Enhancing Website Fingerprinting Attacks against Traffic Drift

Xinhao Deng, **Yixiang Zhang**, Qi Li, Zhuotao Liu, Ke Xu

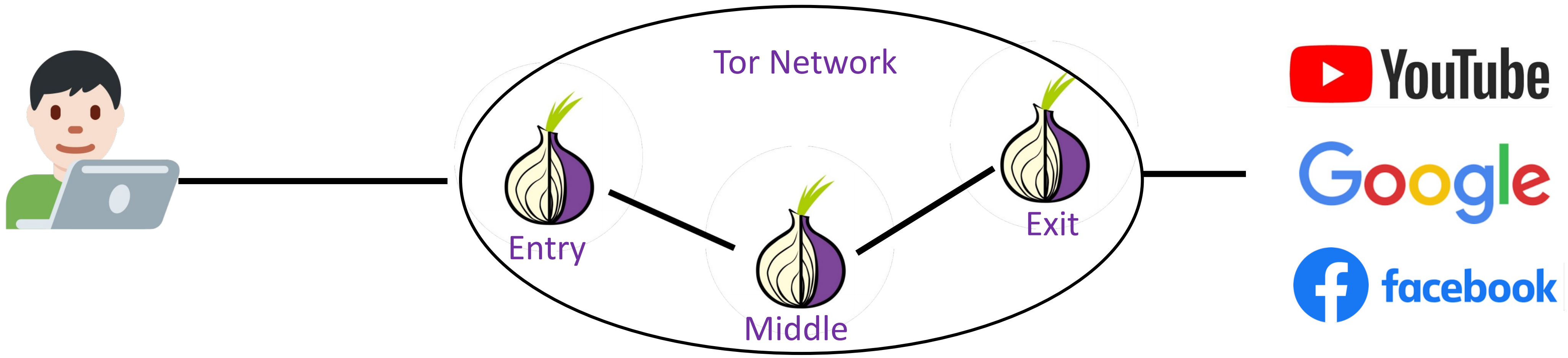
Tsinghua University, Beijing, China

Ant Group, Hangzhou, China

Zhongguancun Laboratory , Beijing, China

Introduction: Tor

