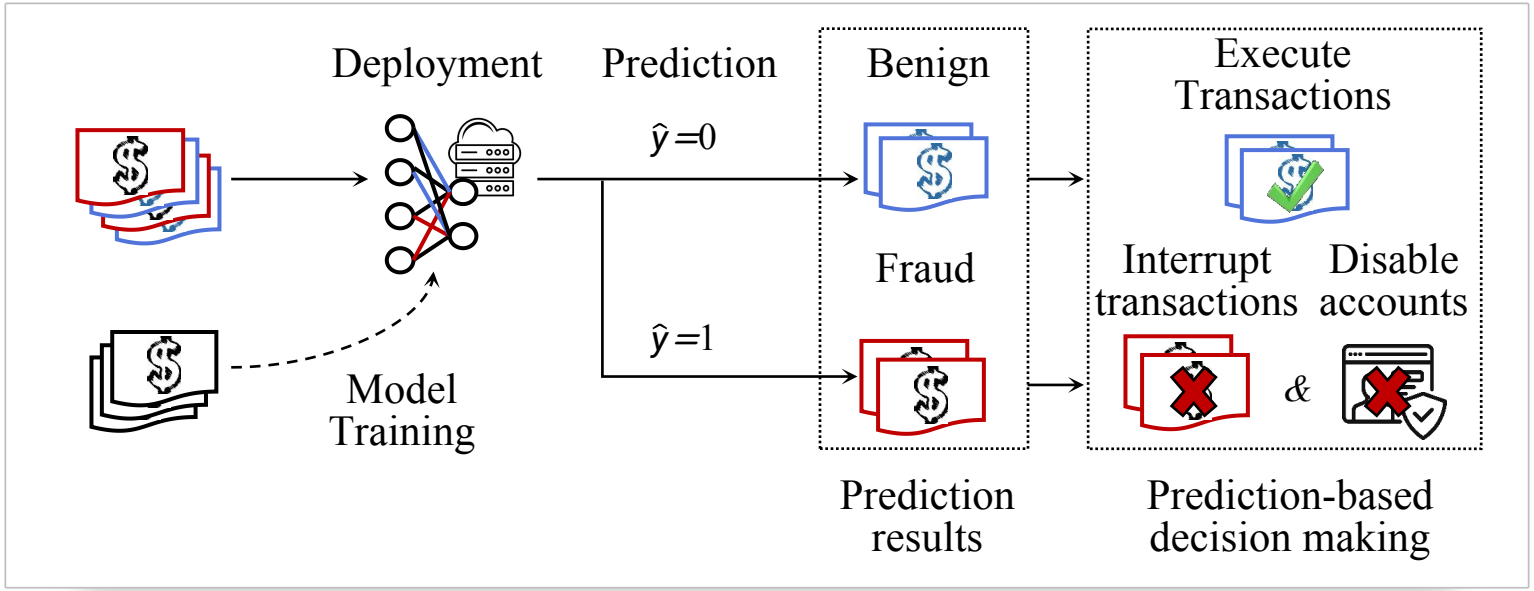


AI-based fraud detection

Detection: Training AI models to detect fraudulent activities

Decision-making: Processing the transactions and accounts based on predictions

Background

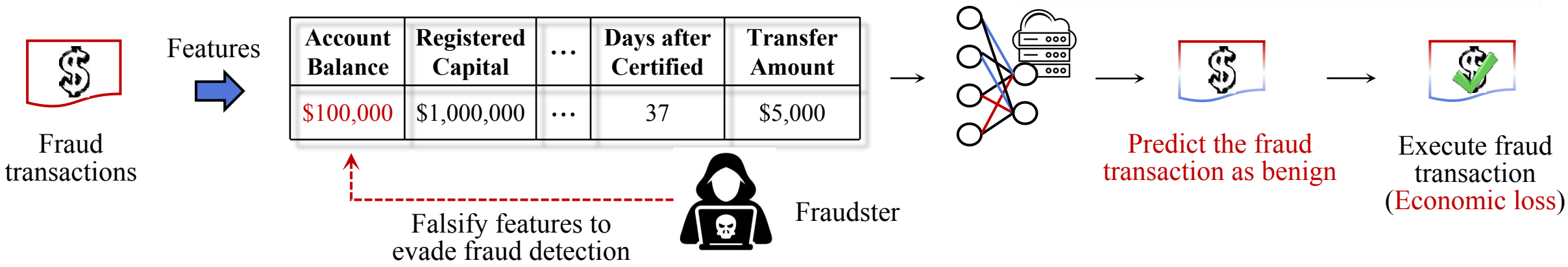


AI-based fraud detection

Detection: Training AI models to detect fraudulent activities

Decision-making: Processing the transactions and accounts based on predictions

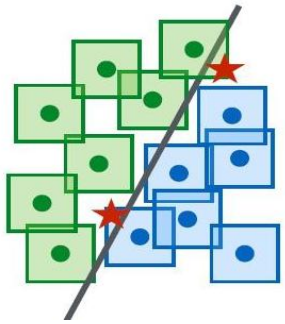
Fraudsters attack the deployed model, envading fraud detection system via feature falsification:



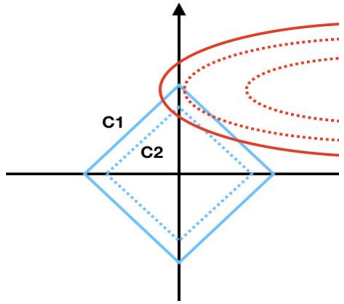
Existing Techniques

Existing Techniques for enhancing the robustness

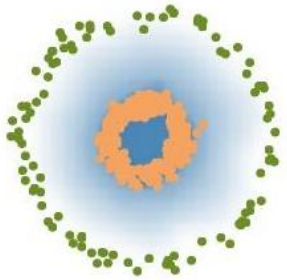
Adversarial training



Model Regularization



Data Augmentation



Difficult to deal with

Without using

Unrestricted falsifications

Monetary Features

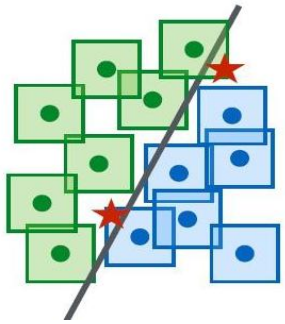
Temporal Features

Feature falsifiction costs resources

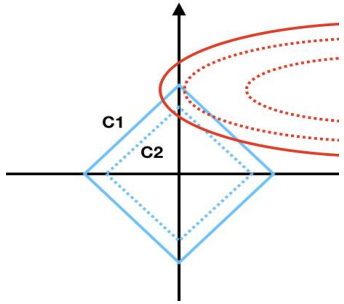
Existing Techniques

Existing Techniques for enhancing the robustness

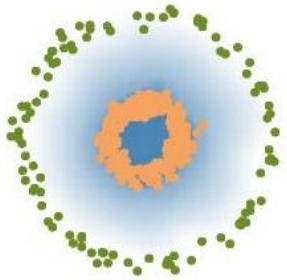
Adversarial training



Model Regularization



Data Augmentation



Difficult to deal with

Without using

Falsification is unrestricted

Monetary Features

Temporal Features

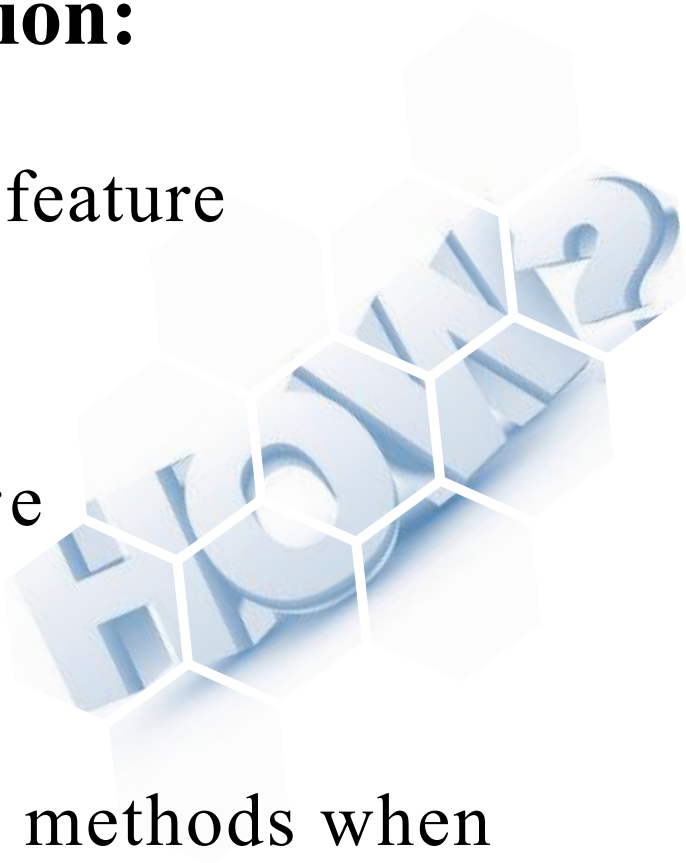
Falsifying these features costs resources

