UMass Amherst | College of Information & Computer Sciences

Defeating DNN-Based Traffic Analysis Systems in Real-Time With Blind Adversarial Perturbations

Milad Nasr, Alireza Bahramali, Amir Houmansadr
University of Massachusetts, Amherst

USENIX Security 2021

Encryption Is Ubiquitous

The content of the network traffic is encrypted!

amazon cloudfront

Telegram

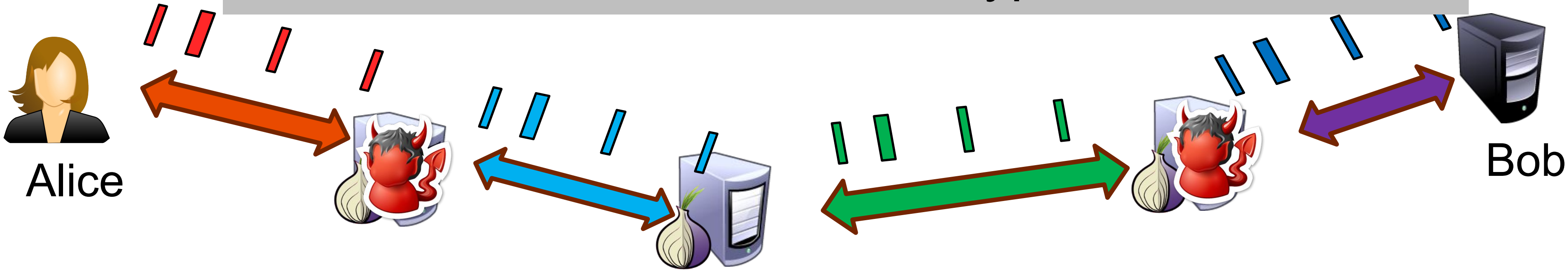
CLOUDFLARE®

facebook

Traffic Analysis: using the metadata of the traffic to do analysis

Example traffic analysis on Tor

Attackers can not link flows using packet contents due to onion encryption



But they can match traffic patterns as Tor is designed to be low-latency

State-of-the-art traffic analysis techniques leverage DNNs

- **Detection rate in traffic correlation improved from 0.2 to 0.9 by using neural networks *[Nasr' 18]***
- **Accuracy in website fingerprinting improved from 60% to 90% by using neural networks *[Bhat' 18 ,Sirinam 19',...]***

The Threat of Adversarial Examples

- Neural networks are vulnerable to the small perturbations to the input a.k.a adversarial examples

