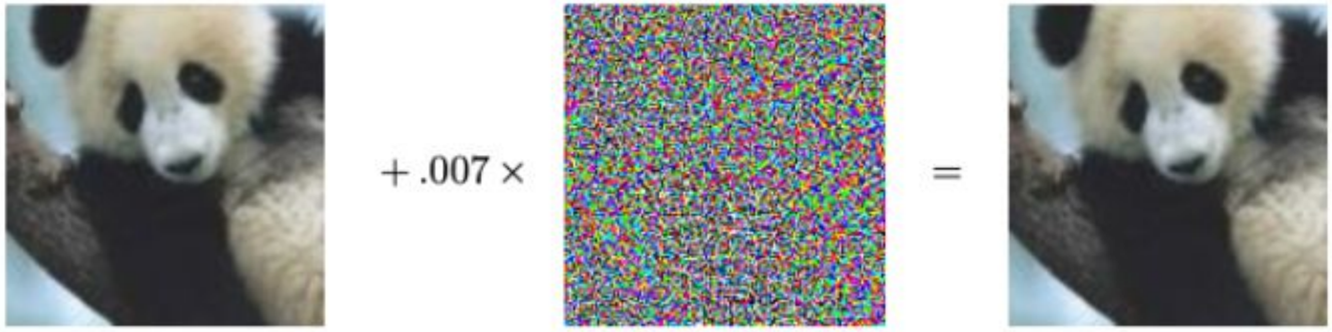
Adversarial Robustness for Tabular Data through Cost and Utility Awareness

Klim Kireev,* Bogdan Kulynych,* Carmela Troncoso

NDSS 2023, San Diego

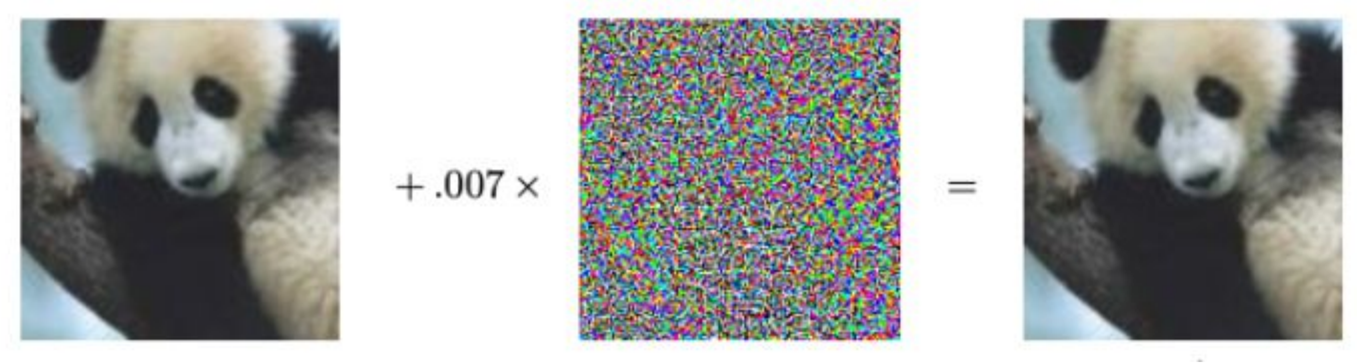
Adversarial examples



“Panda”

“Gibbon”

Adversarial examples



“Panda” “Gibbon”

Comes to mind when someone says “adversarial attack”

Example of a security-critical ML system: Fraud detector

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

System output

Example of a security-critical ML system: Fraud detector

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

System output

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	gmail.com	Italy	No

Example of a security-critical ML system: Fraud detector

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	gmail.com	Italy	No

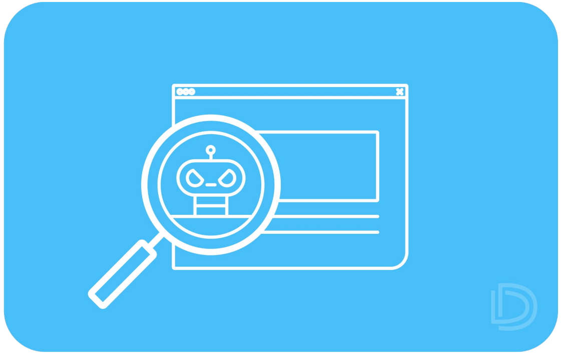
System output

What happened here is also an evasion attack on tabular data

Other security-critical ML application areas

Fraud detection

Credit risk assessment

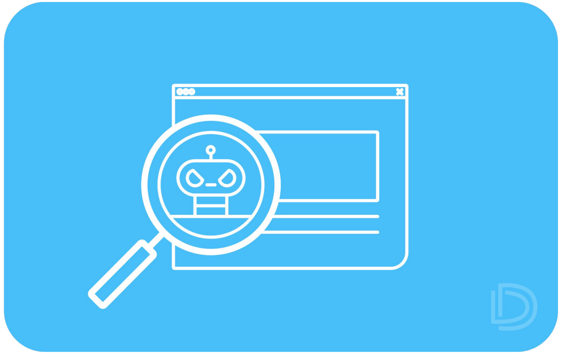


Bot detection

Other security-critical ML application areas

Fraud detection

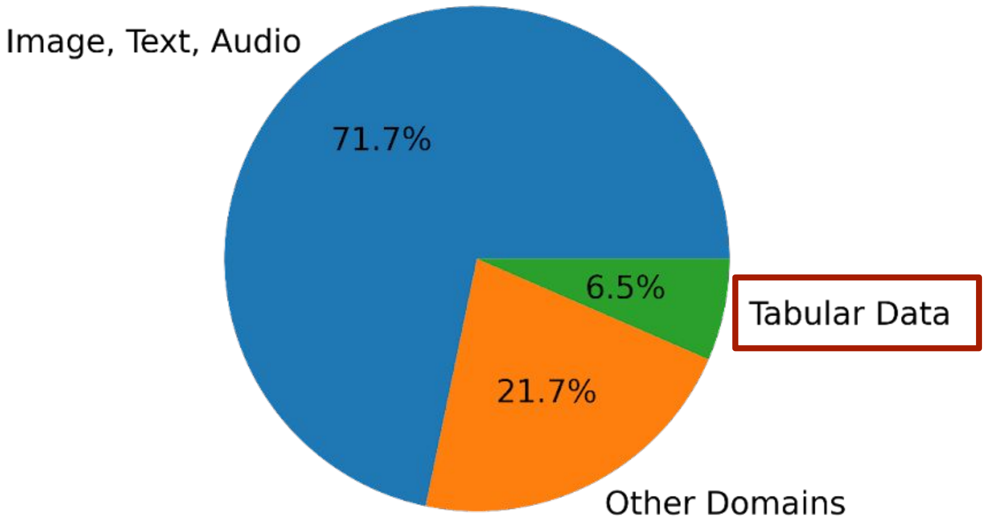
Credit risk assessment



Bot detection

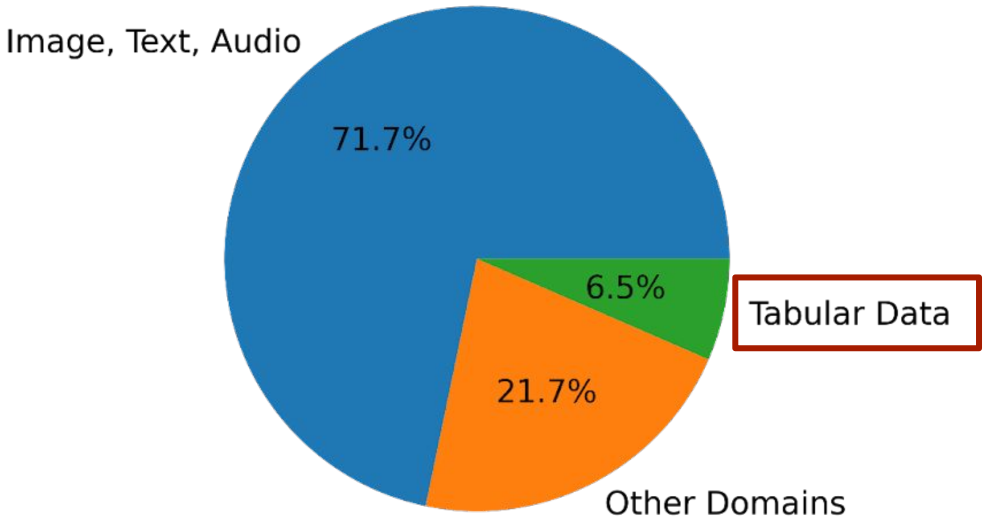
Machine learning systems working on these problems operate on tabular data

Domains studied in the academic literature



Evasion attacks/defences paper distribution on major security conferences over last 3 years according to Apruzzese et al. 2023

Domains studied in the academic literature



But do we need a different approach for tabular data?