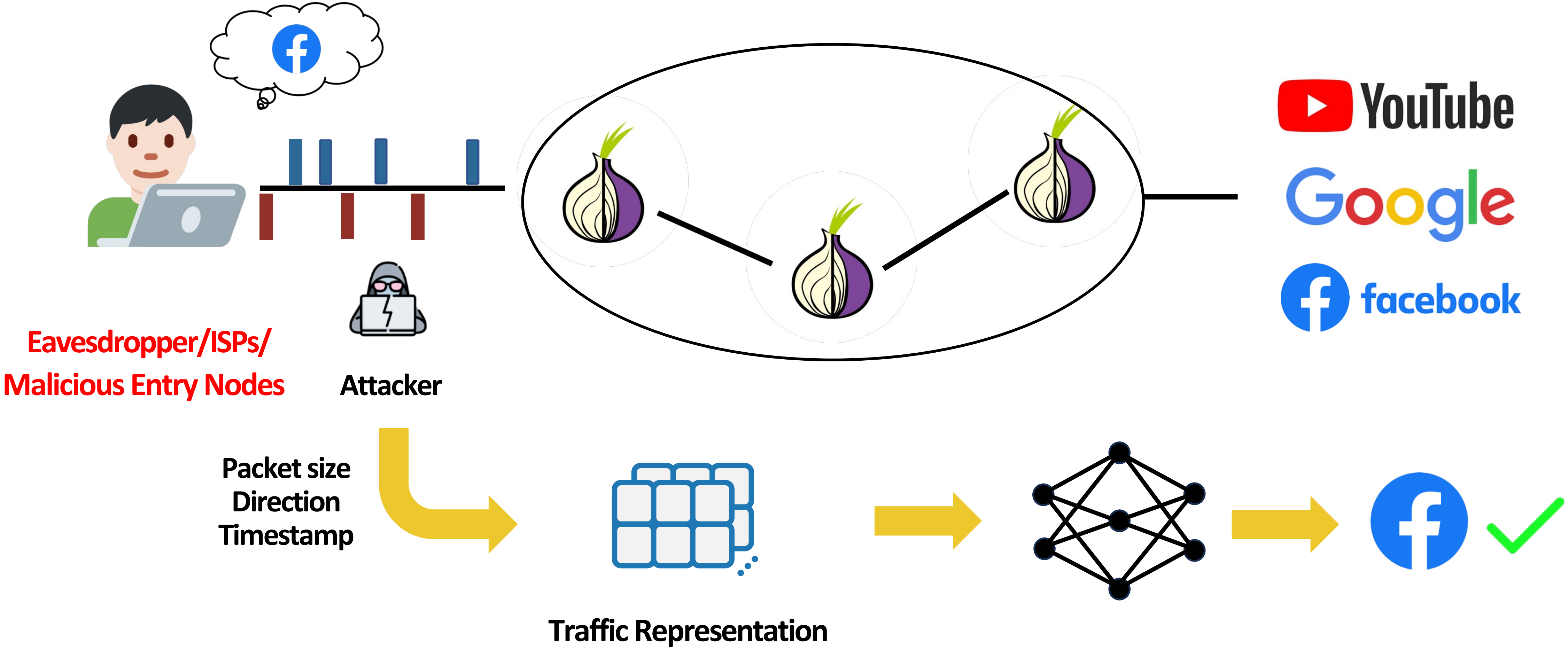
Tor prevents users from being **tracked, monitored** and **censored**.
It routes traffic across a randomly selected 3-hop circuit with layered encryption.



Website Fingerprinting Attack

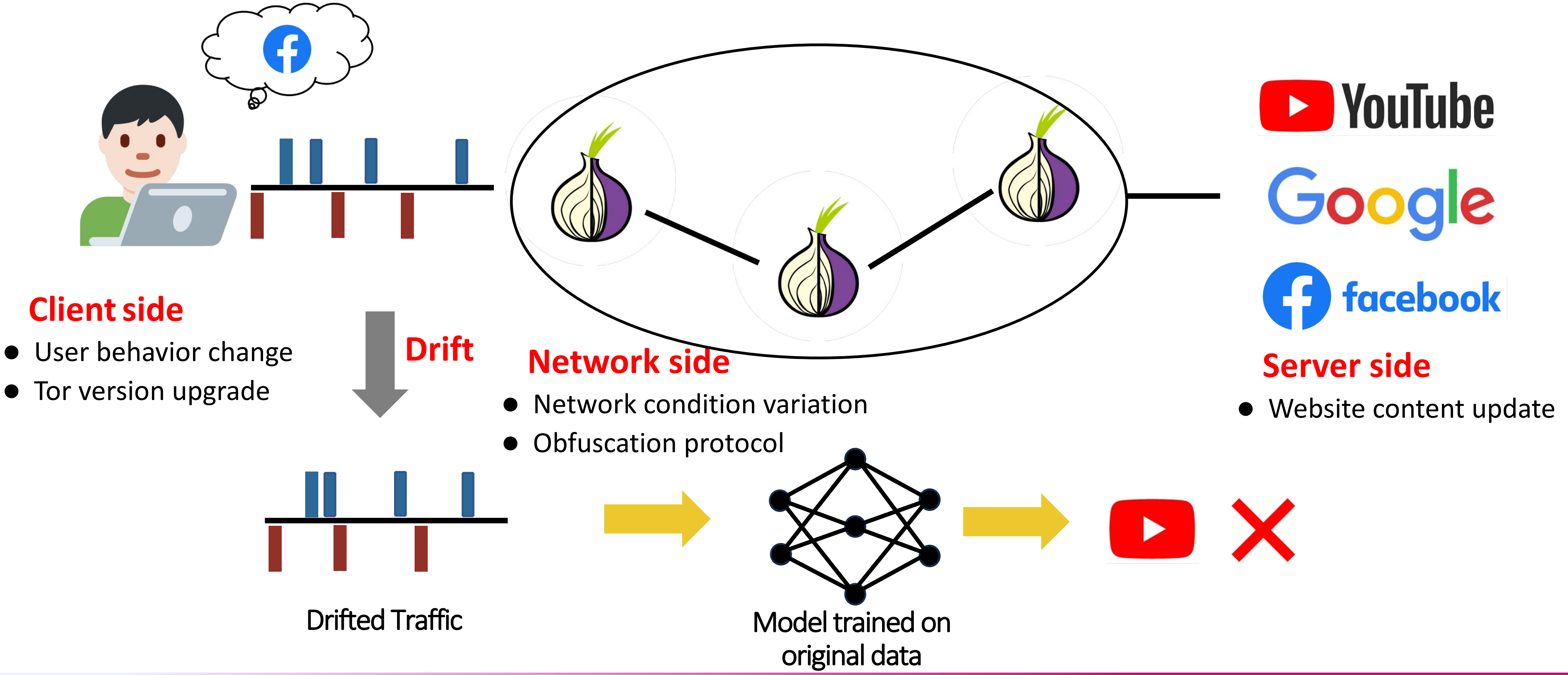
Website Fingerprinting (WF) Attack:

WF attackers try to **infer the website** that a user is visiting **without decrypting the traffic**.



Traffic Drift during WF Attack

Changes on the client side, server side, and network side lead to **traffic drift**, result in **drop of WF performance**.



Existing Mitigation of Traffic Drift

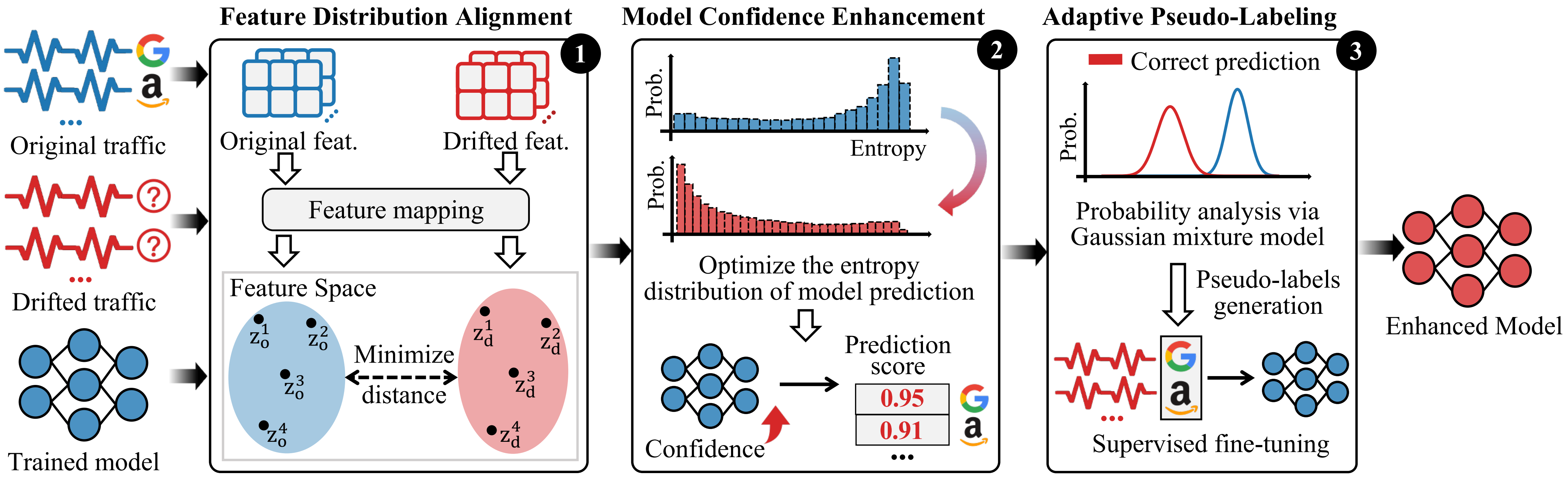
Existing WF attacks are insufficient to effectively handle **diverse** and **complex** traffic drift.

