32ND USENIX SECURITY SYMPOSIUM

Subverting Website Fingerprinting Defenses with Robust Traffic Representation

Meng Shen[1], Kexin Ji[1], Zhenbo Gao[1], Qi Li[2], Liehuang Zhu[1], and Ke Xu[2]
[1]Beijing Institute of Technology (BIT), China
[2]Tsinghua University, China
{shenmeng, jikexin, liehuangz}@bit.edu.cn, gaozhenbo07@foxmail.com, {qli01, xuke}@tsinghua.edu.cn

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

清華大學
Tsinghua University

Aug. 9, 2023

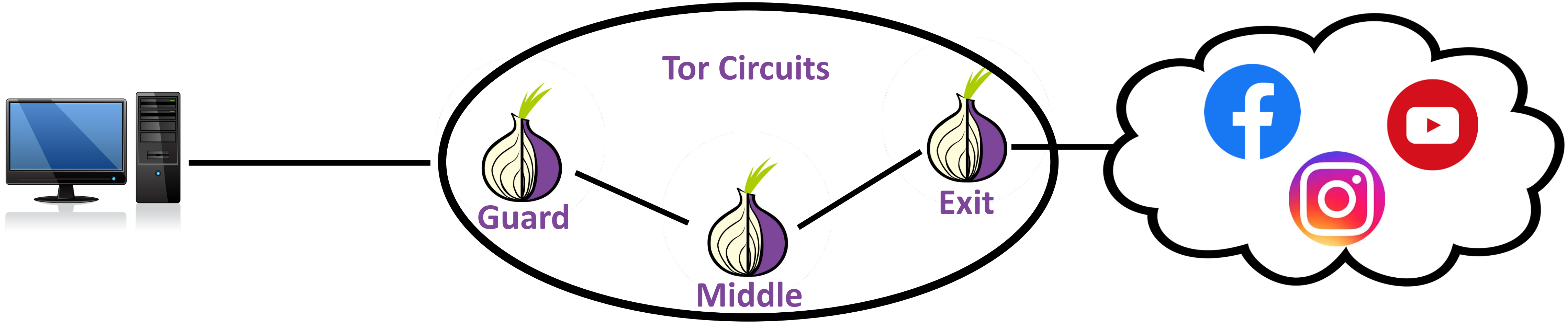
Anonymous Communication

Anonymous communication aims to **hide the identity or communication relationship of both parties** in an open network environment, such as **Anonymous Browsing**, **Secure Communication**, and **File Sharing**.

https://theconversation.com
https://www.slideshare.net
https://www.filecloud.com

Anonymous Communication through Tor

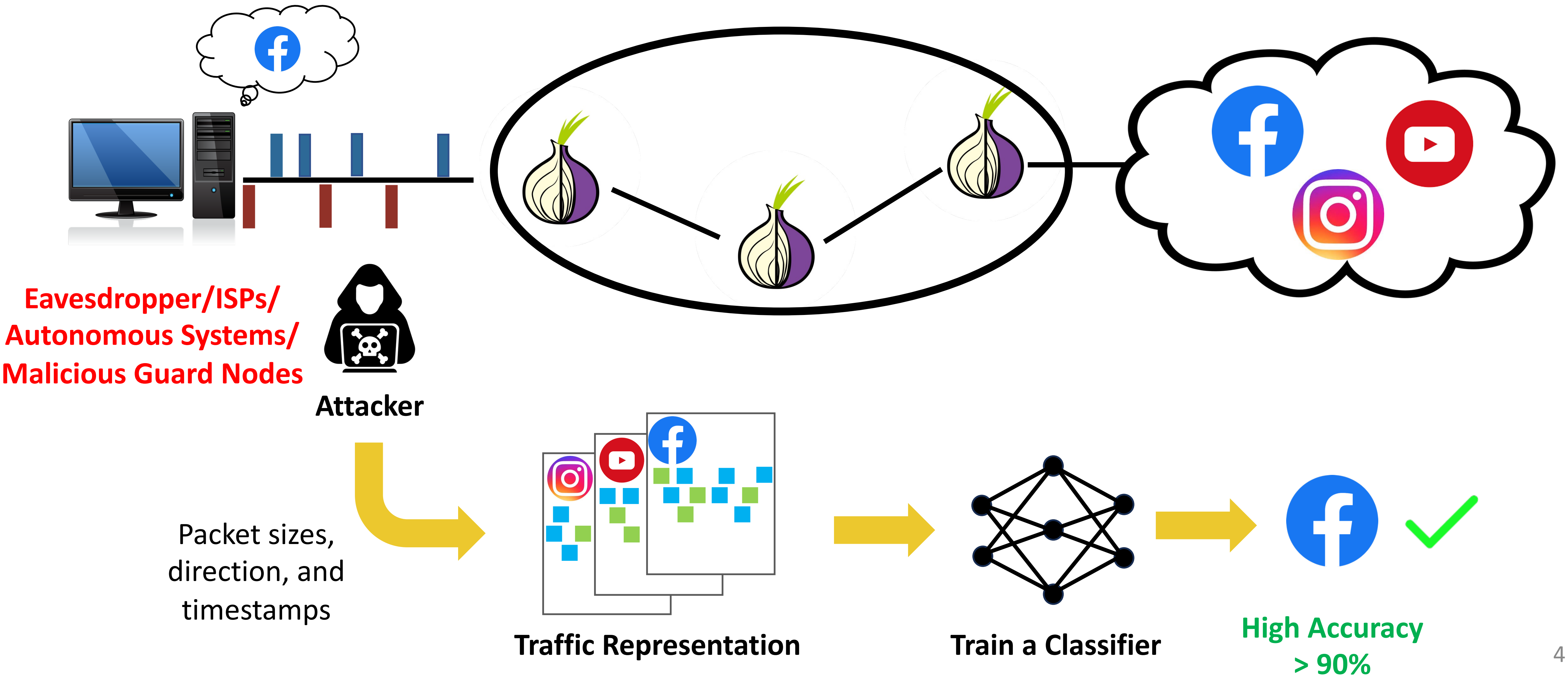
Tor has been **widely used** as an anonymous communication tool to prevent users from being tracked, monitored and censored



Tor routes traffic across a path of **three volunteer-operated nodes** (called **circuits**) with layered encryption

Website Fingerprinting (WF)

WF Attackers try to infer the website that a user is visiting **without breaking the encryption**