Rethinking the principle of falsifying features to evade fraud detection, and design new detection method to combat intelligent fraudsters

Key issues

Key issues in dealing with feature falsification:

- Why fraudsters can evade fraud detection via feature falsification?
- How to defend against potential feature falsification when collected data is limited?
- How to overcome the deficiencies in defense methods when transactions are benign?



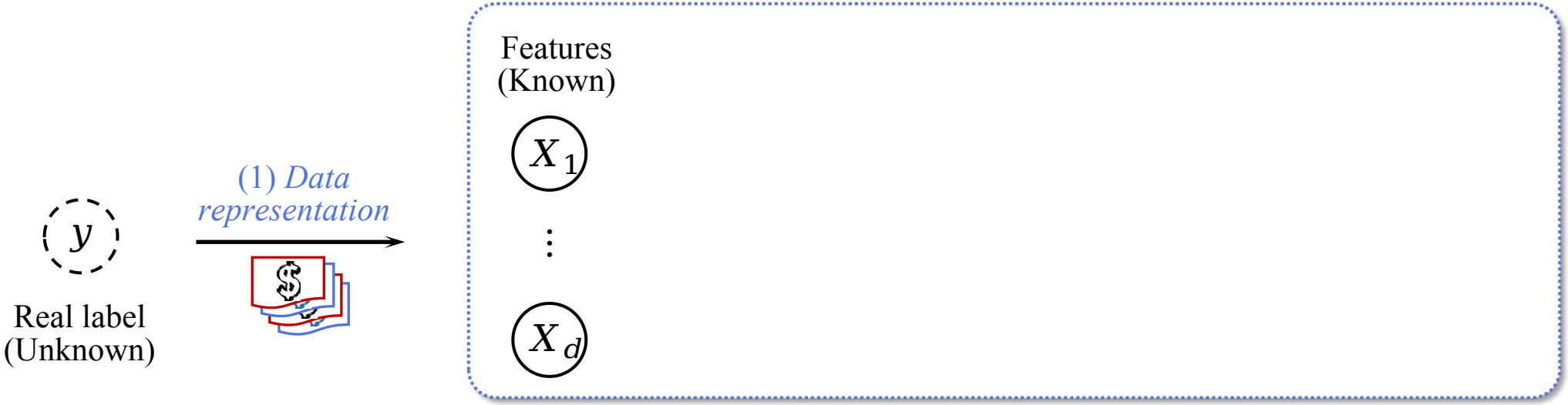
Key issues

Key issues in dealing with feature falsification:

- **Why fraudsters can evade fraud detection via feature falsification?**
- How to defend against potential feature falsification when collected data is limited?
- How to overcome the deficiencies in defense methods when transactions are benign?