Methods	Related Works	Limitation
Periodically Retraining	AWF [NDSS,18] DF [CCS,18] ARES [S&P,23]	**Time-consuming**
Few-shot Fine-tuning	Var-CNN [PETS,19] TF [CCS,19] GANDALF [PETS,21] NetCLR [CCS,23]	**Require labeled traffic that precisely reflects drift**
Online Adaptation	OnlineWF [Security,22] Retracer [WPES,23]	**Inconsistency with entry-node traffic**

All depend on Labeled Drifted Traffic

Our Design Goal

Question: How to adapt WF attacks to **diverse** and **unknown** traffic drift during deployment

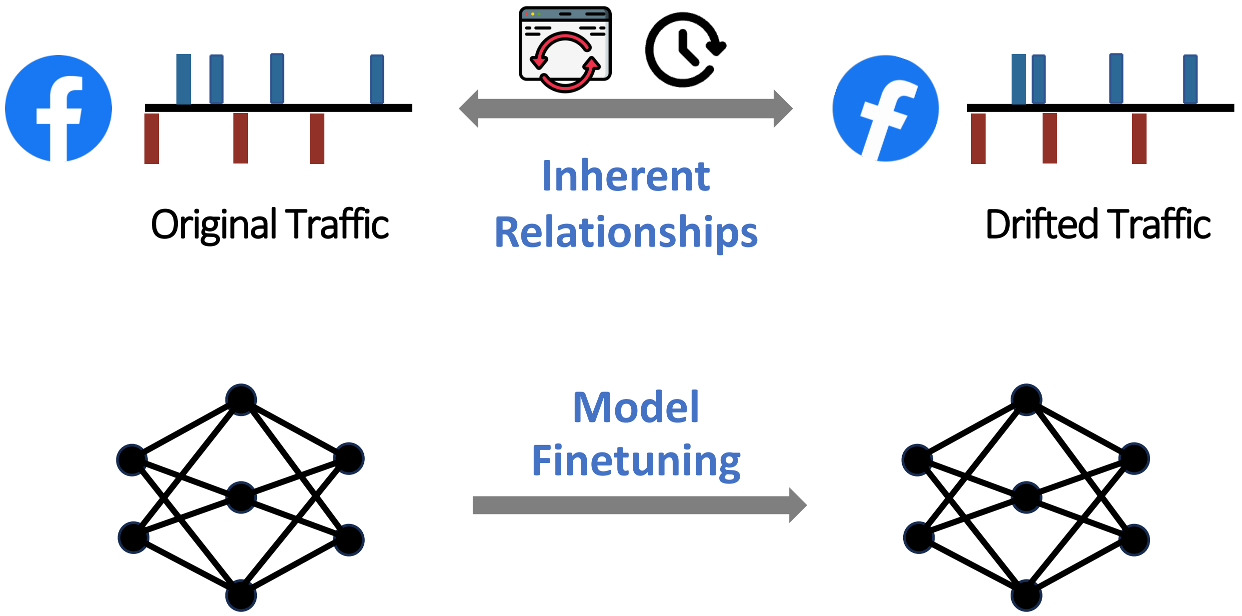


We propose **Proteus**, an adaptive approach that

fine-tunes WF models using traffic **without ground-truth labels**

Our Intuition

Intuition: Traffic of the same website exhibits **inherent relationships** before and after drift