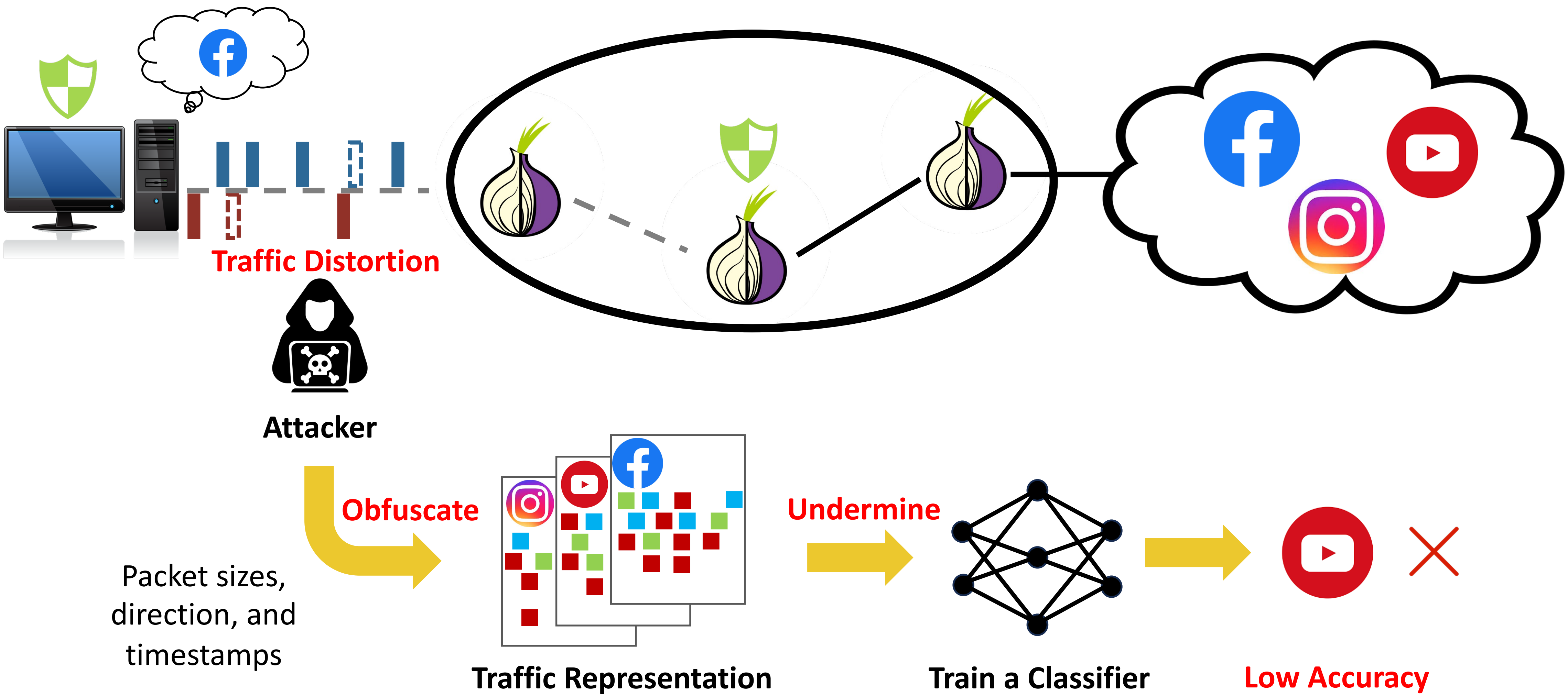


Existing WF Attacks

Category	Feature Granularity	Attacks	Traffic Representation
Traditional Machine Learning	Coarse-grained Statistical Feature	k-NN[1]	Statistical Feature Collection: Mean, Median, Sum, Maximum, ..., Minimum of Packet Sizes, Packet Intervals, ...
		CUMUL[2]	
		k-FP[3]	
Deep Learning	Fine-grained Per-packet Feature	AWF[4]	Packet Direction: +1, -1, -1, -1, +1, -1, ... Packet Timing: 0.13, 0.22, 0.24, … Timing with Direction: +0.13, -0.22, -0.24, ... Inter-Packet Time: 0.13, 0.09, 0.02,...
		DF[5]	
		Var-CNN[6]	
		Tik-Tok[7]	

[1] Wang, et al. Effective attacks and provable defenses for website fingerprinting. USENIX 2020.
[2] Panchenko, et al. Website fingerprinting at internet scale. NDSS 2016.
[3] J. Hayes, et al. k-fingerprinting: A robust scalable website fingerprinting technique. USENIX 2016.
[4] Rimmer, et al. Automated website fingerprinting through deep learning. NDSS 2018.
[5] Sirinam, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. CCS 2018.
[6] Bhat, et al. Var-cnn: A data-efficient website fingerprinting attack based on deep learning. PETS 2019.
[7] Rahman, et al. Tik-tok: The utility of packet timing in website fingerprinting attacks. PETS 2020.

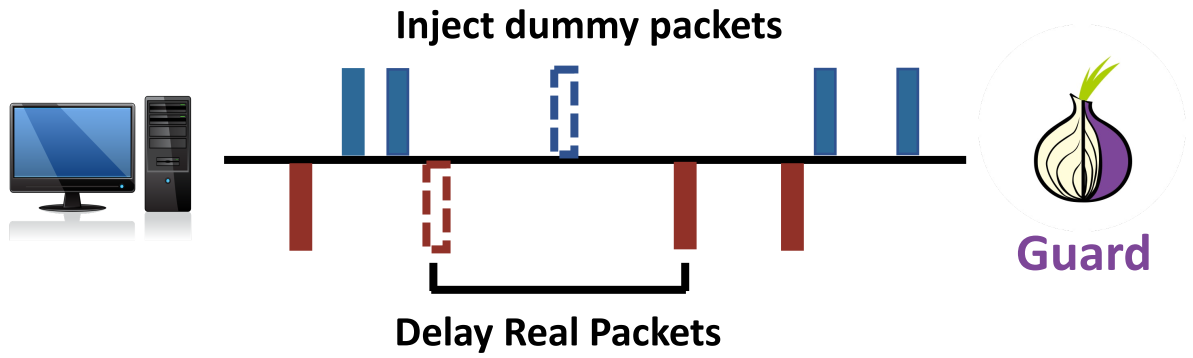
WF Defense


Traffic Distortion
Attacker
Obfuscate
Packet sizes, direction, and timestamps
Traffic Representation
Undermine
Train a Classifier
Low Accuracy

WF Defense

Disturbing Traffic

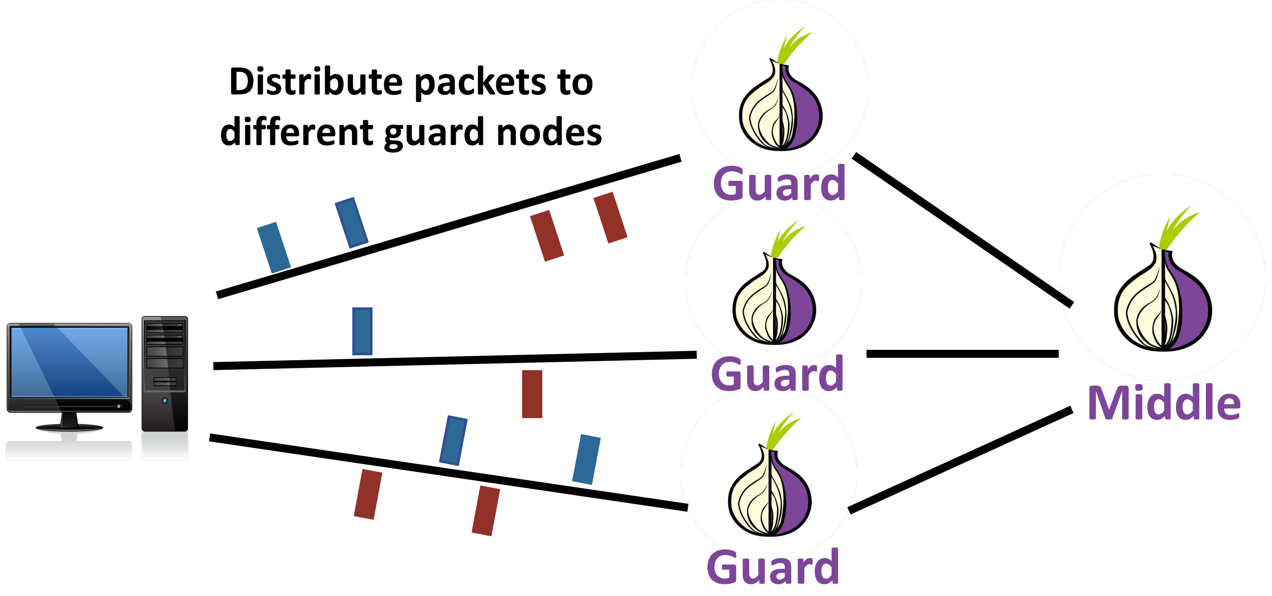
- Tamaraw [Wang, USENIX’14]
- **WTF-PAD [Juarez, ESORICS’16]**
- **Walkie-Talkie [Wang, USENIX’17]**
- **FRONT [Gong, USENIX’20]**
- **Blanket [Nasr, USENIX’21]**
- **RegulaTor [Holland, PETS’22]**