Evasion attacks/defences paper distribution on major security conferences over last 3 years according to Apruzzese et al. 2023

Standard definition of adversarial examples

$$\max_{x' \in \mathcal{F}(x,y)} \ell\big(f(x'),y\big) \quad \text{s.t.} \ \underline{\|x' - x\|_p} \leq \varepsilon$$

L_p distance, L_∞ and L_2 are the most popular choices

Standard definition of adversarial examples

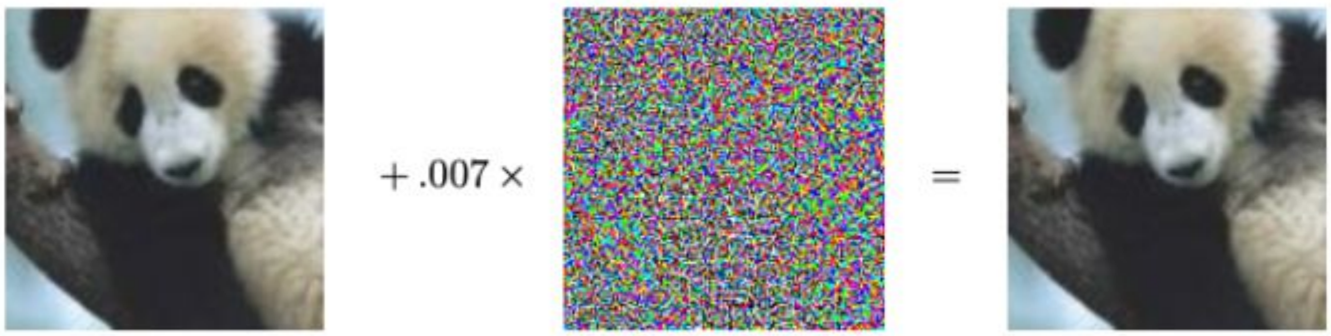
$$\max_{x' \in \mathcal{F}(x,y)} \quad \ell\left(f(x'), y\right) \quad \text{s.t.} \ \underline{\left\|x' - x\right\|_p} \leq \varepsilon$$

L_p distance, L_∞ and L_2 are the most popular choices

This definition was designed for images

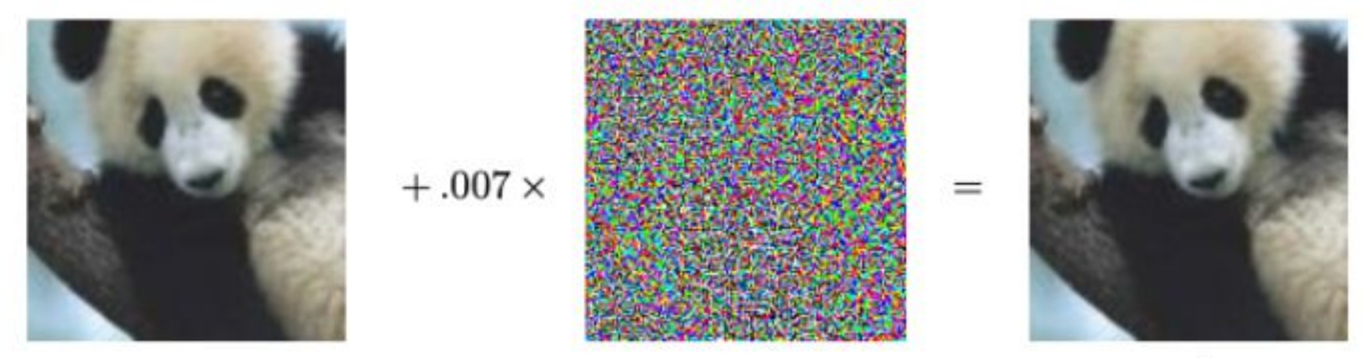
Imperceptibility in a tabular data context

$$\max_{x' \in \mathscr{F}(x, y)} \ell\left(f(x'), y\right) \quad \text{s.t. } \underline{\left\|x' - x\right\|_p} \leq \varepsilon$$



Imperceptibility in a tabular data context

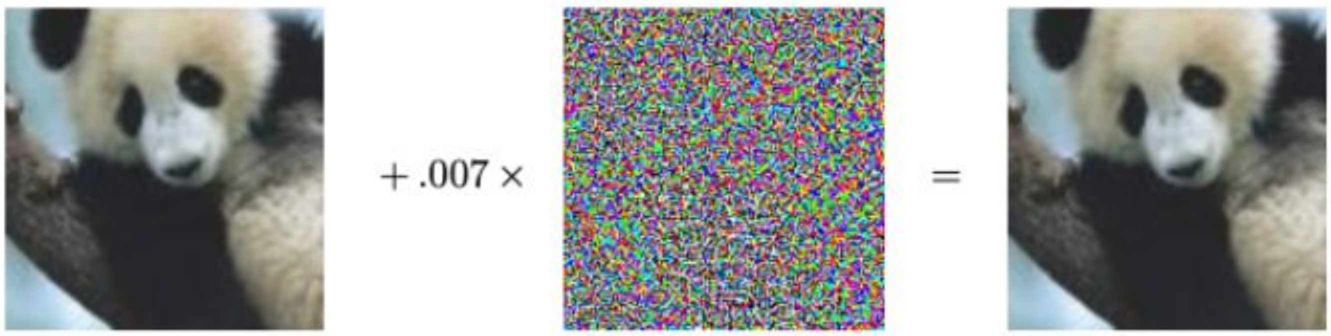
$$\max_{x' \in \mathscr{F}(x,y)} \ell\big(f(x'), y\big) \quad \text{s.t.} \quad \underline{\|x' - x\|_p} \leq \varepsilon$$



It is definitely an imperceptible change

Imperceptibility in a tabular data context

$$\max_{x' \in \mathscr{F}(x,y)} \ell\big(f(x'), y\big) \quad \text{s.t. } \underline{\|x' - x\|_p} \le \varepsilon$$



It is definitely an imperceptible change

"Imperceptibility" implicitly defines threat model

Imperceptibility in a tabular data context

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x'*

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	MasterCard	gmail.com	UK	No

Imperceptibility in a tabular data context

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x'*

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	MasterCard	gmail.com	UK	No

But what about this change? Is it imperceptible?