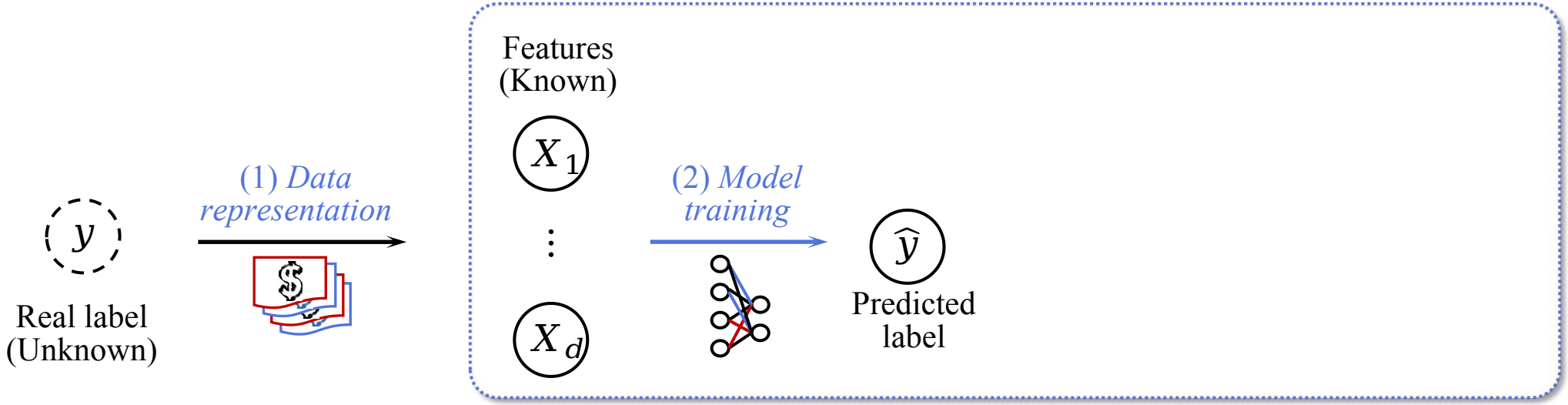
Causal Models of Feature Falsification

- **The entire lifecycle of fraud detection**



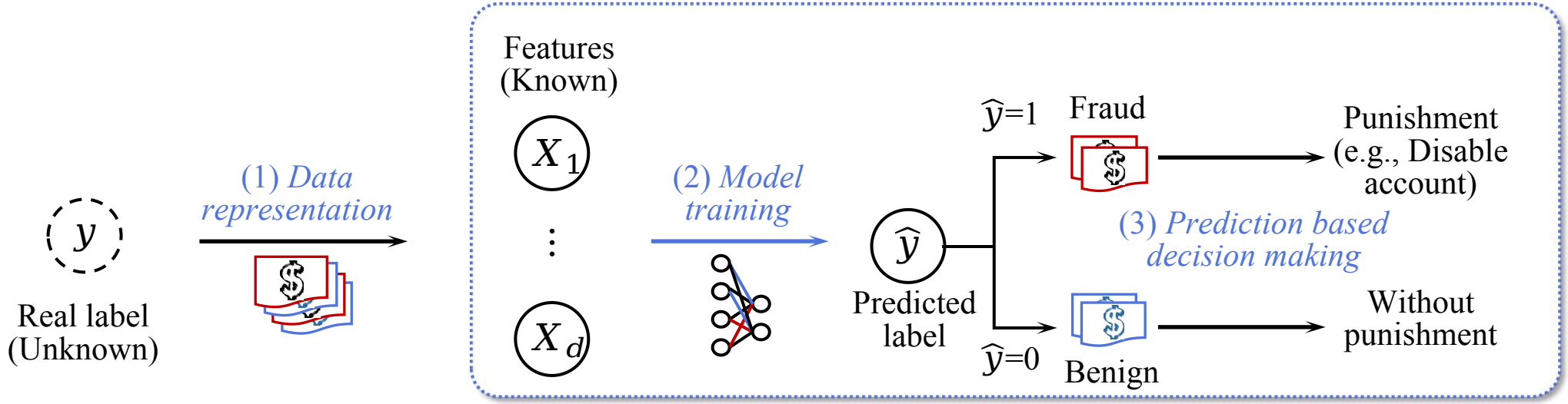
Causal Models of Feature Falsification

- **The entire lifecycle of fraud detection**



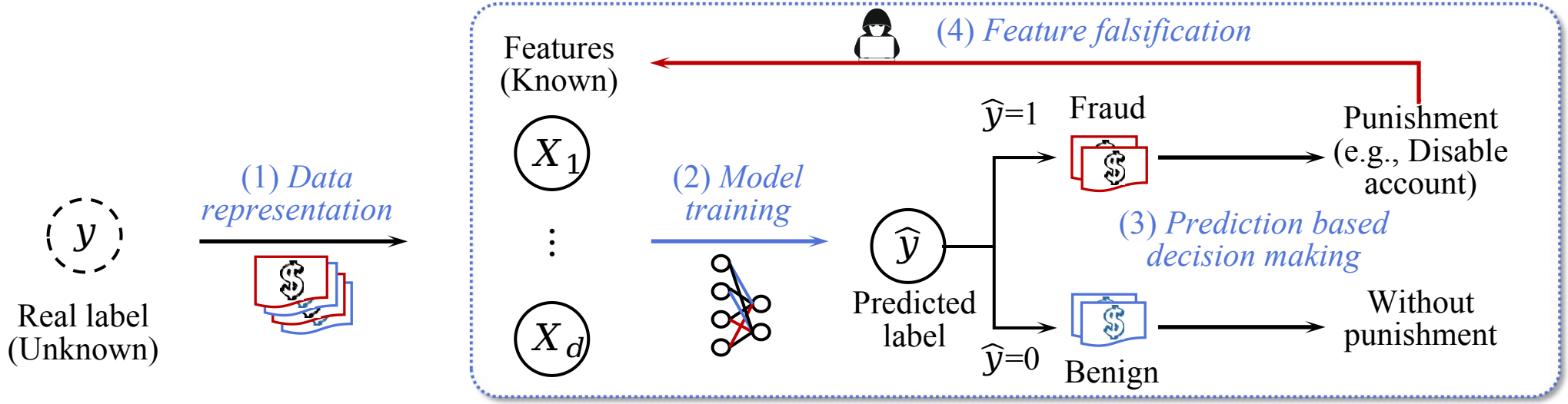
Causal Models of Feature Falsification

- **The entire lifecycle of fraud detection**



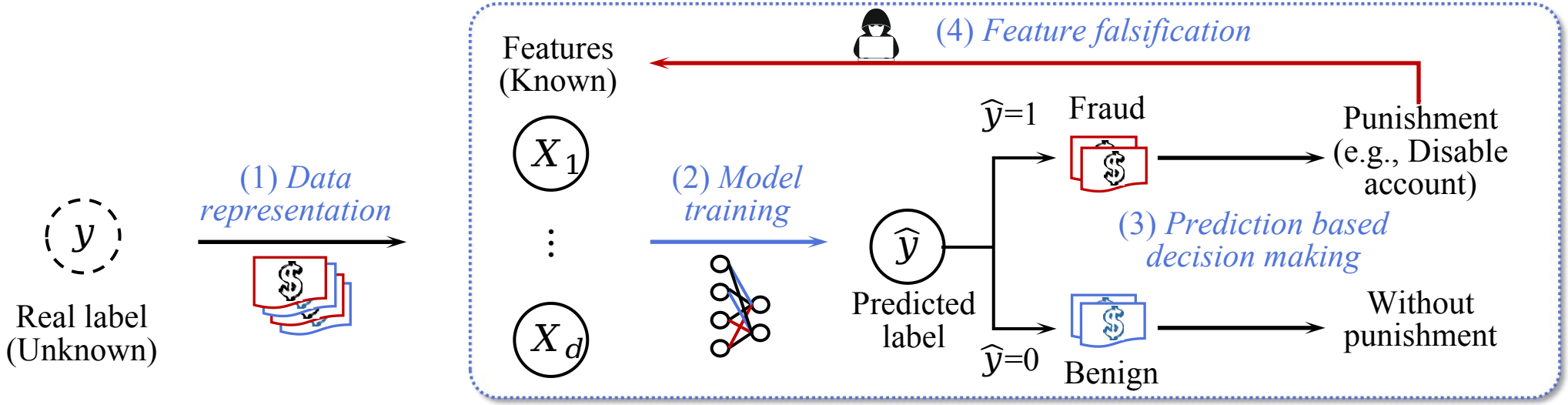
Causal Models of Feature Falsification

- **The entire lifecycle of fraud detection**



Causal Models of Feature Falsification

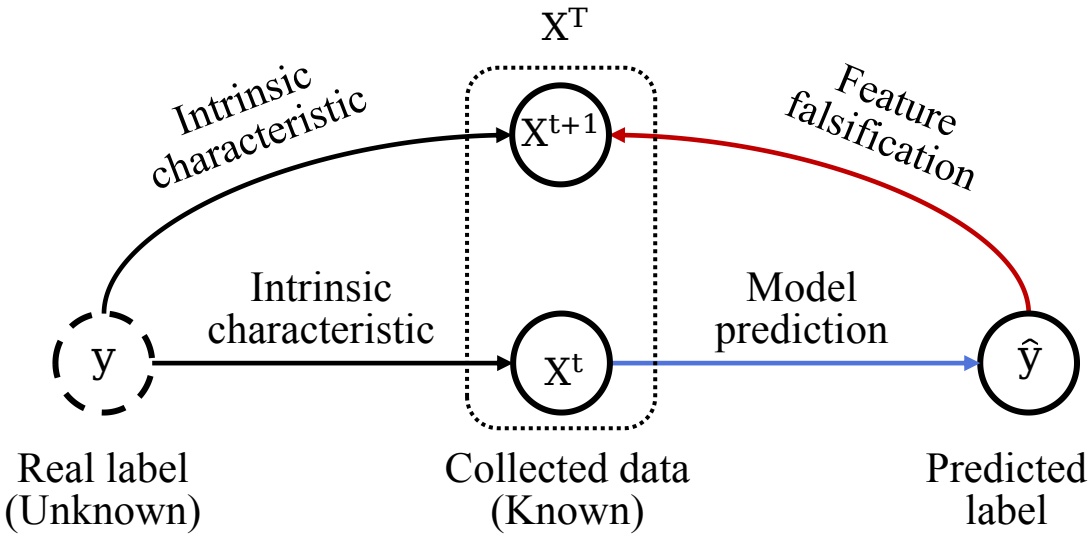
- **The entire lifecycle of fraud detection**



Formalize these processes in fraud detection, leveraging causal analysis to investigate the feature falsification

Causal Models of Feature Falsification

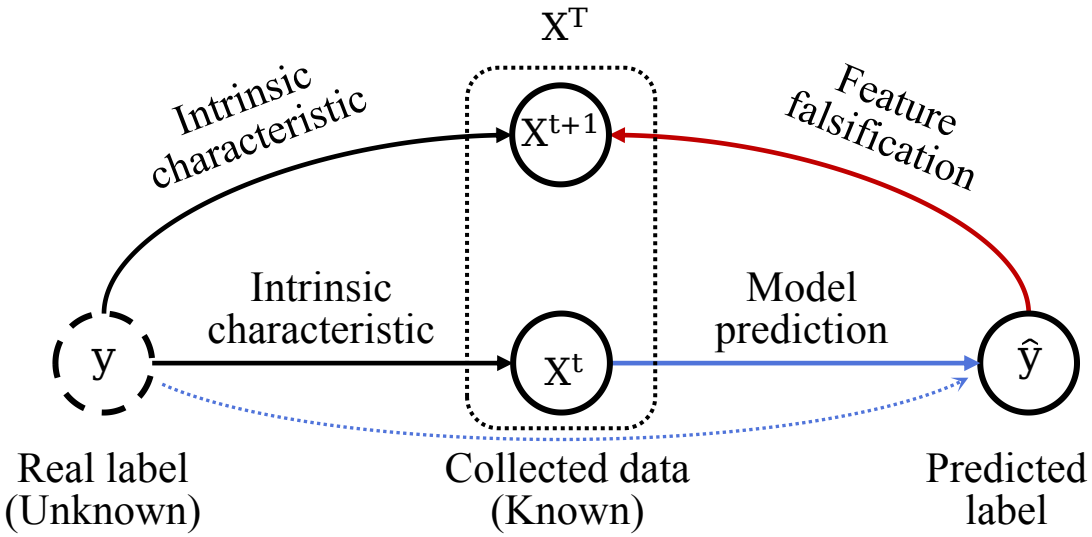
◆ Causal diagram of fraud detection



The four processes in the detection lifecycle can be formalized as a causal diagram, our target is to investigate the correlationship between real label and predicted label