Introduce bandwidth or time overhead

Splitting Traffic

- TrafficSliver [la Cadena, CCS’20]



Introduce implementation overhead

Goal and Challenges

Goal: Fingerprint the Tor traffic accurately even under existing WF defenses

Challenges:

- **Is there a robust traffic representation that can less affected by existing traffic disturbing or splitting strategies?**
- **How to design an effective WF attack achieving high accuracy against existing defenses?**

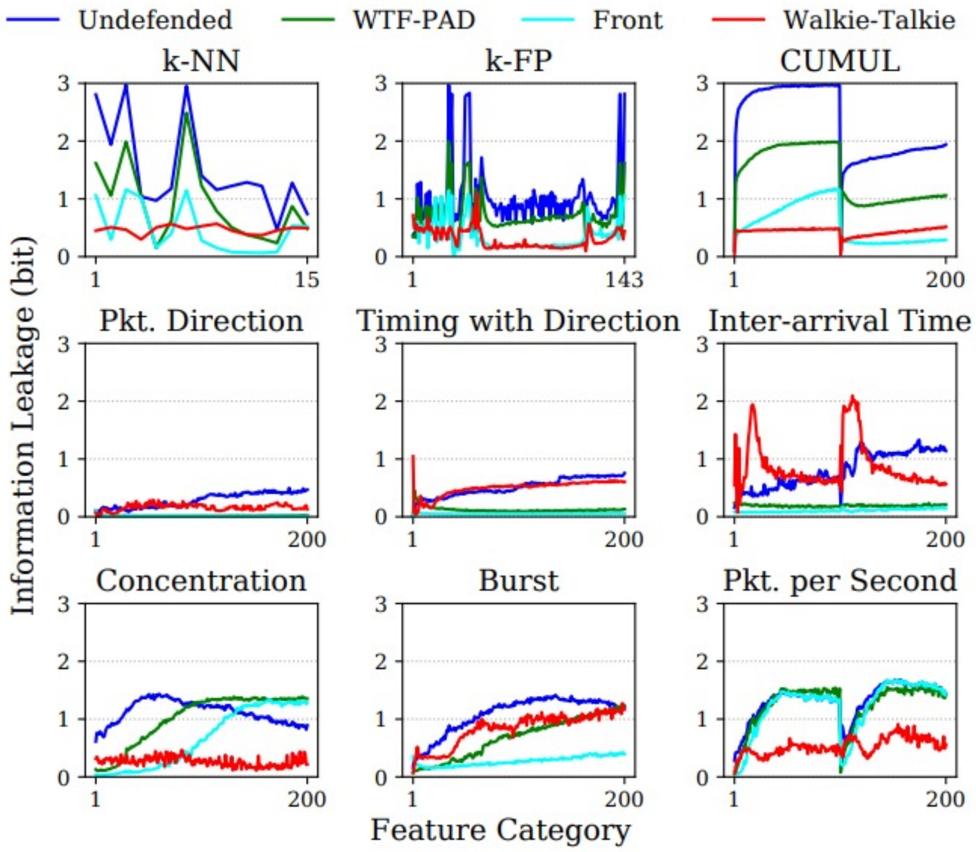
Contribution

- Propose a robust traffic representation called Traffic Aggregation matrix (TAM)
- Present a novel WF attack Robust Fingerprinting (RF)
- Demonstrate RF is superior to SOTA WF attacks in closed- and open-world scenarios
- Develop a countermeasure against RF which more effective to reduce its accuracy

Feature Spaces Exploration

Information Leakage Analysis

- Measure the amount of information attackers can learn from the key feature to fingerprint the Tor traffic
- Typical defenses: WTF-PAD, Front and Walkie-Talkie



[1] Shuai Li, et al. Measuring information leakage in website fingerprinting attacks and defenses. CCS 2018.

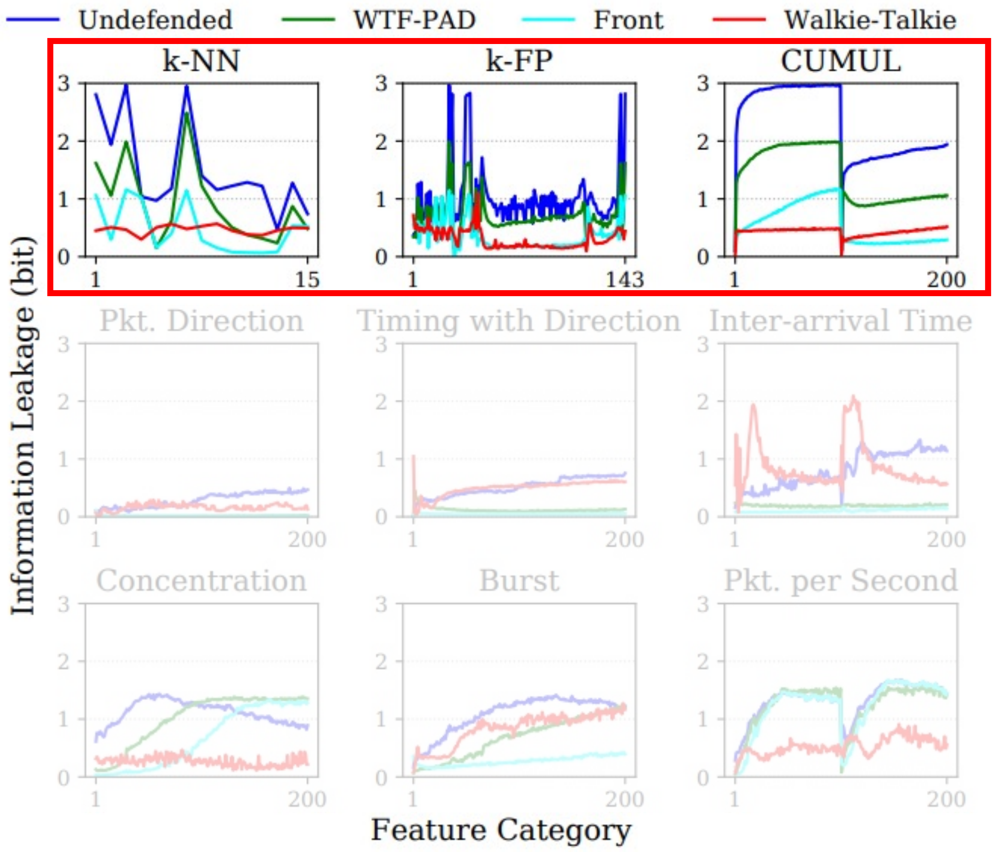
Feature Spaces Exploration

Information Leakage Analysis

- Measure the amount of information attackers can learn from the key feature to fingerprint the Tor traffic
- Typical defenses: WTF-PAD, Front and Walkie-Talkie