How we fix it: Cost-constrained adversary

$$\|x' - x\|_p \leq \varepsilon \quad \Rightarrow \quad \boxed{c(x, x') \leq \varepsilon}$$

How we fix it: Cost-constrained adversary

$$\|x' - x\|_p \leq \varepsilon \quad \Rightarrow \quad \boxed{c(x, x') \leq \varepsilon}$$

We define adversarial capabilities

through financial constraints

How we fix it: Cost-constrained adversary

$$\|x' - x\|_p \leq \varepsilon \quad \Rightarrow \quad \boxed{c(x, x') \leq \varepsilon}$$

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

How we fix it: Cost-constrained adversary

$$\|x' - x\|_p \leq \varepsilon \quad \Rightarrow \quad \boxed{c(x, x') \leq \varepsilon}$$

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x*': $20 $0.5 $14 $c(x, x') = \$\,34.5$

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	MasterCard	gmail.com	UK	No

Value of different adversarial examples in image domains

Value of different adversarial examples in image domains

These two pandas have the same value for an adversary

Value of different adversarial examples in tabular data

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Value of different adversarial examples in tabular data

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x**:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$28	Visa	epfl.ch	Italy	Yes

Value of different adversarial examples in tabular data

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x**:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$28	Visa	epfl.ch	Italy	Yes

What about these transactions?

How we fix it: Adversarial utility

$$\boxed{u_{x,y}(x') \triangleq g(x') - c(x, x')}$$

Gain g(x’) – potential returns from an attack, e.g. Transaction Amount

$$c(x, x') \leq \varepsilon \quad \Rightarrow \quad \boxed{u_{x,y}(x') \geq \tau}$$

How we fix it: Adversarial utility

$$\boxed{u_{x,y}(x') \triangleq g(x') - c(x, x')}$$

Gain g(x') – potential returns from an attack, e.g. Transaction Amount

$$c(x, x') \leq \varepsilon \quad \Rightarrow \quad \boxed{u_{x,y}(x') \geq \boxed{\tau}}$$

Tau is minimum "profit" level of the adversary

How we fix it: Adversarial utility

$$\boxed{u_{x,y}(x') \triangleq g(x') - c(x,x')}$$

Gain g(x') – potential returns from an attack, e.g. Transaction Amount

$$c(x,x') \le \varepsilon \quad \Rightarrow \quad \boxed{u_{x,y}(x') \ge \boxed{\tau}}$$

Tau is minimum "profit" level of the adversary

Cost constraint is replaced with "profit" constraint

How we fix it: Adversarial utility

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x'*: $20 $0.5 $14

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	MasterCard	gmail.com	UK	No

How we fix it: adversarial utility

Transaction *x*:

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes

Transaction *x'*: $20 ⇩ $0.5 ⇩ $14 ⇩

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	MasterCard	gmail.com	UK	No

$$u_{x,y}(x') = \$\,267 - \$\,34.5 = \$\,232.5$$

Contribution I: Threat Models for the Tabular Data

Cost-Bounded Objective

$$\max_{x \in \mathcal{F}(x,y)} \ell(f(x'), y) \quad \text{s.t. } c(x, x') \leq \varepsilon$$

Utility-Bounded Objective

$$\max_{x' \in \mathcal{F}(x,y)} \ell(f(x'), y) \quad \text{s.t. } u_{x,y}(x') \geq \tau$$

Both can have a financial interpretation

Contribution II: Attacks and defense methods

1. Graph search-based attack
2. Relaxation-based adversarial training

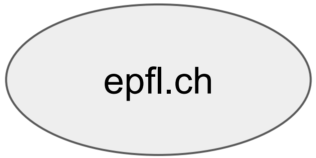
Both for cost-constrained and utility-oriented adversaries!

Evaluation of our methods

Dataset	IEEECIS Fraud detection	HomeCredit default risk	TwitterBot
Goal	Fraud detection	Loan repayment	Bot detection
Gain	Transaction amount	Loan amount	Number of followers

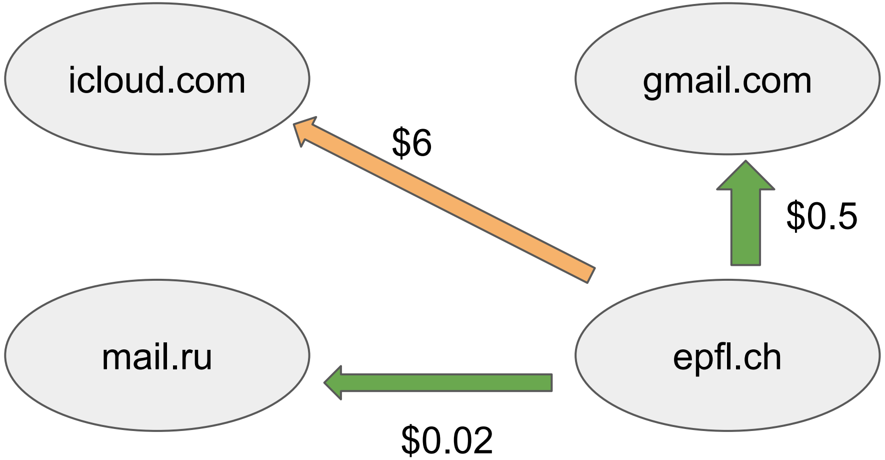
Attack Based on Greedy Graph Search

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes



Attack Based on Greedy Graph Search

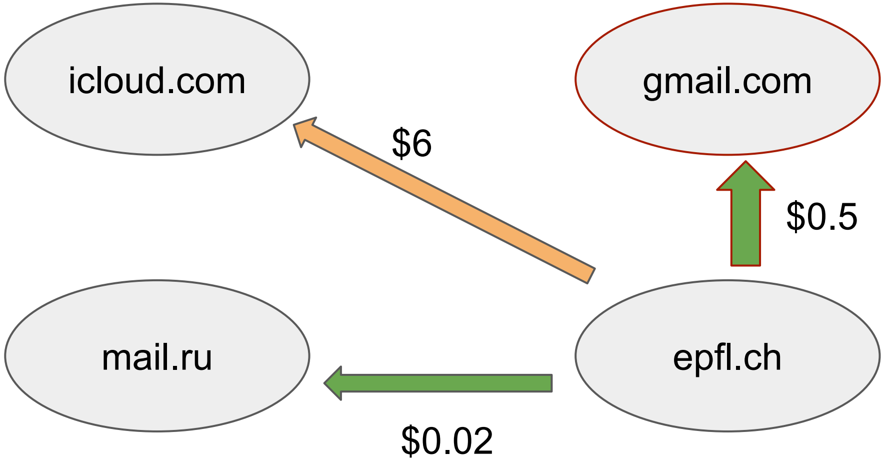
Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	epfl.ch	Italy	Yes