Causal Models of Feature Falsification

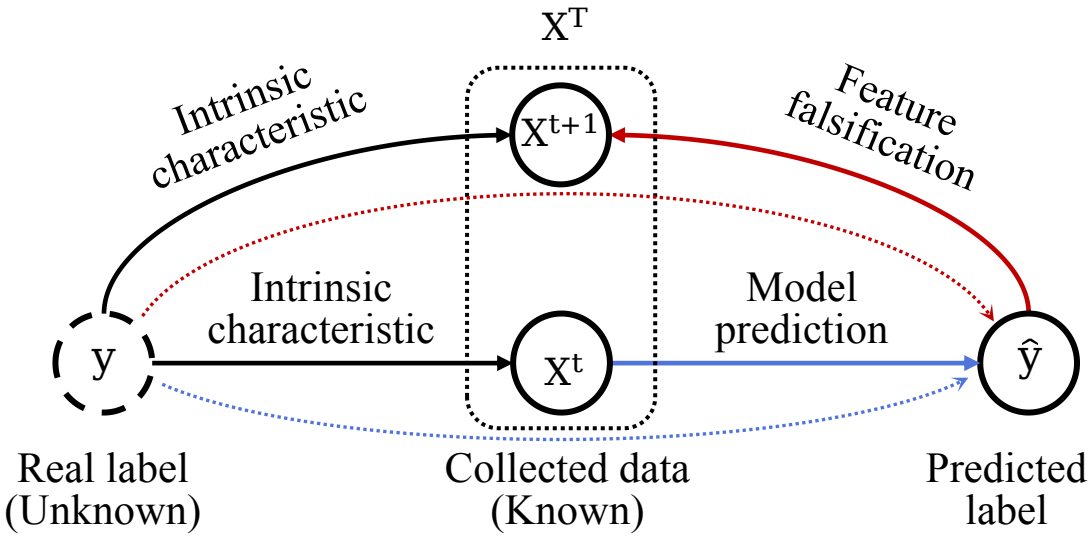
- **Causal diagram of fraud detection**



Causal!

Causal Models of Feature Falsification

◆ Causal diagram of fraud detection



Causing sprious correlation between y and $\hat{y}$

Falsification-inclusive causal path — ***Sprious!***

Model-inclusive causal path — ***Causal!***

$y = 1 \rightarrow \hat{y} = 0$ (i.e., successfully evading the detection system) is a sprious correlation

Selection bias (i.e., the insufficient training data) results in sprious correlations

Key issues

Key issues in dealing with feature falsification:

- Why fraudsters can evade fraud detection via feature falsification?
- **How to defend against potential feature falsification when collected data is limited?**
- How to overcome the deficiencies in defense methods when transactions are benign?

Two-player game in the detection

Motivation: Feature selection is a classical method for addressing selection bias

Two-player game in Fraud detection

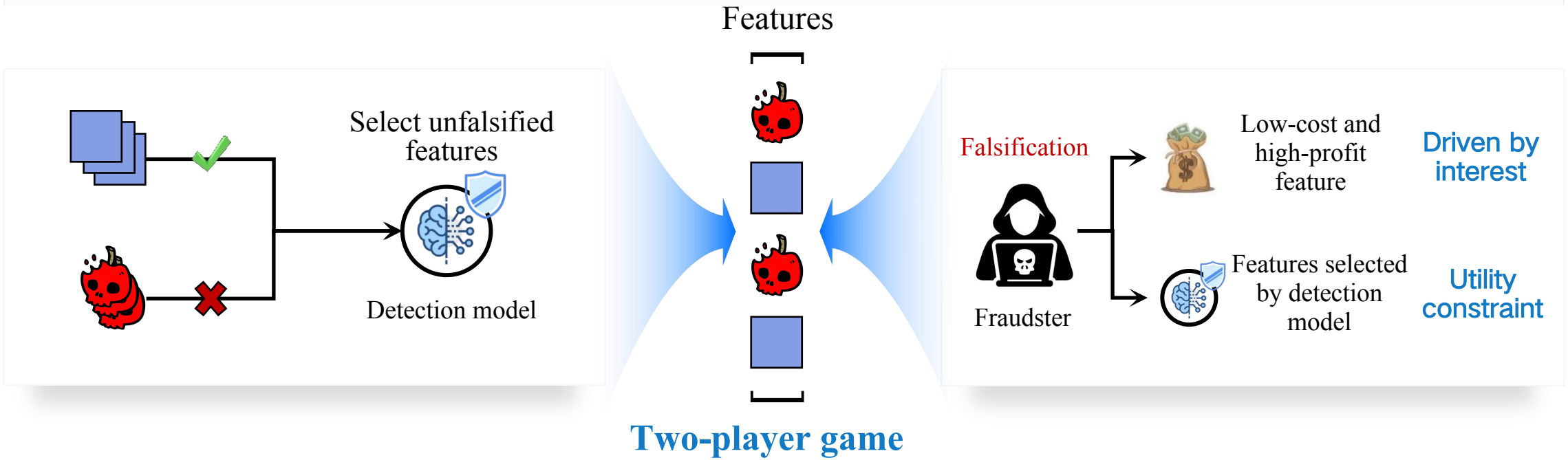
Motivation: Feature selection is a classical method for addressing selection bias

Key issue: Which feature should be selected?

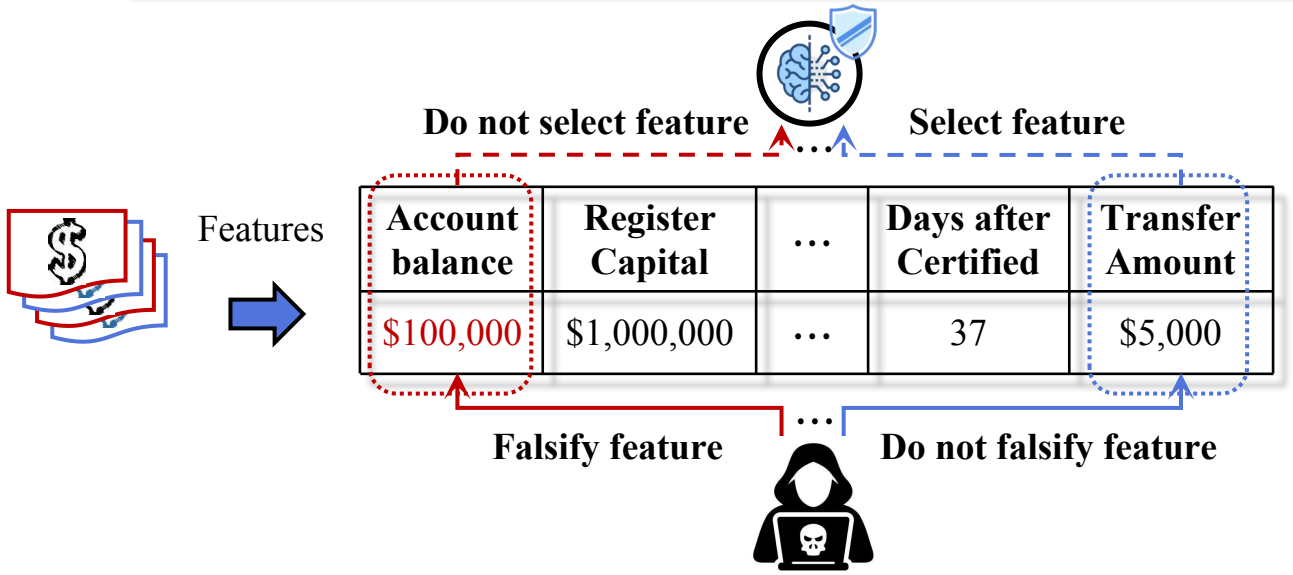
Two-player game in Fraud detection

Motivation: Feature selection is a classical method for addressing selection bias