Coarse-grained statistical features

- The information leakage is hidden by different defenses
- Trivial contributions to website fingerprinting



[1] Shuai Li, et al. Measuring information leakage in website fingerprinting attacks and defenses. CCS 2018.

Feature Spaces Exploration

Information Leakage Analysis

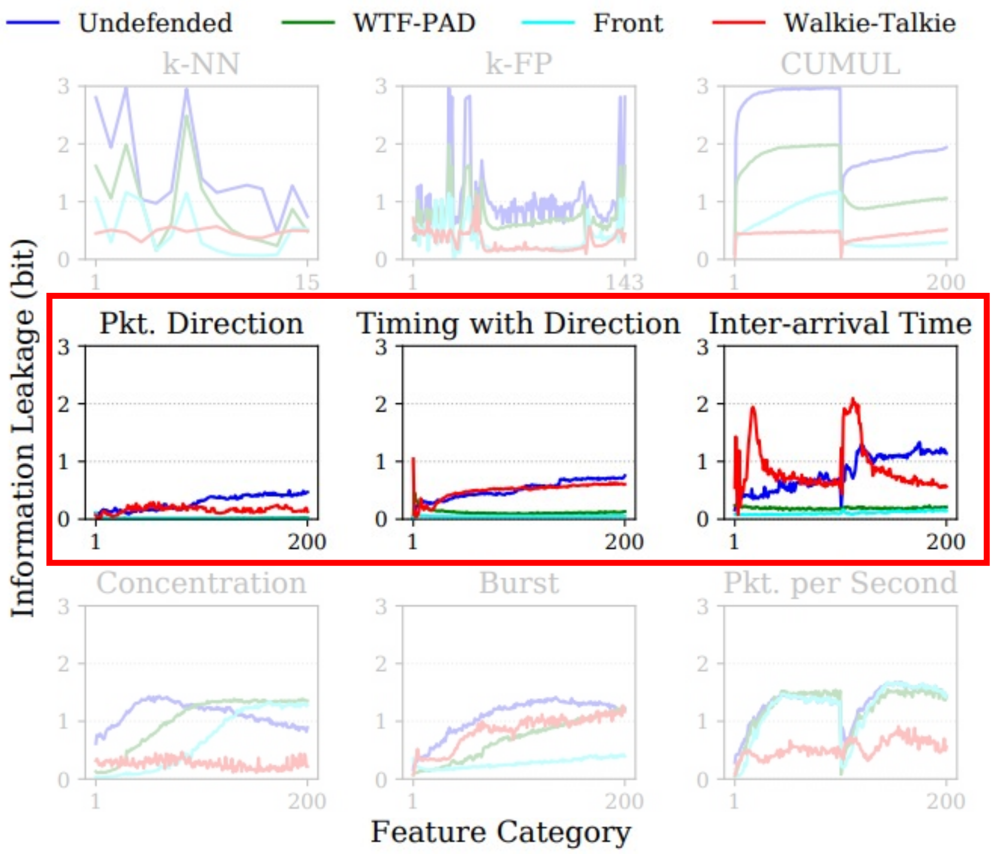
- Measure the amount of information attackers can learn from the key feature to fingerprint the Tor traffic
- Typical defenses: WTF-PAD, Front and Walkie-Talkie

Coarse-grained statistical features

- The information leakage is hidden by different defenses
- Trivial contributions to website fingerprinting

Fine-grained per-packet feature sequences

- Affected by defenses due to the randomness in packets padding and delaying



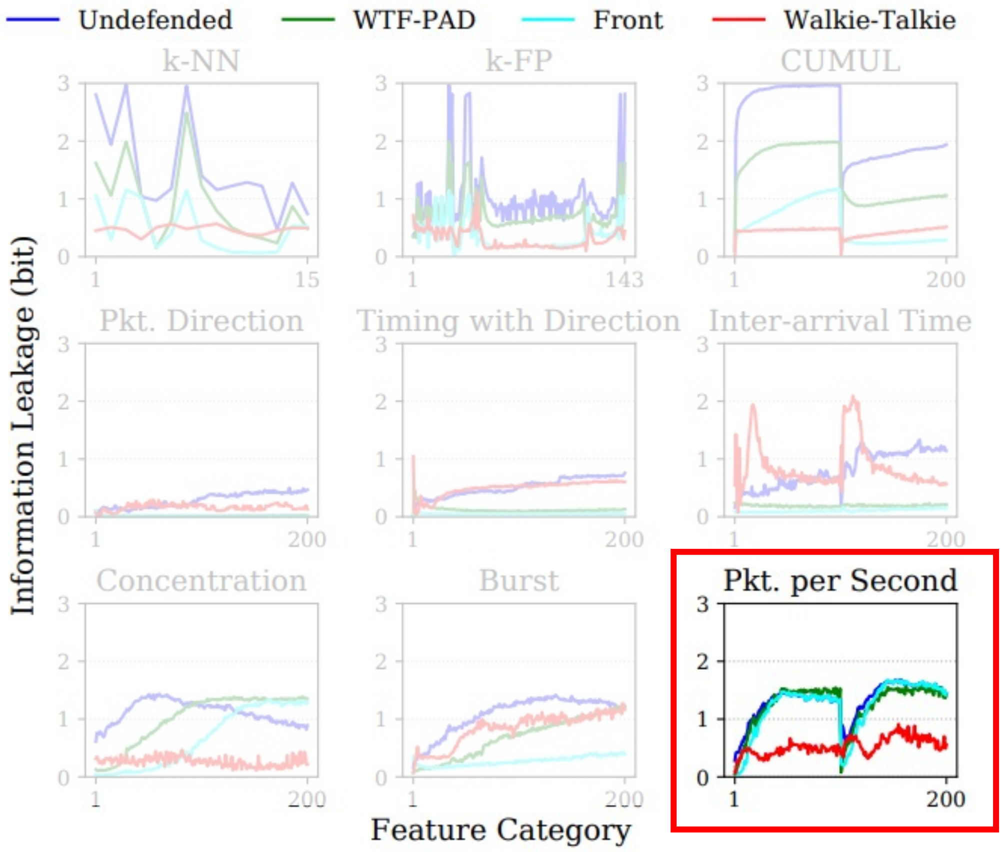
A feature with an intermediate granularity?