The attack is essentially a graph search

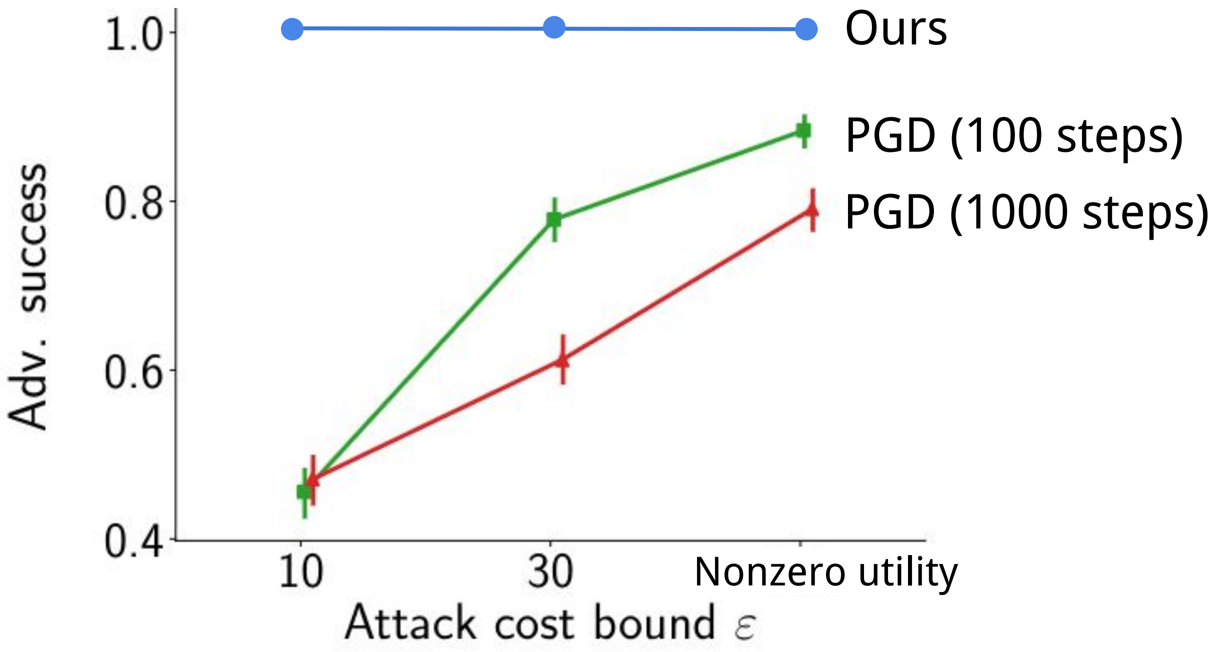
Attack Based on Greedy Graph Search

Transaction Amount	Card Type	Recipient Email	Billing country	Fraud
$267	Visa	gmail.com	Italy	No

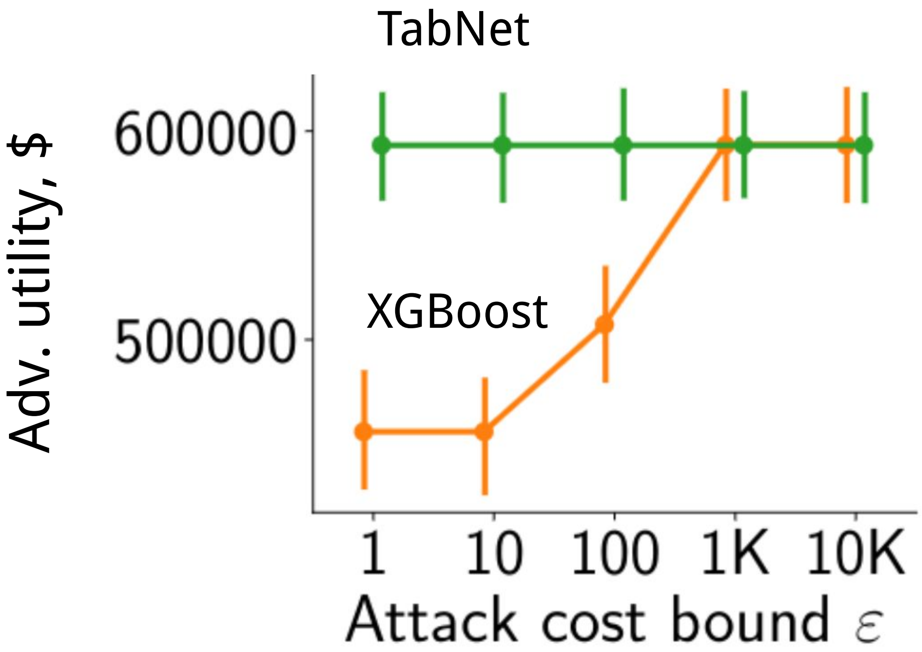


The attack is essentially a graph search

Standard attack (PGD) fails within our threat models



Attacks bring profit to the adversary and are model-agnostic!



Defenses: Adversarial Training

$$\min_{\theta} \max_{x \in \mathcal{F}(x,y)} \ell(f_{\theta}(x'), y) \quad \text{s.t. } c\left(x, x'\right) \leq \varepsilon$$

The standard way to obtain robust models is training on adversarial examples

However…

Defenses: Adversarial Training

$$\min_{\theta} \max_{x \in \mathcal{F}(x,y)} \ell(f_{\theta}(x'), y) \quad \text{s.t. } c(x, x') \leq \varepsilon$$

The standard way to obtain robust models is training on adversarial examples

However...

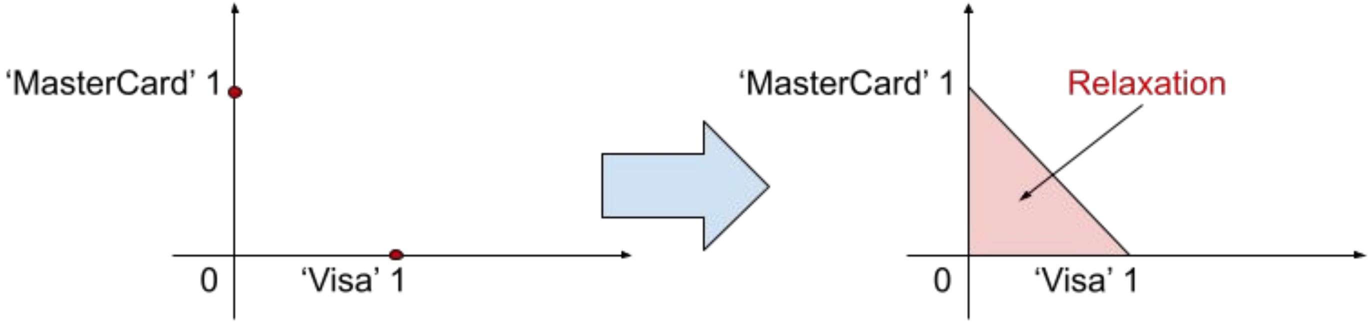
Graph-based attack takes 1-10 seconds per one sample