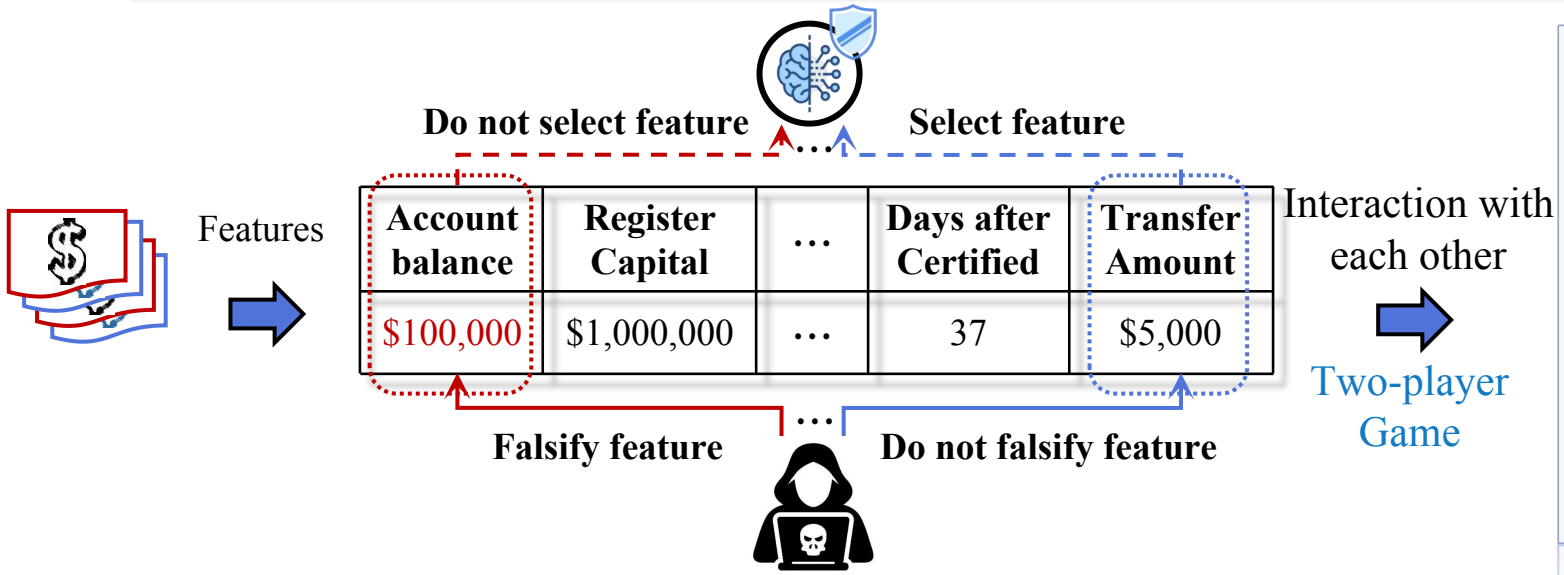
Key issue: Which feature should be selected?