[1] Shuai Li, et al. Measuring information leakage in website fingerprinting attacks and defenses. CCS 2018.

Feature Spaces Exploration (Cont'd)

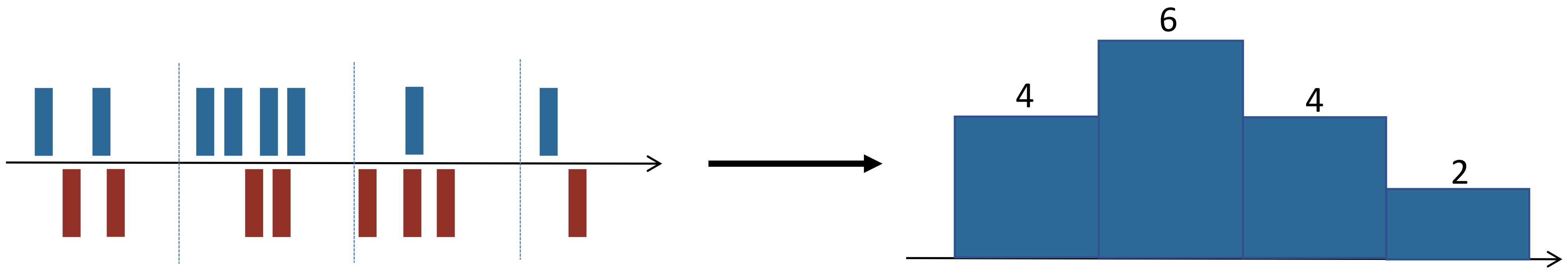
Packet-per-second

- **Uncovered by WTF-PAD and Front**
- A potential **robust representation** which is cannot be easily disturb by defenses

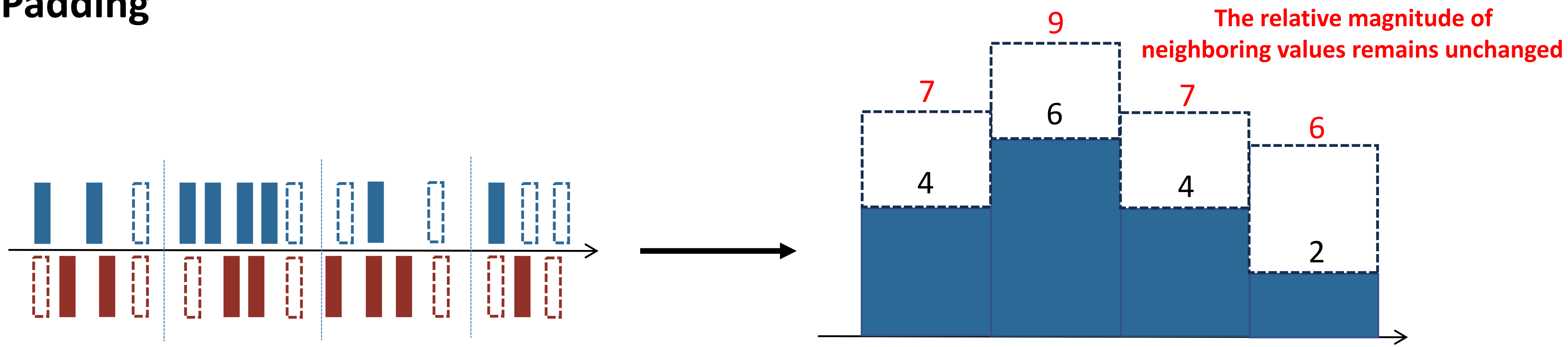


Deeper Look at Packet-per-second

Original Traffic



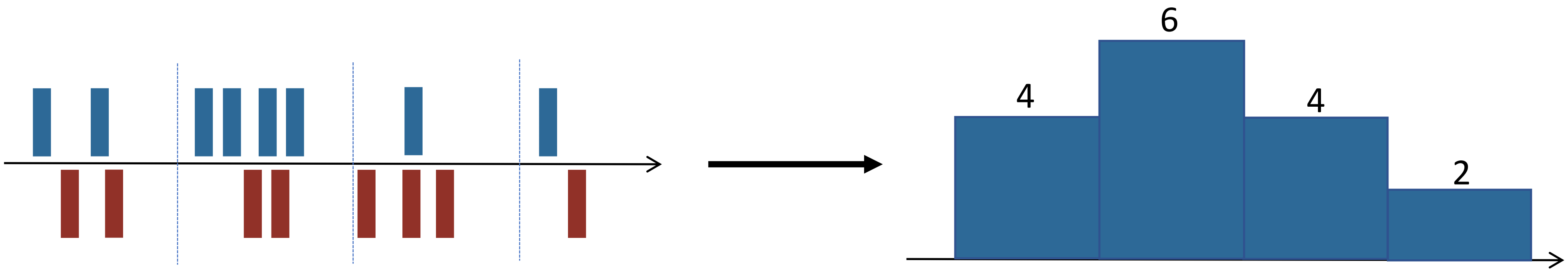
Packet Padding



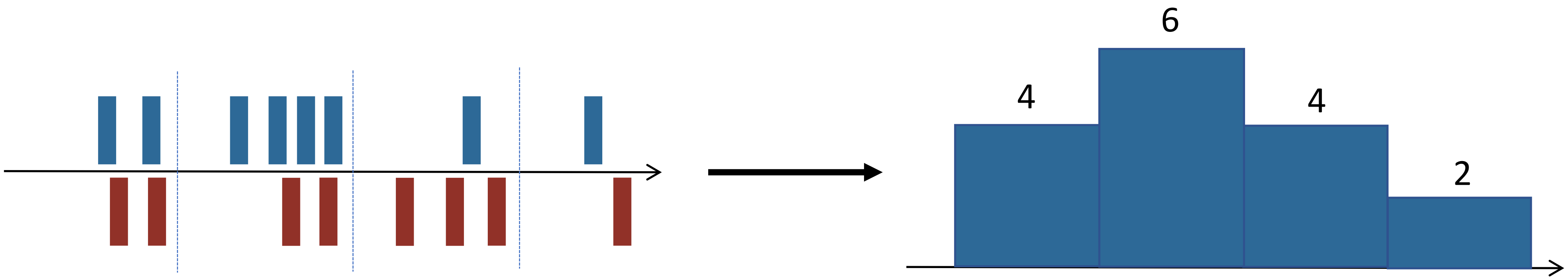
Accommodate the changes in the total number of packets by multiple intervals

Deeper Look at Packet-per-second (Cont'd)

Original Traffic



Packet Delaying



Resist moderate changes in time series

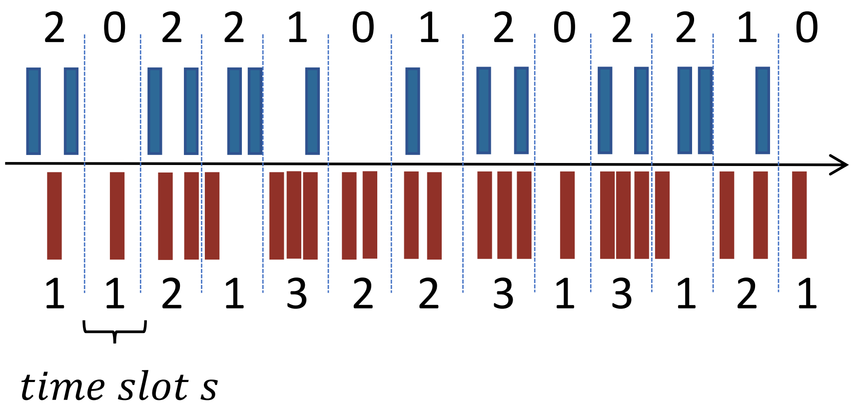
Traffic Aggregation Matrix

Definition

- TAM $M = \{m_{ij} \mid i \in \{1,2\}, j \in [1,\ N]\}$

Construction

- Divide the entire traffic into $\boldsymbol{N}$ small fixed-length time slots $\boldsymbol{s}$
  - Reduce the information loss
  - Tolerate packet padding and delaying
- Counts the number of outgoing and incoming packets per time slot
- Merges the values into the $2 \times N$ matrix.