Constraint relaxation

$$\{\text{‘Visa’}, \text{‘MasterCard’}\} \mapsto \{[1,0], [0, 1]\}$$

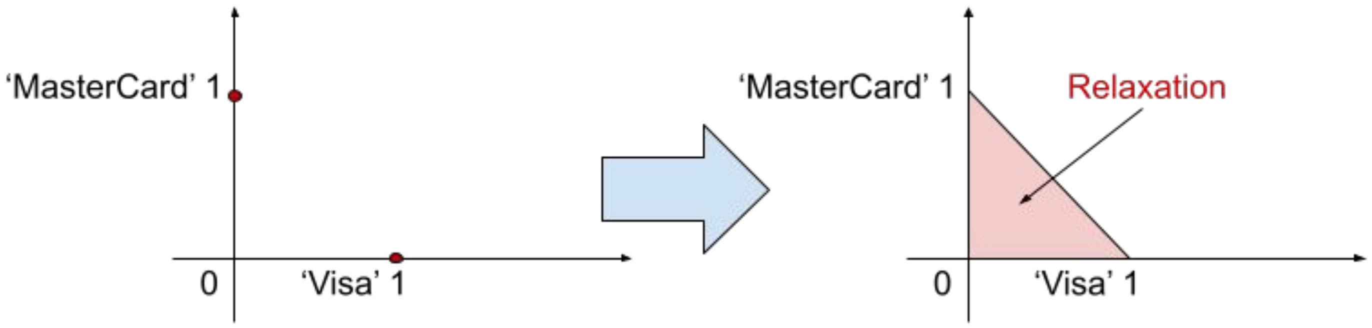
Constraint relaxation

{'Visa', 'MasterCard'} ↦ {[1,0], [0, 1]}



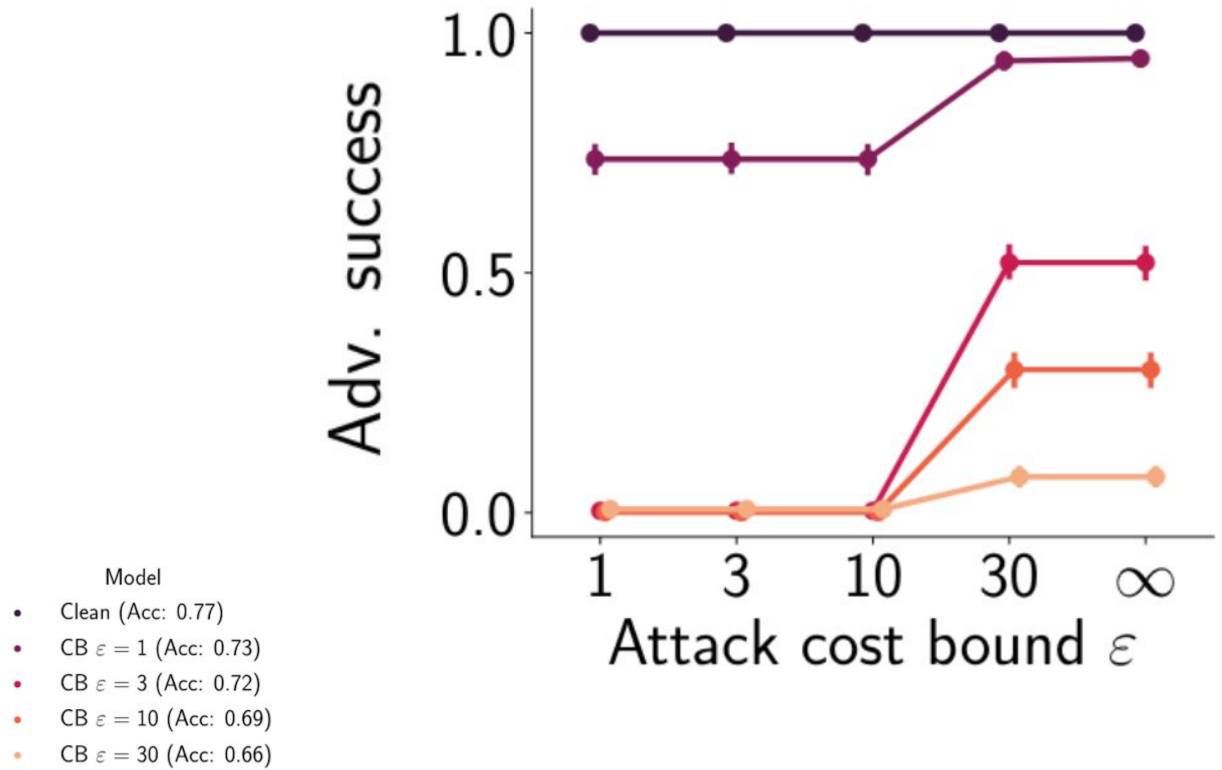
Constraint relaxation

{'Visa', 'MasterCard'} $\mapsto$ {[1,0], [0, 1]}



We relax the discrete graph search problem to continuous optimization

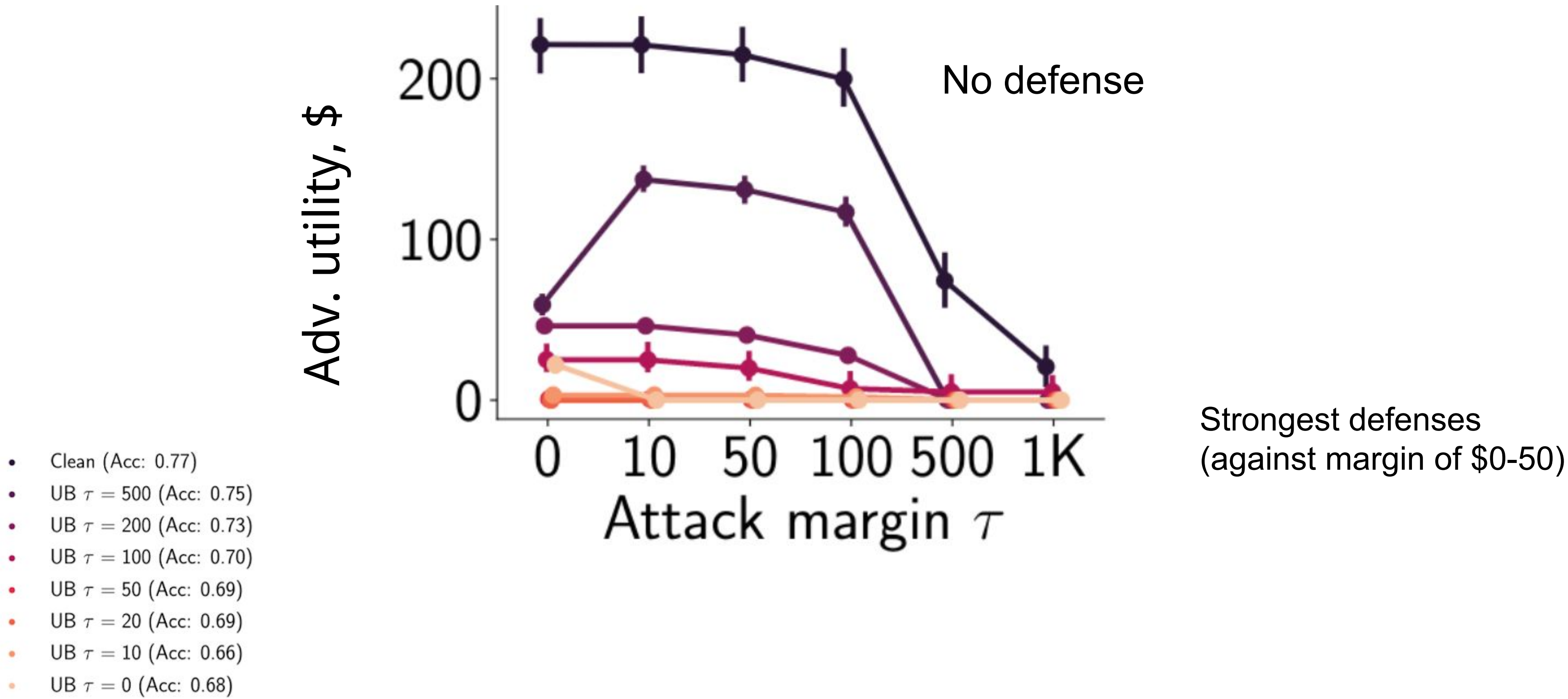
Evaluation: Cost-bounded Adversarial Training



No defense

Strongest defense
(against cost bound of 30)

Evaluation: Utility-bounded Adversarial Training

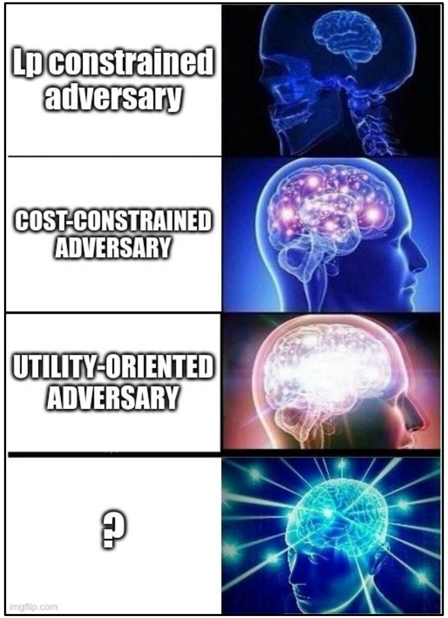


Adversarial Robustness for Tabular Data Through Cost and Utility Awareness

arxiv.org/abs/2208.13058

1. **Threat models suitable for tabular adversaries:**
   a. Cost-constrained adversary to capture financial costs
   b. Utility-oriented adversary to also recognize different profit from different examples

2. **Attacks and defenses within these threat models:**
   a. Efficient, model-agnostic graph-based attack
   b. Adversarial training as defense. The version which trains against Utility-oriented adversaries increases security in both threat models!