Two-player game in Fraud detection



Two-player game in Fraud detection

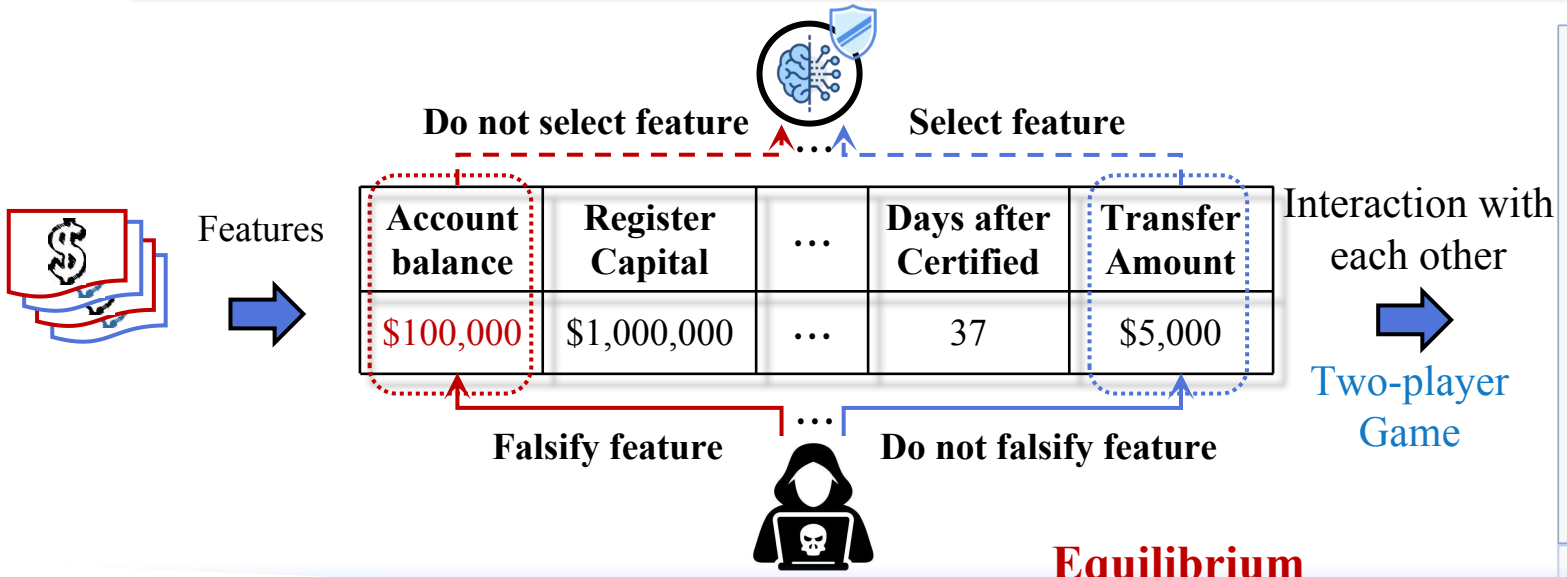


Fraudster

Detction model

	Falsify	Do not falsify
Select	8.5	12.5
Do not select	14.795	10.355

Two-player game in Fraud detection



Fraudster

Detction model

	Falsify	Do not falsify
Select	8.5	12.5
Do not select	14.795	10.355

Equilibrium

Expert knowledge representations

Features: $X_1 \cdots X_d$
Costs: $C_1 \cdots C_d$

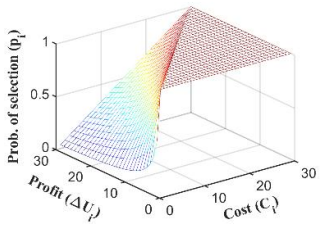
Costs of feature falsification

Features: $X_1 \cdots X_d$
Profits: $\Delta U_1 \cdots \Delta U_d$

Profits of feature falsification

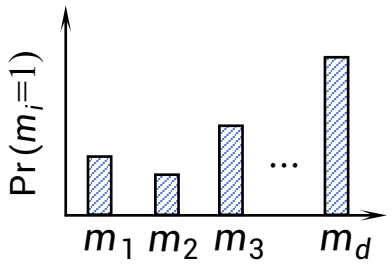
Probabilistic mask construction

(1) Cost-profit based probability calculation



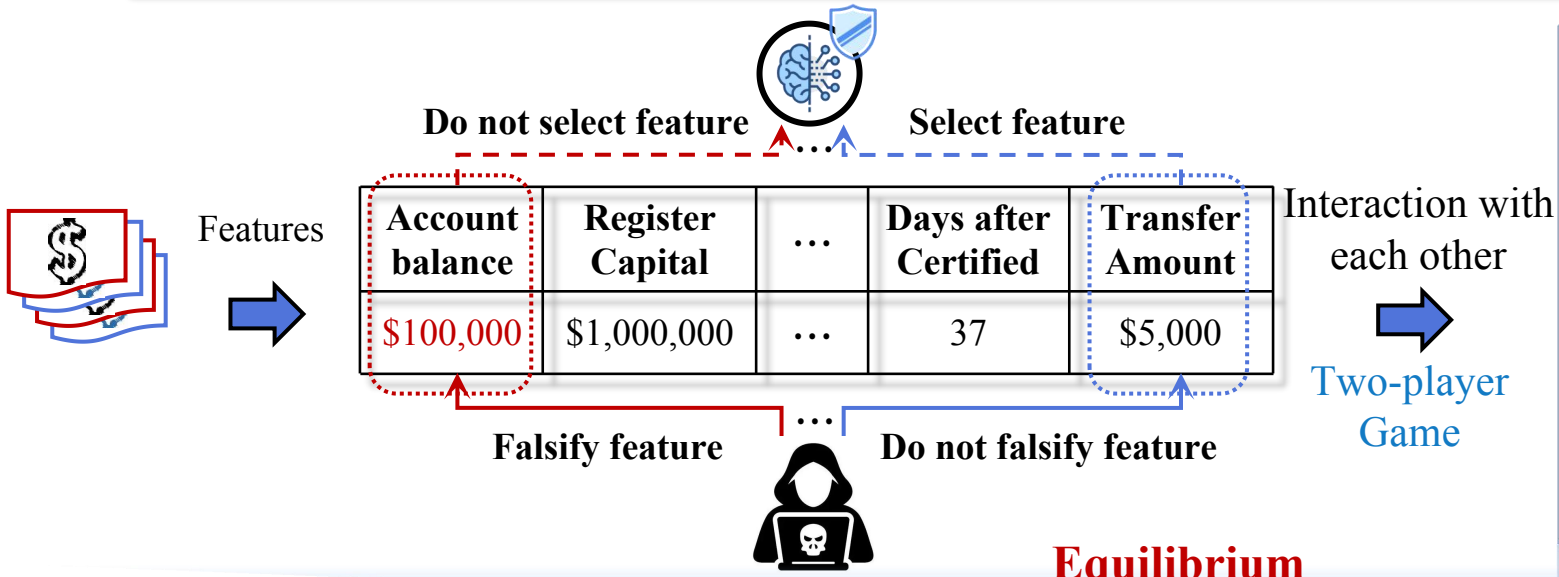
$$p_i = \min\left(1, \frac{C_i}{\Delta U_i}\right)$$

(2) Bernoulli variables based feature masks



Masks:
$m_i \sim \text{Bernoulli}(p_i)$

Two-player game in Fraud detection



Fraudster

Detction model

	Falsify	Do not falsify
Select	8.5	12.5
Do not select	14.795	10.355

Equilibrium

Expert knowledge representations

Features: $X_1 \cdots X_d$
Costs: $C_1 \cdots C_d$

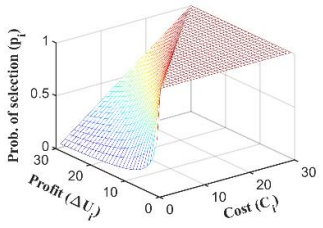
Costs of feature falsification

Features: $X_1 \cdots X_d$
Profits: $\Delta U_1 \cdots \Delta U_d$

Profits of feature falsification

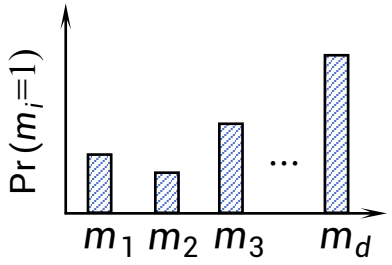
Probabilistic mask construction

(1) Cost-profit based probability calculation



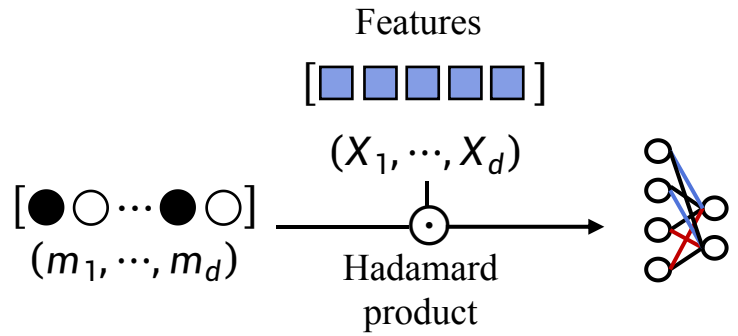
$$p_i = \min\left(1, \frac{C_i}{\Delta U_i}\right)$$

(2) Bernoulli variables based feature masks



Masks:
$m_i \sim \text{Bernoulli}(p_i)$

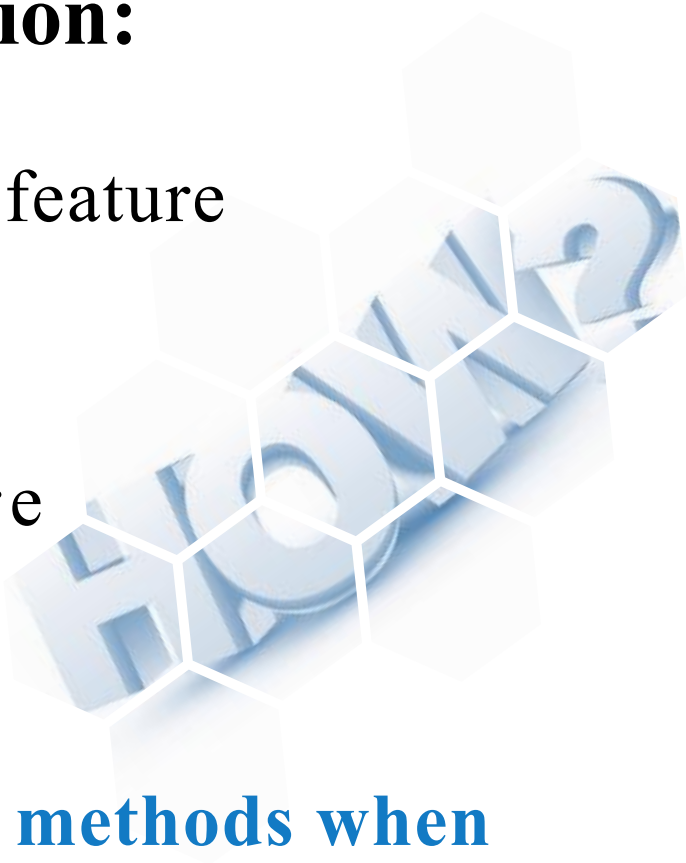
Training\Detecting with probabilistically masked features



Key issues

Key issues in dealing with feature falsification:

- Why fraudsters can evade fraud detection via feature falsification?
- How to defend against potential feature falsification when collected data is limited?
- **How to overcome the deficiencies in defense methods when transactions are benign?**