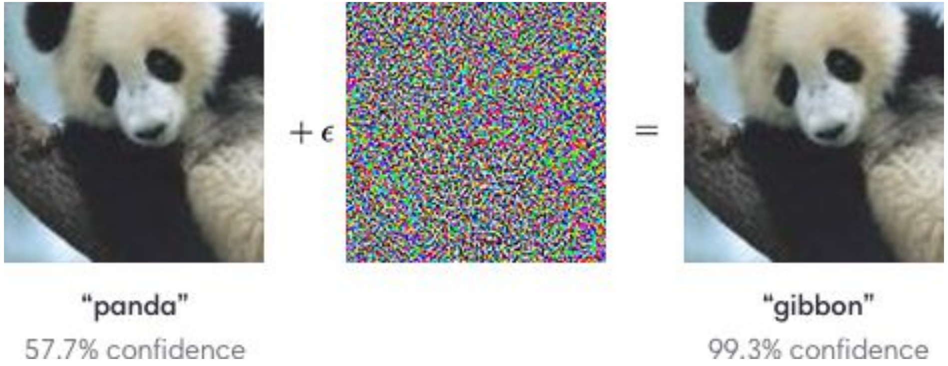


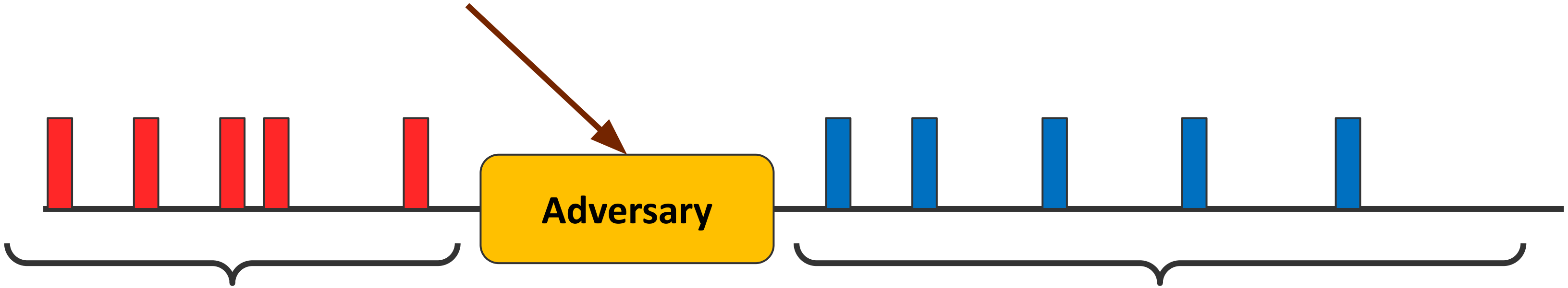
Image from openai.com

Our Goal:

Whether and how adversarial examples can be applied on DNN-based traffic analysis systems

Applying Adversarial Examples on Traffic Analysis Applications Is Very Challenging

Perturbations should be applied in **real-time**

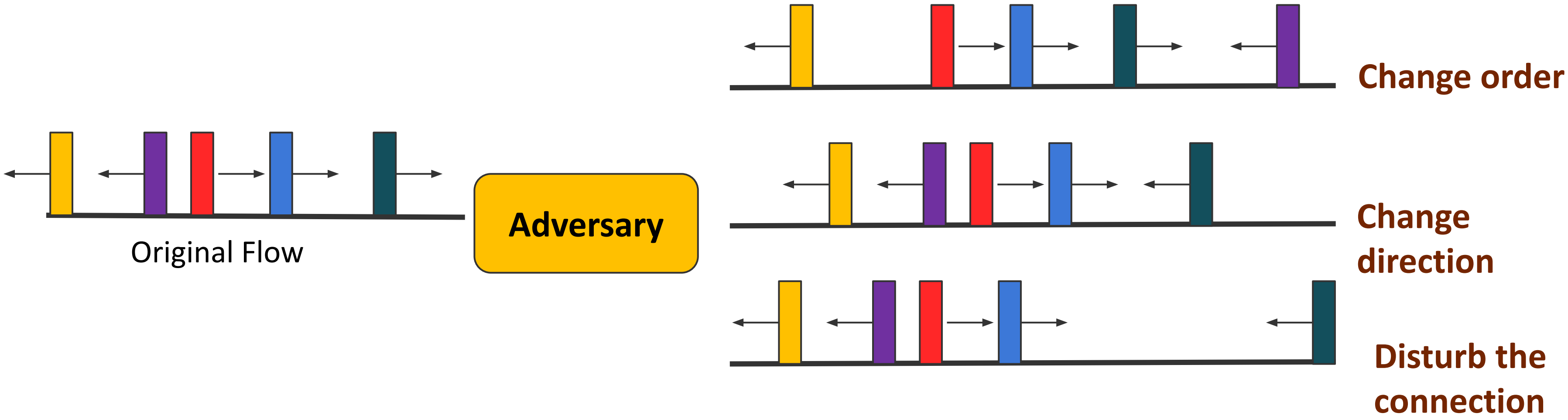


Adversary doesn't know the pattern of coming packets!

Adversary cannot change the pattern after sending!

Adversary is Blind!

Applying Adversarial Examples on Traffic Analysis Applications Is Very Challenging



Network flows should cannot be modified arbitrarily. Protocol specifications and constraints should be preserved!