Metrics

Adversarial success rate - the proportion of correctly classified samples from the test set for which an adversary mounted a successful attack

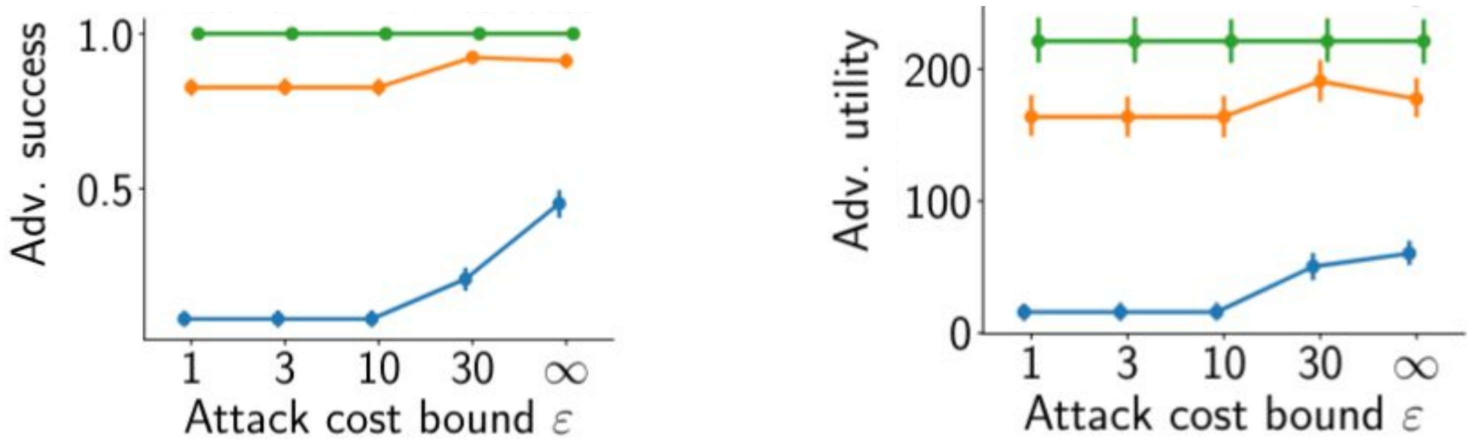
It is the principal metric for a cost-constrained adversary

Average utility - average utility of successfully generated adversarial examples

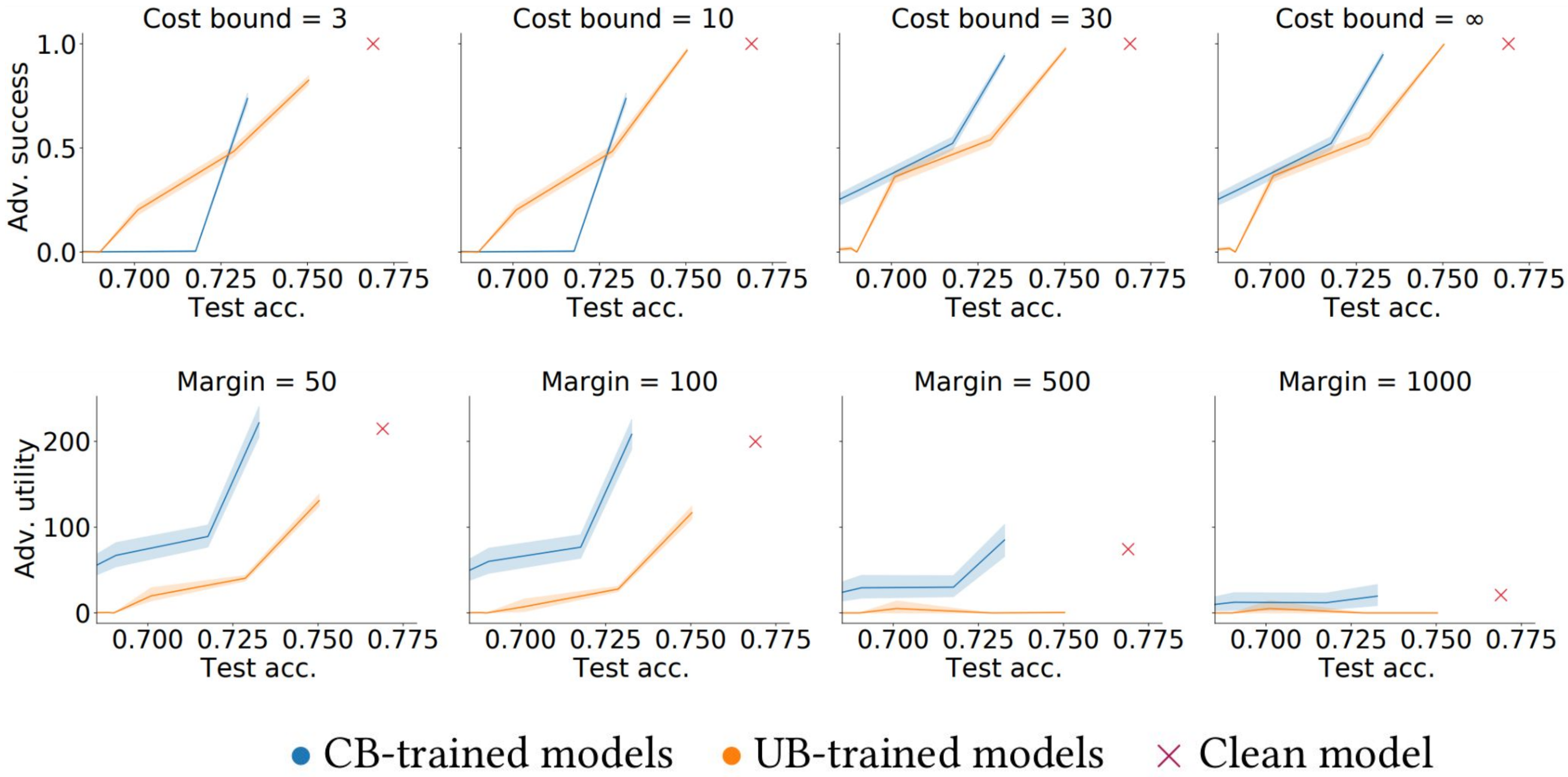
We propose it to evaluate a utility-oriented ardersary

Attacks bring profit to the adversary and are model-agnostic

IEEECIS. Model (test acc.): ● LR (0.62) ● XGBT (0.83) ● TabNet (0.77)



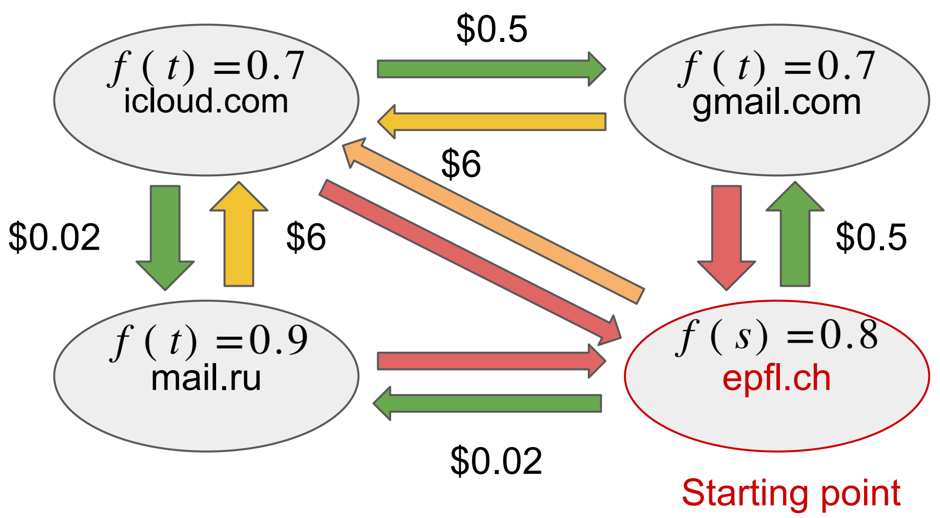
Trade-offs



Cost bound = 3
Cost bound = 10
Cost bound = 30
Cost bound = ∞
Adv. success
1.0
0.5
0.0
0.700 0.725 0.750 0.775
Test acc.
Margin = 50
Margin = 100
Margin = 500
Margin = 1000
Adv. utility
200
100
0
CB-trained models
UB-trained models
Clean model