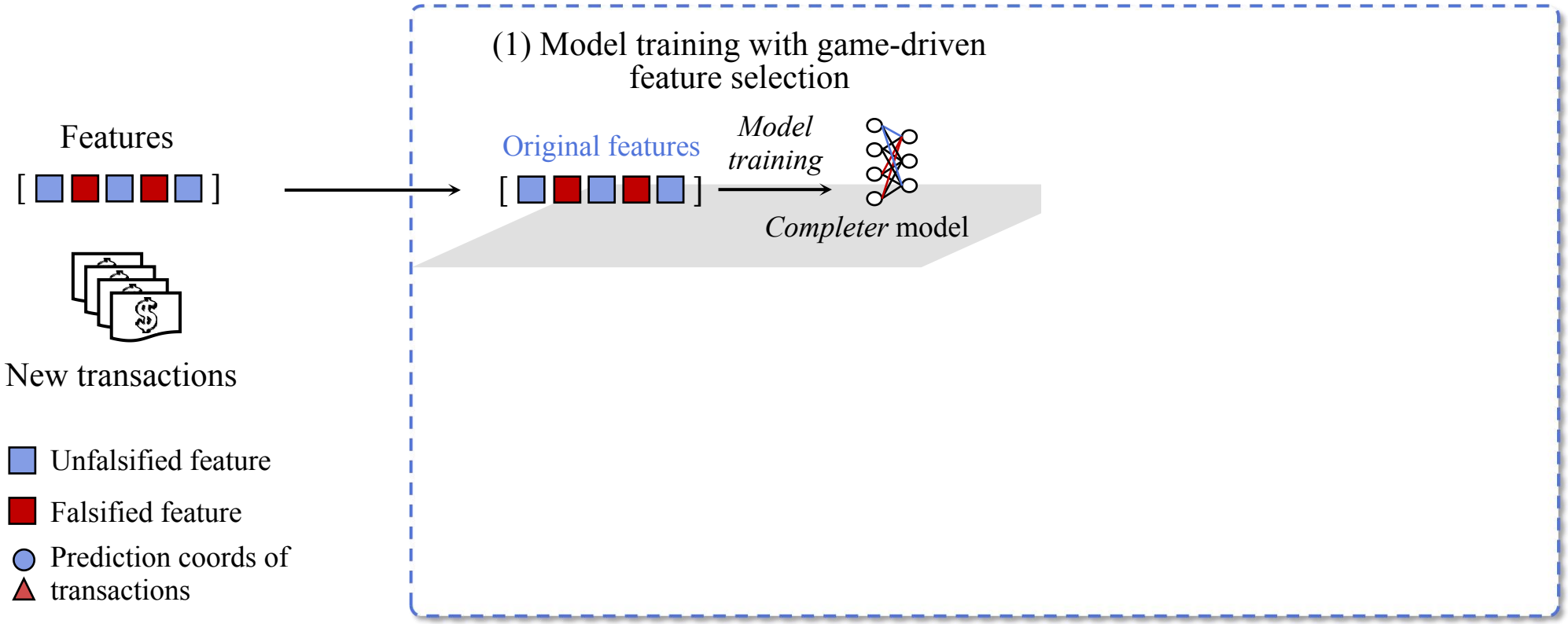


Desgin of GAMER

Solution: Use different models to detect transactions that they specialized in

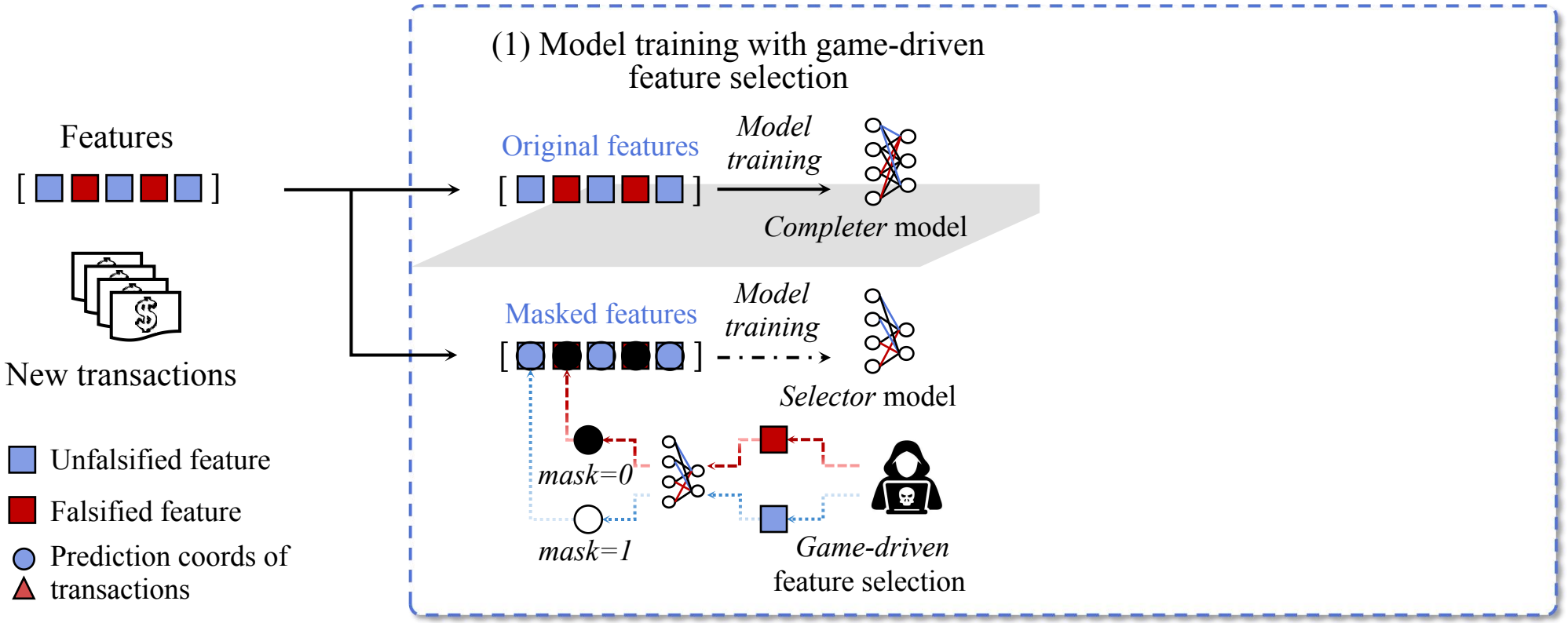
Desgin of GAMER

Solution: Use different models to detect transactions that they specialized in



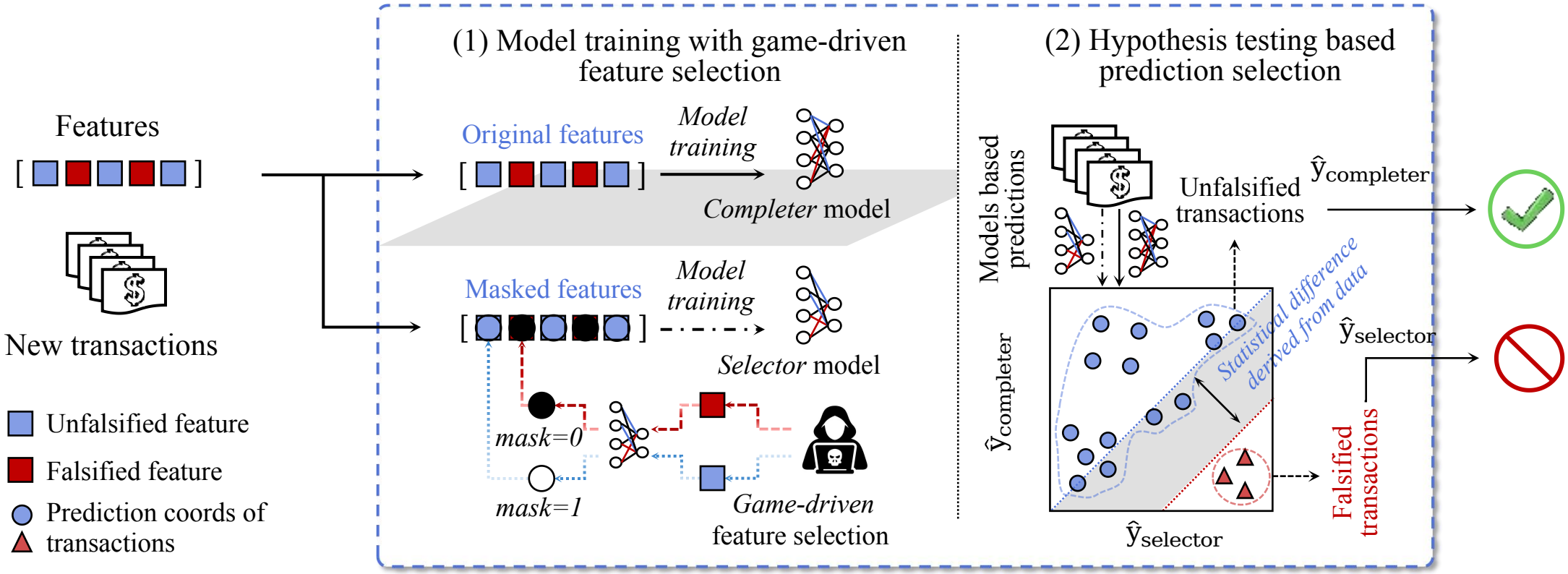
Desgin of GAMER

Solution: Use different models to detect transactions that they specialized in



Desgin of GAMER

Solution: Use different models to detect transactions that they specialized in



Evaluation

- Simulation:
  - We **randomly** set costs to each feature and repeat each experiment 30 times to **avoid the bias introduced by cost assignation**

- Real-world data:
  - Real-world data are collected from **inter-enterprise transactions**
  - The labels are collected from **whether transactions are complained about by users**
  - Each data has **302 features**, which capture various factors such as transaction amount, transaction frequency, etc