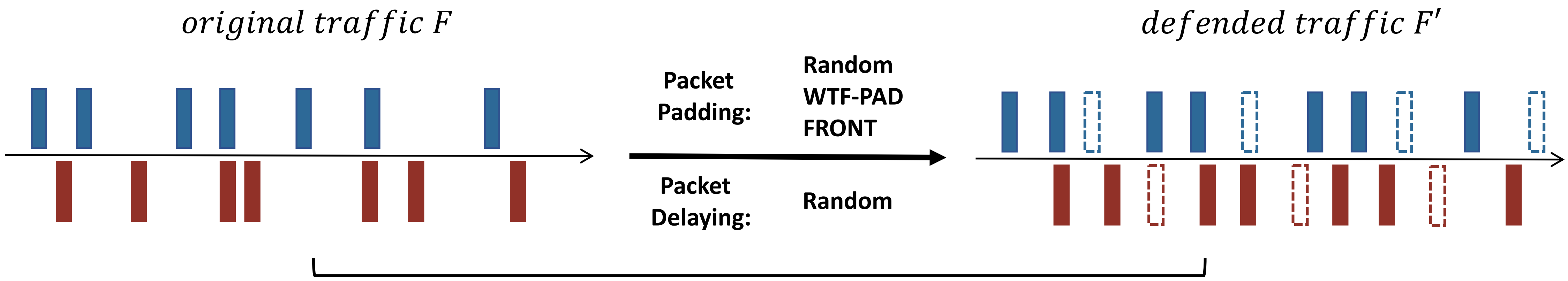
$M \in \mathbb{R}^{2\times N}$

2	0	2	2	⋯	0	**Incoming**
1	1	2	1	⋯	1	**Outgoing**

Traffic Aggregation Matrix (TAM)

Analysis of the Robustness Against Padding and Delaying

- Undefended Dataset[1] (randomly select 100 traces from 1000 traces for each of the 95 websites)
- Representations to compare: **Direction, Time with Direction**
- Intra-class distance metric: **Maximum Mean Discrepancy (MMD)** [2]

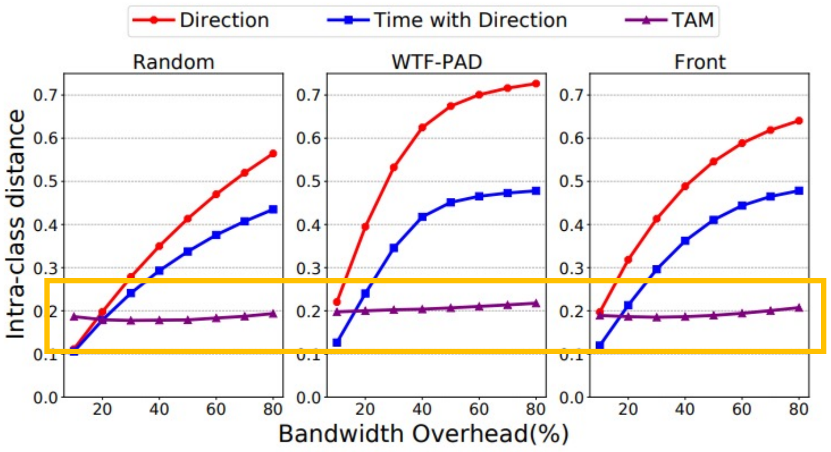


A robust representation should keep **the intra-class distance** between F and F' **as short as possible**

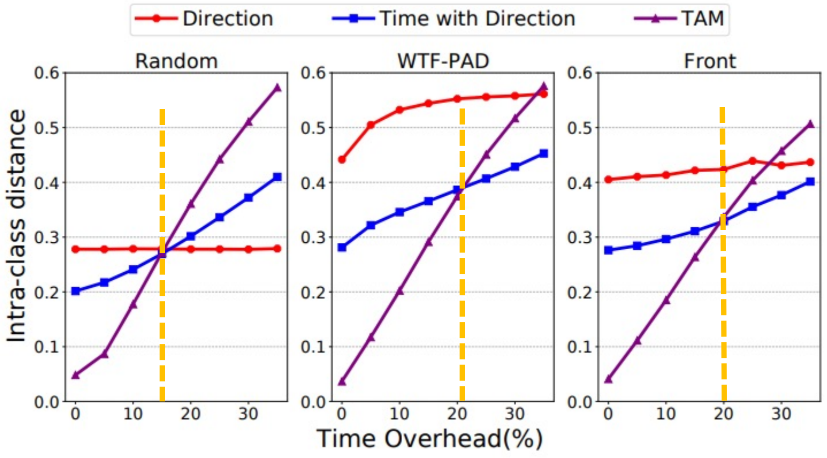
[1] Payap Sirinam, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. CCS 2018.
[2] Arthur Gretton, et al. A Kernel Two-Sample Test. JMLR 2012.

Analysis of the Robustness Against Padding and Delaying (Cont'd)

- Undefended Dataset[1] (randomly select 100 traces from 1000 traces for each of the 95 websites)
- Representations to compare: **Direction, Time with Direction**
- Intra-class distance metric: **Maximum Mean Discrepancy (MMD)** [2]



Vary bandwidth (fix time overhead to 10%)



Vary time (fix bandwidth overhead to 30%)

TAM is a more robust traffic representation under large bandwidth and moderate time overhead

[1] Payap Sirinam, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. CCS 2018.
[2] Arthur Gretton, et al. A Kernel Two-Sample Test. JMLR 2012.