Attacks

Transaction Amount	Card Type	Recipient Email	Billing country
$267	Visa	epfl.ch	Italy

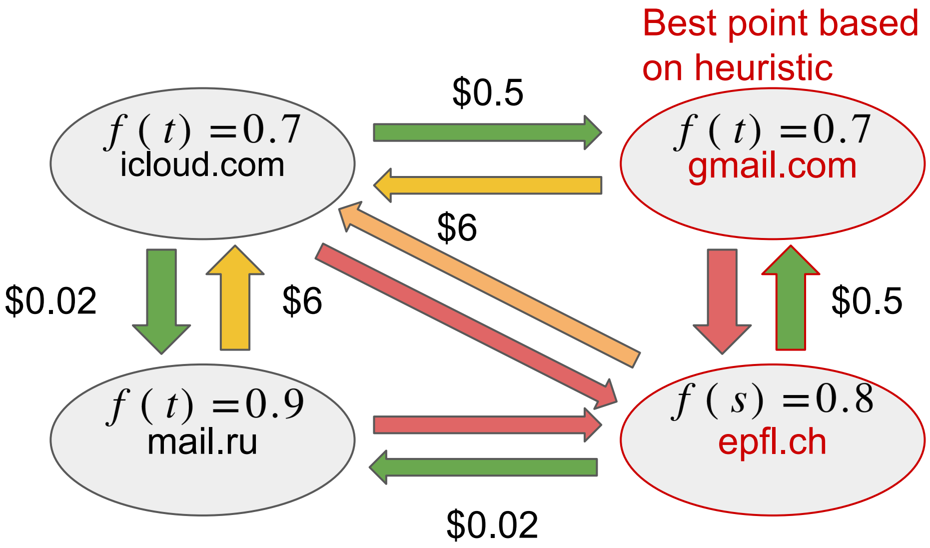


For the heuristic we stopped at

$$h(s, t) = \frac{f(t) - f(s)}{c(s, t) + \sigma}$$

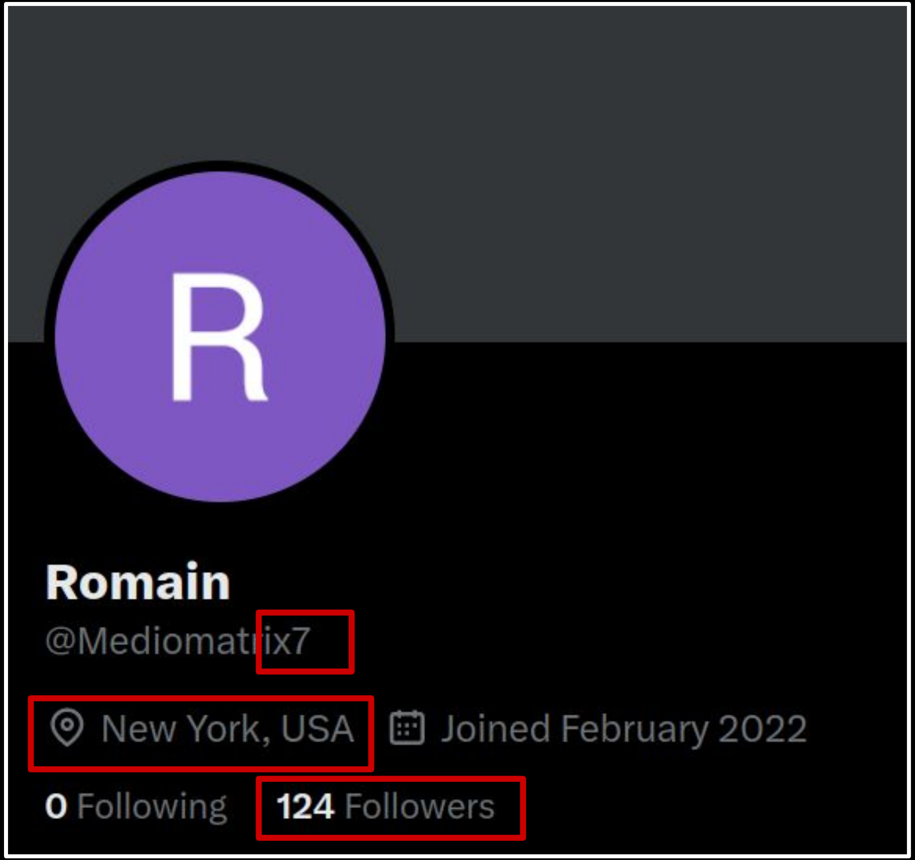
Attacks

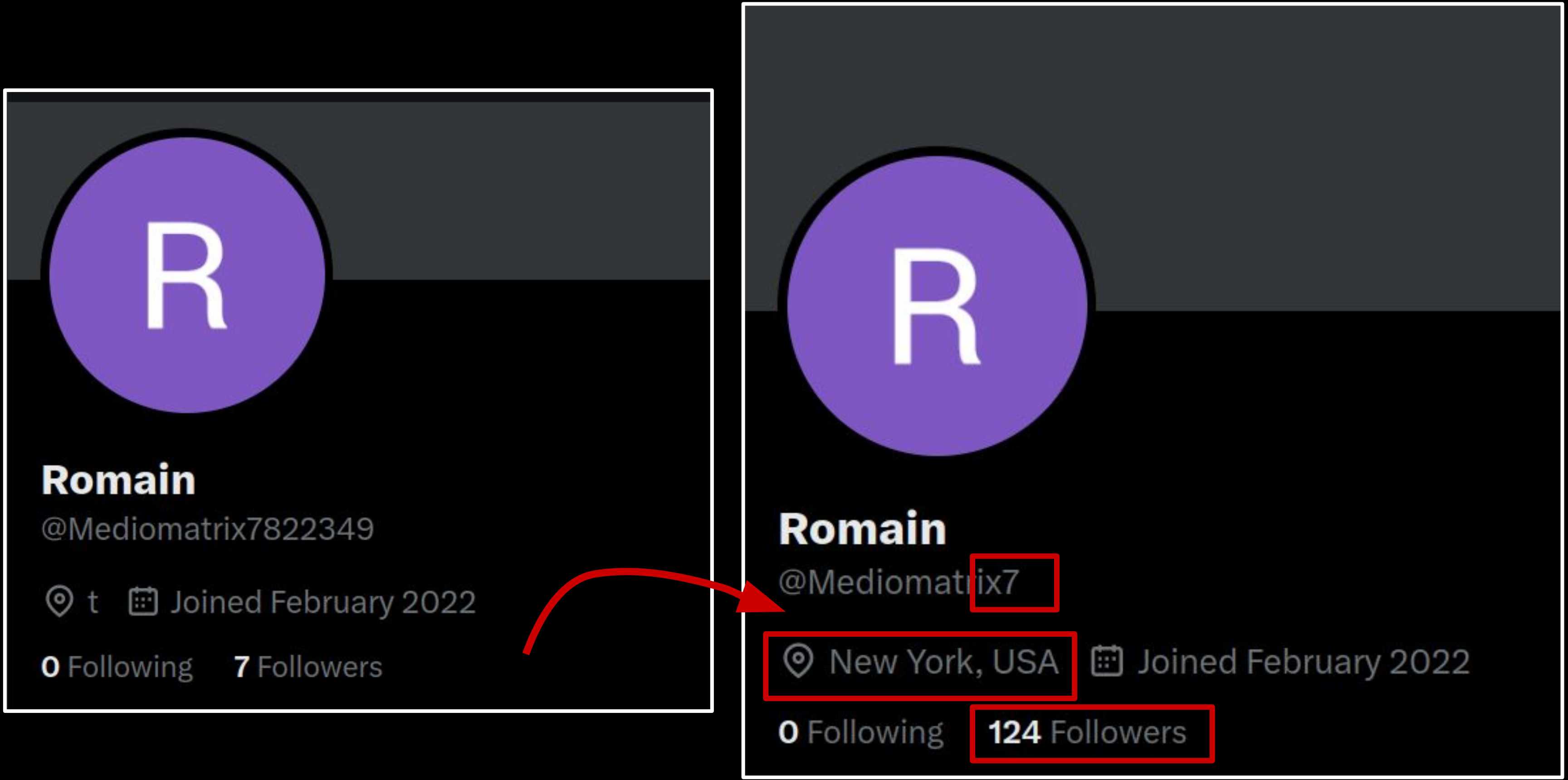
Transaction Amount	Card Type	Recipient Email	Billing country
$267	Visa	epfl.ch	Italy