Overview of Our Contributions

- A **generic** framework for applying **blind** adversarial perturbations on live traffic analysis systems
- Implemented a Tor pluggable transport called BLANKET
- We apply the attack on recent traffic analysis works

Our generic framework

$$\arg\min_{\boldsymbol{\delta}} \forall \boldsymbol{x} \in D^S : f(\boldsymbol{x}+\boldsymbol{\delta}) \neq f(\boldsymbol{x})$$

$$s.t. \quad x + \delta \in C$$

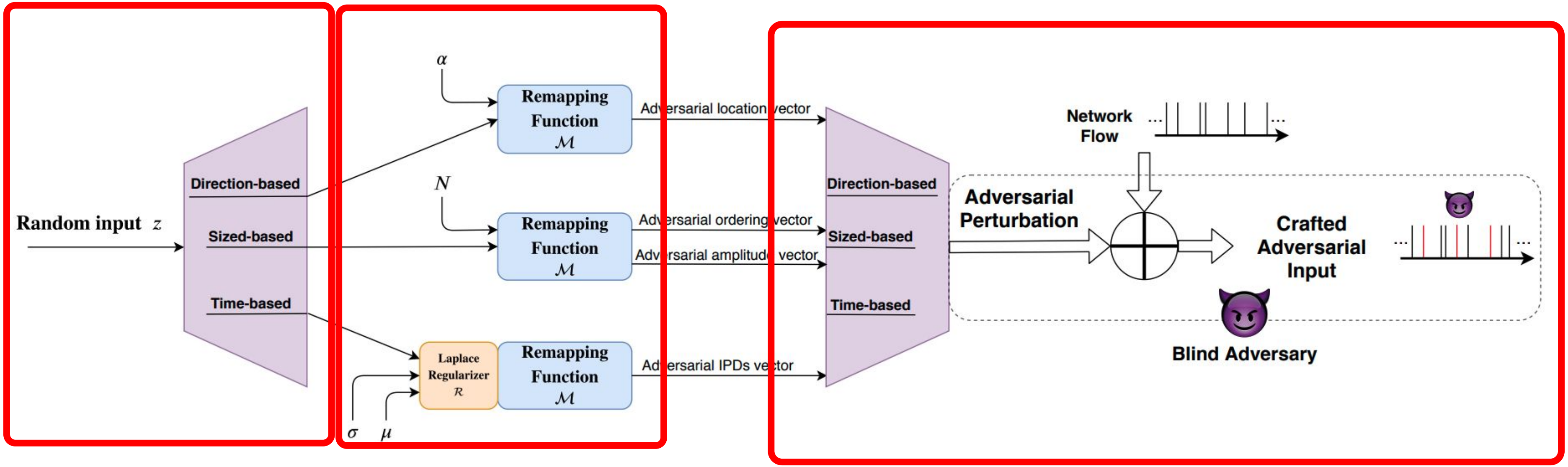
Perturbation

Traffic pattern

Target Model

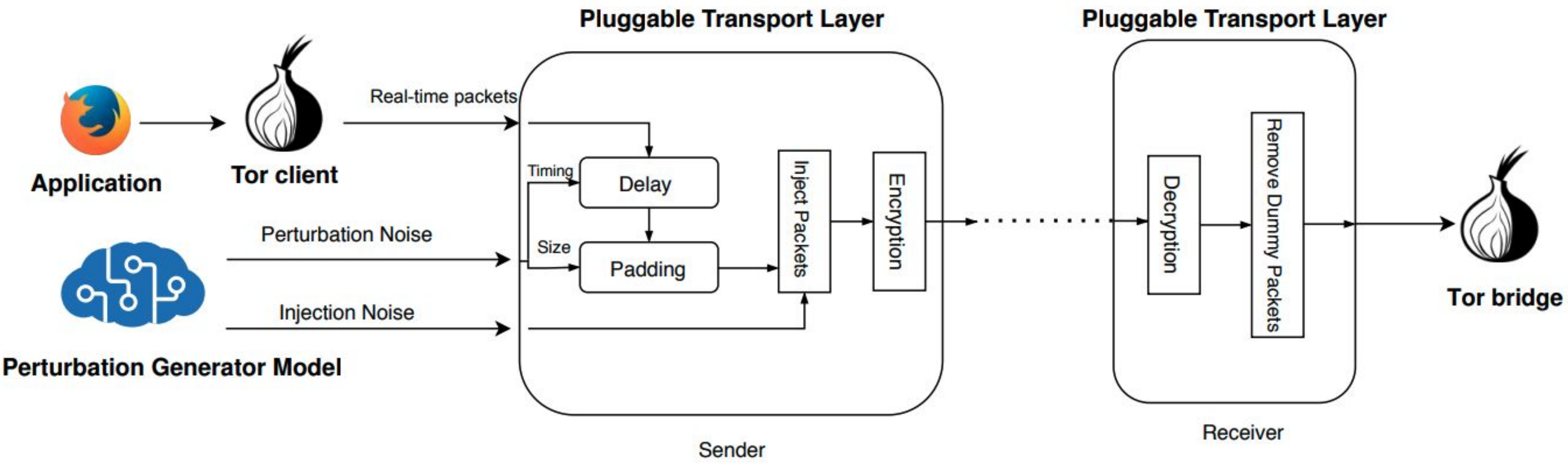
Constraints (packet sizes, timing, protocol specifications)