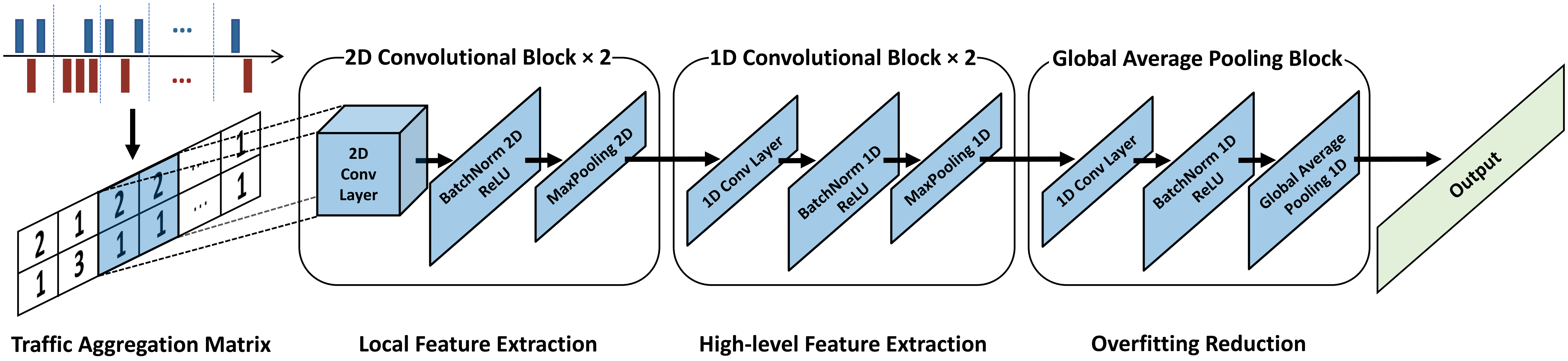
Design of Robust Fingerprinting

Robust Traffic Representation

- Aggregates multi-dimensional information: packet direction, number, and time.
- Tolerate packet padding and delaying

Effective CNN-based Classifier

- Extract robust discriminative features automatically



Experiment Setup

Public datasets:

	Monitored		Unmonitored
• **Undefended Dataset** [Sirinam, CCS'18]:	**95 websites × 1000**	**+**	**40,000 websites × 1**
• **Walkie-Talkie Dataset** [Rahman, PETS'20]:	**100 websites × 400**	**+**	**10,000 websites × 4**

WF attacks:

- **ML-based: k-NN** [Wang, USENIX'14], **CUMUL** [Panchenko, NDSS'16], and **k-FP** [Hayes, USENIX'16]
- **DL-based: AWF** [Rimmer, NDSS'18], **DF** [Sirinam, CCS'18], **Tik-Tok** [Bhat, PETS'19] and **Var-CNN** [Rahman, PETS'20]

WF defenses:

- **Traffic Disturbing: WTF-PAD** [Juarez, ESORICS'16], **Front** [Gong, USENIX'20], **RegulaTor** [Holland, PETS'22]
  Tamaraw [Wang, USENIX'14], **Blanket** [Nasr, USENIX'21], and **Walkie-Talkie** [Wang, USENIX'17]
- **Traffic Splitting: Traffic-Sliver** [la Cadena, CCS'20]
  - **By Direction (BD)**
  - **Batch Weighted Random (BWR)**

Attacks Comparison in the Closed-world Scenario

Attacks	Undefended	Disturbing Traffic Defenses					Splitting Traffic Defenses	
		WTF-PAD	Front	RegulaTor	Blanket	Walkie-Talkie	BD	BWR
k-FP	94.45	68.33	52.66	49.27	-	39.81	77.39	36.35
DF	98.40	90.85	76.85	20.96	98.00	71.02	20.69	19.99
Tik-Tok	98.45	93.80	84.79	47.07	98.13	72.85	92.74	57.63
Var-CNN	98.87	94.70	79.24	47.68	98.49	87.53	95.50	31.09
RF	**98.83**	**96.58**↓2.25	**93.34**↓5.49	**67.43**↓31.4	**98.62**↓0.21	**93.87**↓4.96	**95.70**↓3.13	**79.68**↓19.15

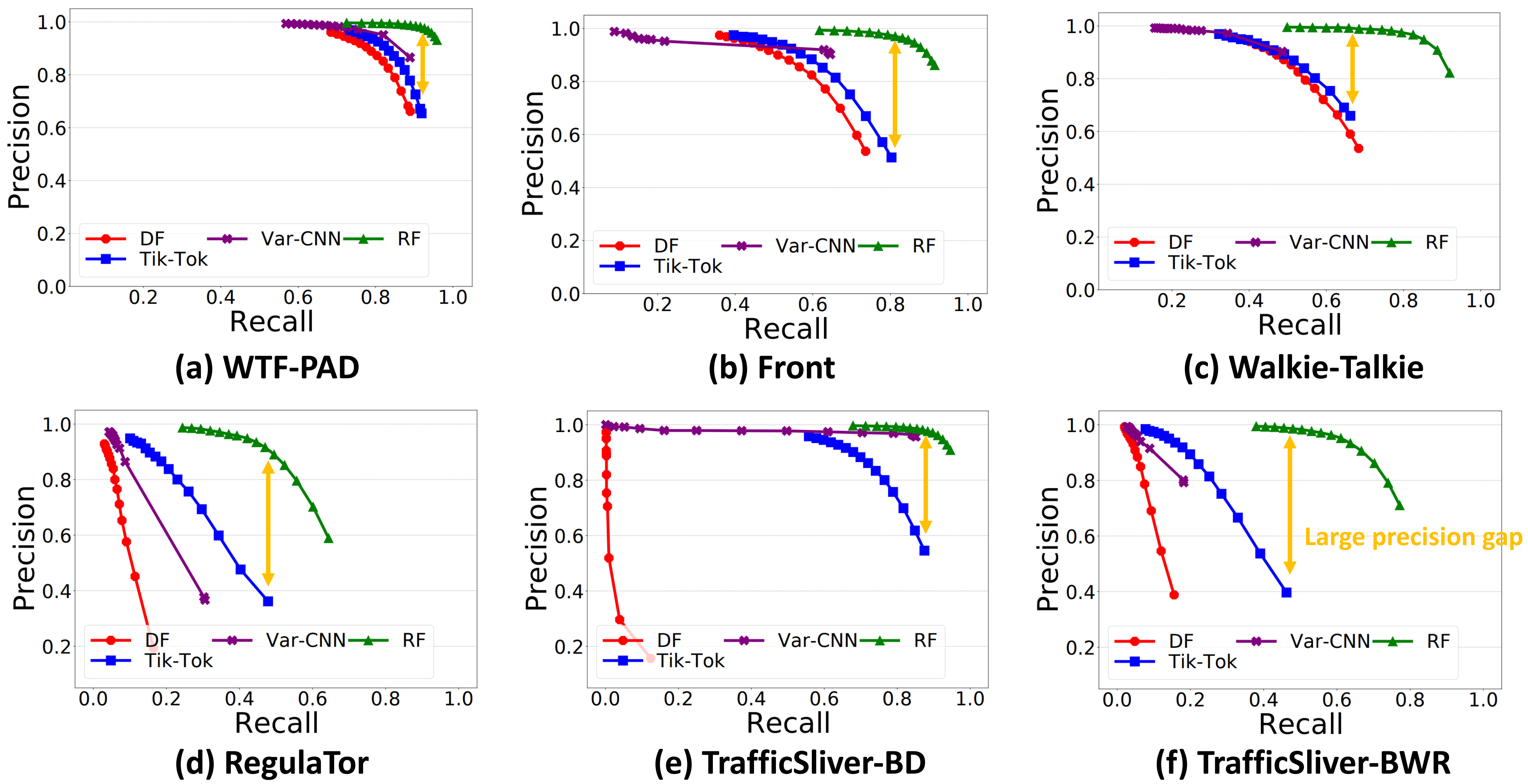
- RF has the slightest decrease in accuracy on all defenses, especially **for WTF-PAD, Front, Blanket, Walkie-Talkie and TrafficSliver-BD**, which decrease by less than **6%**