Real-world Transactions

Simulation

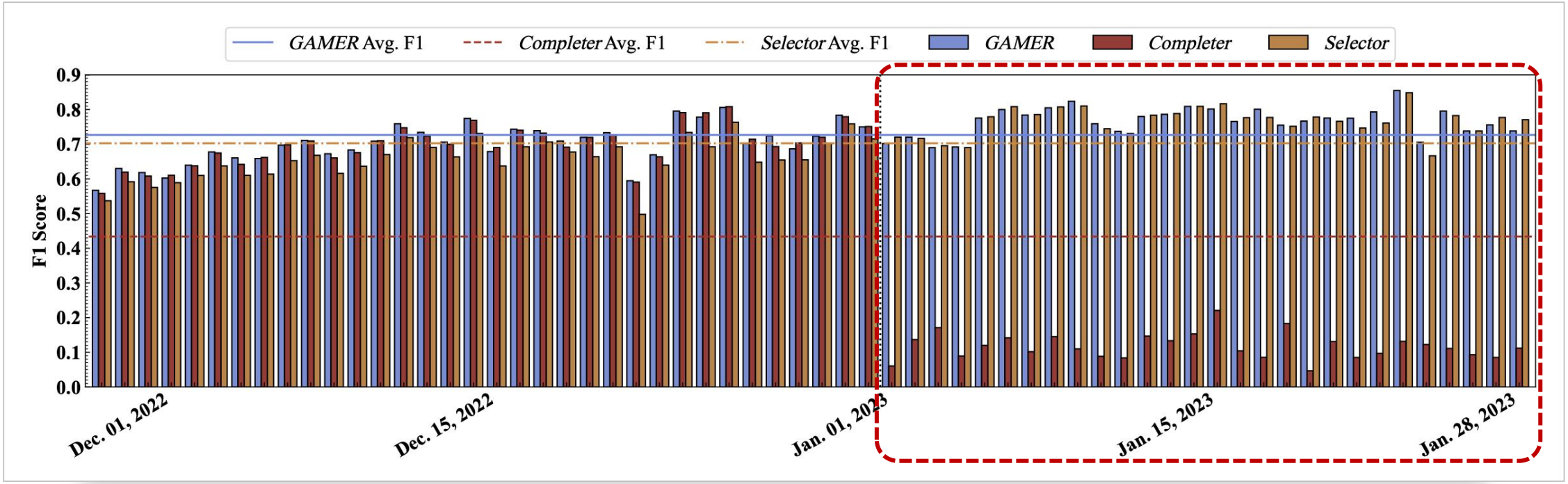
Rational Fraudsters (i.e., taking the cost-profit into account)

Dataset		AUC					F1 Score				
	At. Cost / Method	100	200	300	400	500	100	200	300	400	500
CreditCard	Non-robust	$0.9629_{(-)}$	$0.8547_{(-)}$	$0.7883_{(-)}$	$0.7199_{(-)}$	$0.6609_{(-)}$	$0.8729_{(-)}$	$0.7504_{(-)}$	$0.5483_{(-)}$	$0.3712_{(-)}$	$0.2827_{(-)}$
	CW [12]	$\mathbf{0.9970}_{(\uparrow)}$	$0.9566_{(\uparrow)}$	$0.8914_{(\uparrow)}$	$0.8269_{(\uparrow)}$	$0.7651_{(\uparrow)}$	$0.9107_{(\uparrow)}$	$0.8802_{(\uparrow)}$	$0.8476_{(\Uparrow)}$	$0.7937_{(\Uparrow)}$	$0.7342_{(\Uparrow)}$
	TARDES [85]	$0.9042_{(\downarrow)}$	$0.8748_{(\uparrow)}$	$0.8436_{(\uparrow)}$	$0.8178_{(\uparrow)}$	$0.8025_{(\Uparrow)}$	$0.7565_{(\downarrow)}$	$0.7323_{(\downarrow)}$	$0.7163_{(\uparrow)}$	$0.7077_{(\Uparrow)}$	$0.7041_{(\Uparrow)}$
	PGD [45]	$0.9265_{(\downarrow)}$	$0.8995_{(\uparrow)}$	$0.8746_{(\uparrow)}$	$0.8493_{(\uparrow)}$	$0.8306_{(\Uparrow)}$	$0.7467_{(\Downarrow)}$	$0.7286_{(\Downarrow)}$	$0.7150_{(\Downarrow)}$	$0.7032_{(\Downarrow)}$	$0.6955_{(\Downarrow)}$
	FGSM [25]	$0.9297_{(\downarrow)}$	$0.8940_{(\uparrow)}$	$0.8625_{(\uparrow)}$	$0.8482_{(\uparrow)}$	$0.8405_{(\Uparrow)}$	$0.8590_{(\downarrow)}$	$0.8452_{(\uparrow)}$	$0.8365_{(\Uparrow)}$	$0.8335_{(\Uparrow)}$	$0.8317_{(\Uparrow)}$
	GAMER	$0.9828_{(\uparrow)}$	$\mathbf{0.9584}_{(\uparrow)}$	$\mathbf{0.9402}_{(\Uparrow)}$	$\mathbf{0.9247}_{(\Uparrow)}$	$\mathbf{0.9164}_{(\Uparrow)}$	$\mathbf{0.9403}_{(\uparrow)}$	$\mathbf{0.9115}_{(\Uparrow)}$	$\mathbf{0.8954}_{(\Uparrow)}$	$\mathbf{0.8785}_{(\Uparrow)}$	$\mathbf{0.8687}_{(\Uparrow)}$

Irrational Fraudsters

Attack method	Without attack	AutoAttack [19]			Square attack [2] (Black-box attack)		
Def. method / Metric	F1	F1	ASR	$\frac{At.\ Cost}{ASR}$	F1	ASR	$\frac{At.\ Cost}{ASR}$
Non-robust	$\mathbf{0.8298}_{(-)}$	$0.0323_{(-)}$	$0.9812_{(-)}$	$11.55_{(-)}$	$0.0446_{(-)}$	$0.9738_{(-)}$	$11.58_{(-)}$
CW [12]	$0.6090_{(\downarrow 26.6\%)}$	$0.3743_{(\uparrow)}$	$0.7210_{(\downarrow 26.5\%)}$	$4.75_{(\downarrow 58.9\%)}$	$0.4142_{(\uparrow)}$	$0.6964_{(\downarrow 28.5\%)}$	$3.79_{(\downarrow 67.3\%)}$
TARDES [85]	$0.5922_{(\downarrow 28.6\%)}$	$0.4749_{(\uparrow)}$	$0.4891_{(\downarrow 50.1\%)}$	$4.34_{(\downarrow 62.4\%)}$	$0.4796_{(\uparrow)}$	$0.4751_{(\downarrow 51.2\%)}$	$4.41_{(\downarrow 61.9\%)}$
PGD [45]	$0.6006_{(\downarrow 27.6\%)}$	$0.4678_{(\uparrow)}$	$0.5442_{(\downarrow 44.5\%)}$	$4.50_{(\downarrow 61.0\%)}$	$0.4677_{(\uparrow)}$	$0.5381_{(\downarrow 44.8\%)}$	$4.76_{(\downarrow 58.8\%)}$
FGSM [25]	$0.5234_{(\downarrow 36.9\%)}$	$0.0323_{(-)}$	$0.9670_{(\downarrow 1.44\%)}$	$9.77_{(\downarrow 15.4\%)}$	$0.0367_{(\downarrow)}$	$0.9609_{(\downarrow 1.32\%)}$	$10.20_{(\downarrow 11.9\%)}$
GAMER	$0.8012_{(\downarrow 3.45\%)}$	$\mathbf{0.6356}_{(\uparrow)}$	$\mathbf{0.4145}_{(\downarrow 57.8\%)}$	$\mathbf{23.26}_{(\uparrow 101.4\%)}$	$\mathbf{0.7308}_{(\uparrow)}$	$\mathbf{0.2826}_{(\downarrow 70.9\%)}$	$\mathbf{34.67}_{(\uparrow 199.4\%)}$

Evaluation on real-world data



Using the equilirium in detection, GAMER increases the F1 score by 67.5% on average for two-month fraud detection

Conclusion

- Insufficient training data (i.e., selection bias) causes the sprious correlations between real label and predicted label ($y = 1 \rightarrow \hat{y} = 0$)
  - Using feature selection to overcome selection bias
- Using Two-player game to decide the selection strategies of detection model
- Using different model to detect the transactions they specilized in

Email: tanqi@szu.edu.cn

Thanks! Questions?