Overview



$$\arg\max_{G} \mathop{\mathbb{E}}_{z\sim uniform(0,1)} [(\sum_{\boldsymbol{x}\in\mathcal{D}^S} l(f(\mathcal{M}(\boldsymbol{x},G(z))),f(\boldsymbol{x}))) + \mathcal{R}(G(z))]$$

Experimental Setup



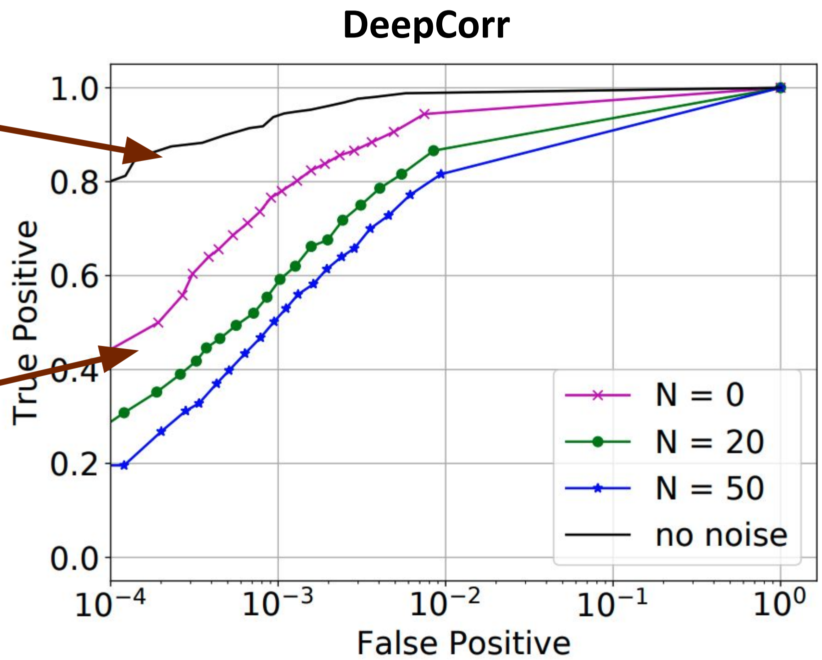
Experimental Setup

Target Systems:

- **DeepCorr**: Traffic correlation (Timing, Sizes and Directions)*[Nasr 19']*
- **Var-CNN**: Website fingerprinting (Timing, Directions and statistical informations)*[Bhat 18']*
- **Deep Fingerprinting**: Website fingerprinting (Timing, Directions)*[Sirinam 18']*

Using BLANKET To Defeat Traffic Correlation

Performance without an attack



By increasing the perturbation intensity the performance drops significantly

Deep learning based traffic correlation methods are **vulnerable** to BLANKET

Using BLANKET To Defeat Website Fingerprinting

Large Drop in Average Accuracy
for specific target

VarCNN 93% Average accuracy (Timing and Sizes)

α, μ, σ,	BW Overhead (%)	$\mathcal{A}$: SU-DU (%)	Max ST-DU (#, %)
20, 0, 5	0.04	79.0	–, 100.0
100, 0, 10	2.04	83.9	–, 100.0
500, 0, 20	11.11	97.0	–, 100.0
1000, 0, 30	25.0	98.6	–, 100.0
2000, 0, 50	66.66	99.0	–, 100.0

DF 92% Average accuracy (Directions)

α	Bandwith Overhead (%)	SU-DU (%)	Max ST-DU (#, %)
20	0.04	24.2	–, 100.0
100	2.04	49.6	–, 100.0
500	11.11	91.8	–, 100.0
1000	25.0	95.7	–, 100.0
2000	66.66	97.7	–, 100.0

Large Drop in Average Accuracy

Can we counter BLANKET?