Attacks Comparison in the Closed-world Scenario (Cont'd)

Attacks	Undefended	Disturbing Traffic Defenses					Splitting Traffic Defenses	
		WTF-PAD	Front	RegulaTor	Blanket	Walkie-Talkie	BD	BWR
k-FP	94.45	68.33	52.66	49.27	-	39.81	77.39	36.35
DF	98.40	90.85	76.85	20.96	98.00	71.02	20.69	19.99
Tik-Tok	**98.45**	**93.80**	**84.79**	**47.07**	**98.13**	**72.85**	**92.74**	**57.63**
Var-CNN	98.87	94.70	79.24	47.68	98.49	87.53	95.50	31.09
RF	**98.83**↑0.38	**96.58**↑2.78	**93.34**↑8.55	**67.43**↑20.36	**98.62**↑0.49	**93.87**↑21.02	**95.70**↑2.96	**79.68**↑22.05

- RF outperforms all other WF attacks. Particularly, RF achieves a **best accuracy improvement of 22.05%** and an **average accuracy improvement of 8.9%** over **the SOTA attack Tik-Tok**

Attacks Comparison in the Open-world Scenario



- RF **consistently** and **significantly** outperforms other SOTA attacks on all defenses

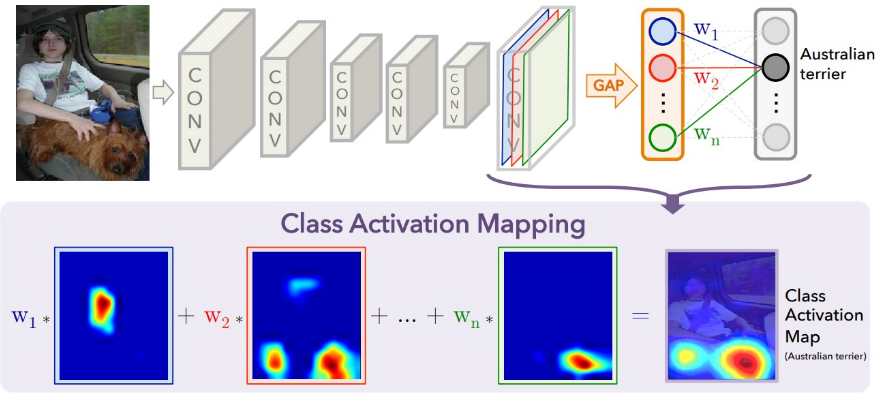
Countermeasure

Design Goals

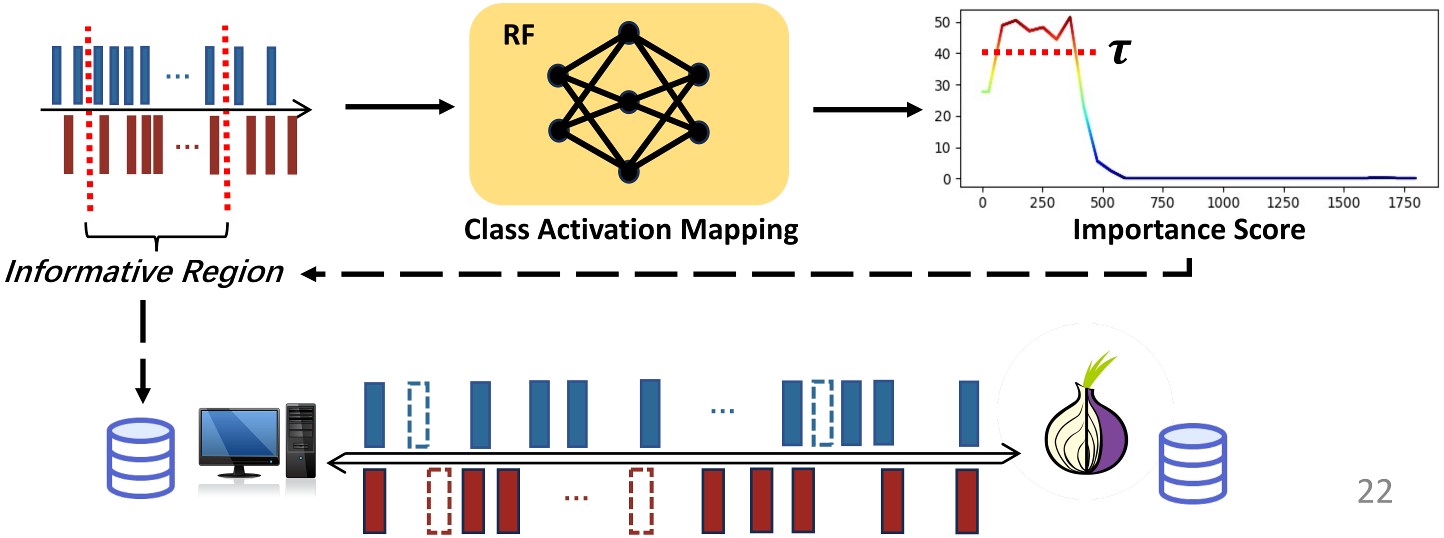
- **Effective:** Effectively reduce the accuracy of WF attacks.
- **Lightweight:** Introduces moderate bandwidth and time overhead.
- **Practical:** Can be applied to live traffic.

Basic Idea

- **Informative Region Extraction:** Use *Class Activation Mapping (CAM)* to learn packet sequences containing informative features from historical traffic of a collection of websites
- **Traffic Morphing:** Morph the original traffic from a certain website by packet padding and delaying to mimic multiple packet sequences from another website.



Class Activation Mapping [Zhou, CVPR'16]



Performance Evaluation

Defense	Overhead (%)		Accuracy (%)	
	Bandwidth	Time	RF	Var-CNN
BD	0	0	95.70	95.50
BWR	0	0	79.68	31.09
WTF-PAD	63	0	96.58	94.70
Front	103	0	93.34	79.24
Walkie-Talkie	31	34	93.87	87.53
RBB	43	14	97.63	86.35
Blanket	47	23	98.62	98.49
RegulaTor	77	5	67.43	47.68
Our Defense	**73**	**14**	**52.59**	**27.65**

- Our defense has the **best performance** and **moderate overhead** in defeating RF
- A **zero-delay** defense with **better performance** against RF is more desirable

Conclusion

Contributions

- Propose a robust traffic representation called Traffic Aggregation matrix (TAM)
- Present a novel WF attack Robust Fingerprinting (RF)
- Demonstrate RF is superior to SOTA WF attacks in closed- and open-world scenarios
- Develop a countermeasure against RF which more effective to reduce its accuracy

Future Work

- Explore more robust traffic representations
- Evaluate WF attacks against more real-world deployed defenses
- Investigate more effective zero-delay defenses against RF

Thank You!

Kexin Ji

jikexin@bit.edu.cn

Beijing Institute of Technology, China

Source Code and Datasets Available:

https://github.com/robust-fingerprinting/RF