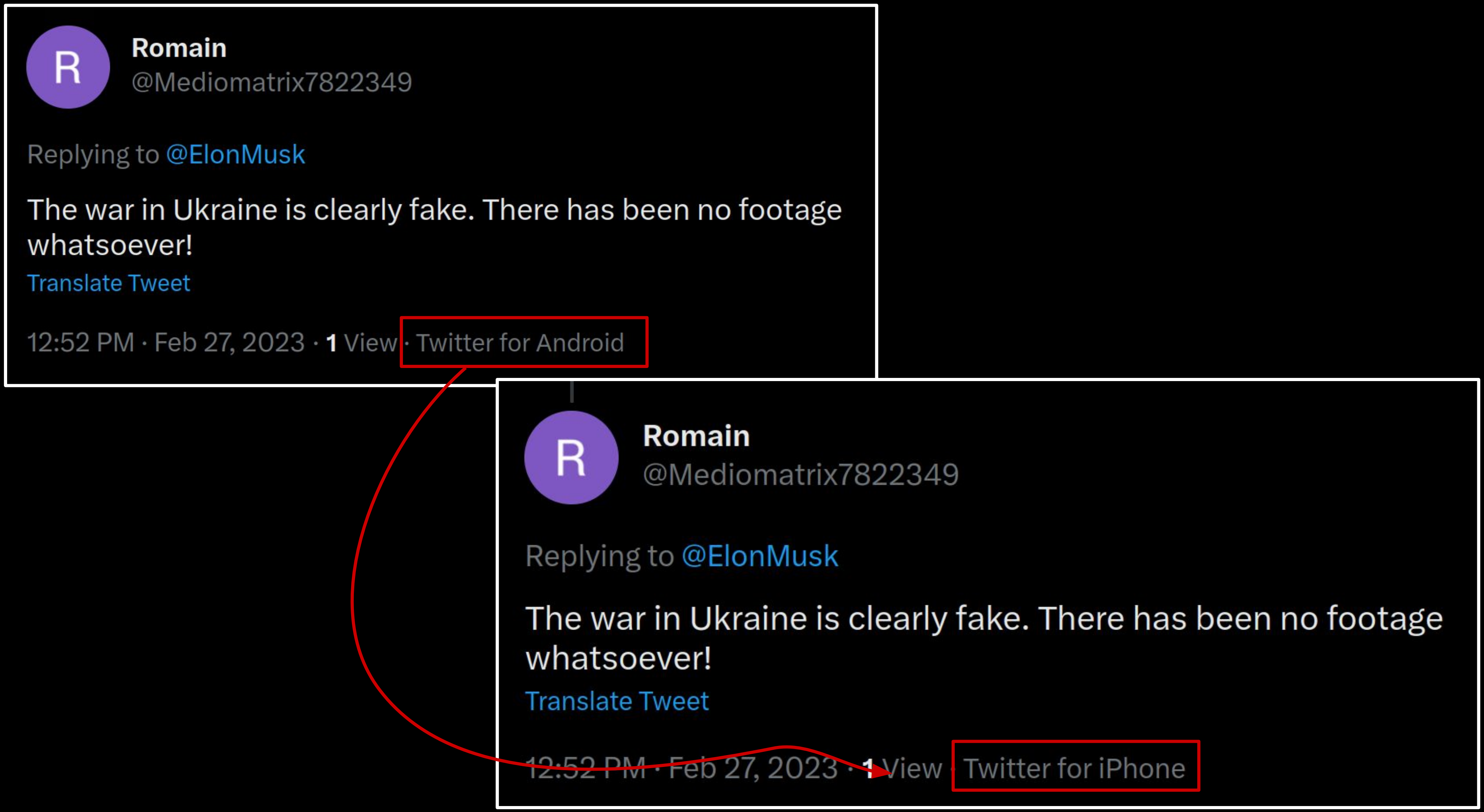


TransactionID	TransactionDT	TransactionA...	ProductCD	card1	card2	card3
3663549	18403224	31.95	W	10409	111.0	150.0
3663550	18403263	49.0	W	4272	111.0	150.0
3663551	18403310	171.0	W	4476	574.0	150.0
3663552	18403310	284.95	W	10989	360.0	150.0
3663553	18403317	67.95	W	18018		
3663554	18403323	57.95	W	12839		
3663555	18403350	87.0	W	16560		
3663556	18403387	390.0	W	15066		
3663557	18403405	103.95	W	2803		
3663558	18403416	117.0	W	12544		
3663559	18403474	261.95	W	16982		
3663560	18403504	107.95	W	9500		
3663561	18403508	335.0	W	18366		

CNT_CHIL...	AMT_INCO...	AMT_CRE...	AMT_ANN...	REGION_P...	DAYS_BIR...
0	135000.0	568800.0	20560.5	0.01885	-19241
0	99000.0	222768.0	17370.0	0.035792	-18064
0	202500.0	663264.0	69777.0	0.019101	-20038
2	315000.0	1575000.0	49018.5	0.026392	-13976
1	180000.0	625500.0	32067.0	0.010032	-13040
0	270000.0	959688.0	34600.5	0.025164	-18604
2	180000.0	499221.0	22117.5	0.0228	-16685
0	166500.0	180000.0	14220.0	0.005144	-9516
0	315000.0	364896.0	28957.5	0.04622	-12744
1	162000.0	45000.0	5337.0	0.018634	-10395
0	67500.0	675000.0	25447.5	0.0031219999999999	-23670
0	135000.0	261621.0	16848.0	0.008019	-15524
0	247500.0	296280.0	23539.5	0.018634	-12278
0	90000.0	360000.0	18535.5	0.0145199999999999	-19687


R
Romain
@Mediomatrix7
New York, USA
Joined February 2022
0 Following
124 Followers


R
Romain
@Mediomatrix7822349
t
Joined February 2022
0 Following
7 Followers
R
Romain
@Mediomatrix7
New York, USA
Joined February 2022
0 Following
124 Followers


Romain
@Mediomatrix7822349
Replying to @ElonMusk
The war in Ukraine is clearly fake. There has been no footage whatsoever!
Translate Tweet
12:52 PM · Feb 27, 2023 · 1 View · Twitter for Android
Romain
@Mediomatrix7822349
Replying to @ElonMusk
The war in Ukraine is clearly fake. There has been no footage whatsoever!
Translate Tweet
12:52 PM · Feb 27, 2023 · 1 View · Twitter for iPhone

Adversarial Cost

min.	avg.	max.
\$0.02	\$35.7	\$281.6