Traffic Correlation

Adversary Strength	Original	No Def	Madry et al. [34]	IGR [48]	RC [7]	Our Defense
$\mu = 0, \sigma = 10$	79%	63%	70%	62%	63%	74%
$\mu = 0, \sigma = 50$	79%	21%	25%	23%	22%	32%
$\mu = 0, \sigma = 100$	79%	13%	18%	13%	14%	23%

Website Fingerprinting

Adversary Strength	Original	No Def	Madry et al. [34]	IGR [48]	RC [7]	Our Defense
$\alpha = 20$	92%	60%	84%	62%	54%	84%
$\alpha = 100$	92%	28%	48%	23%	23%	60%
$\alpha = 500$	92%	8%	19%	2%	7%	24%

Our adversarial perturbation mechanism is hard to protect against!

Comparing BLANKET With Traditional Attacks on Traffic Analysis

Name	Bandwidth Overhead	Latency OverHead	Accuracy
WTF-PAD (DF)	64%	0%	3%
Walkie-Talkie (DF)	31%	36%	5%
BLANKET (DF)	**25%**	**0%**	**1%**
WTF-PAD (VarCNN)	27%	0%	88%
BLANKET (VarCNN)	**25%**	**0%**	**2%**

While there exist other attacks on traffic analysis, BLANKET outperforms all regarding latency, overhead, and performance

Conclusions

- A **generic** framework for applying **blind** adversarial perturbations on live traffic analysis systems
- Implemented a Tor pluggable transport called BLANKET
- We apply the attack on recent traffic analysis works

UMass Amherst | College of Information & Computer Sciences

COMPUTING FOR THE COMMON GOOD

Massbrowser.cs.umass.edu
@massbrowser
Join us!

Milad Nasr
milad@cs.umass.edu

https://people.cs.umass.edu/~milad/ | @srxzr

UMass Amherst | College of Information & Computer Sciences

COMPUTING FOR THE COMMON GOOD

References:

Nasr, Milad, Alireza Bahramali, and Amir Houmansadr. "Deepcorr: Strong flow correlation attacks on tor using deep learning." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

Bhat, Sanjit, et al. "Var-CNN: A Data-Efficient Website Fingerprinting Attack Based on Deep Learning." Proceedings on Privacy Enhancing Technologies 1: 19.

Sirinam, Payap, et al. "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

Packet Timing Constraints

$$\mathcal{M}^{T}(\boldsymbol{x}, G(z), \mu, \sigma) = \boldsymbol{x} + \frac{G(z) - \max(\overline{G(z)} - \boxed{\mu}, 0) - \min(\overline{G(z)} + \mu, 0)}{\mathrm{std}(G(z))} \min(\mathrm{std}(G(z)), \boxed{\sigma})$$

Average of distributions

Standard deviation of distributions

Packet Size Constraints

Algorithm 3 Size remapping function

$a \leftarrow G(z)$

$\boldsymbol{x} \leftarrow$ training input
$N \leftarrow$ maximum sum of added sizes
$n \leftarrow$ maximum added size to each packet
$s \leftarrow$ cell sizes

for i in argsort(-a) **do**
  if $N \leq 0$ **then**
    break
  end if
  $\delta = \lfloor \min(s\frac{a[i]}{s}, n, N) \rfloor$
  $N = N - \delta$
  $\boldsymbol{x}[i] = \boldsymbol{x}[i] + \delta$
end for
return $\boldsymbol{x}$

Transferability

Traffic Correlation (Alexnet to DeepCorr)

Adversary Strength	Transferability (%)
$N = 10$	75.32
$N = 20$	83.11
$N = 50$	90.24

Website Fingerprinting (DF to VarCNN)

Adversary Strength	Transferability (%)
$\alpha = 100$	30.65
$\alpha = 500$	85.90
$\alpha = 1000$	96.53