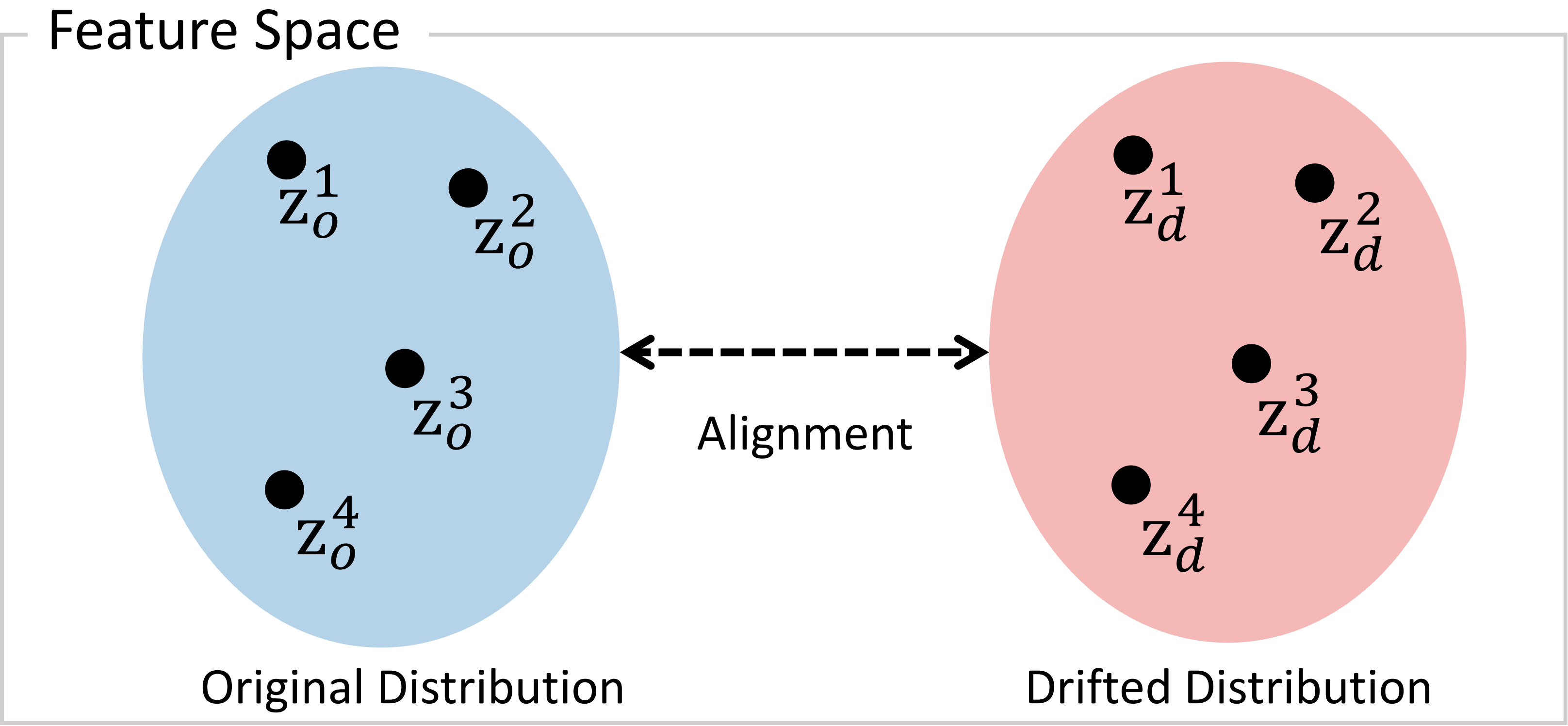
Stage1: Aligning overall feature distribution (unsupervised)
Stage2: Improving confidence of prediction results (unsupervised)
Stage3: Selecting reliable pseudo labels (for supervised fine-tuning)

Stage1: Feature Distribution Alignment

Goal: Aligning overall feature distribution before and after drift

Challenges:

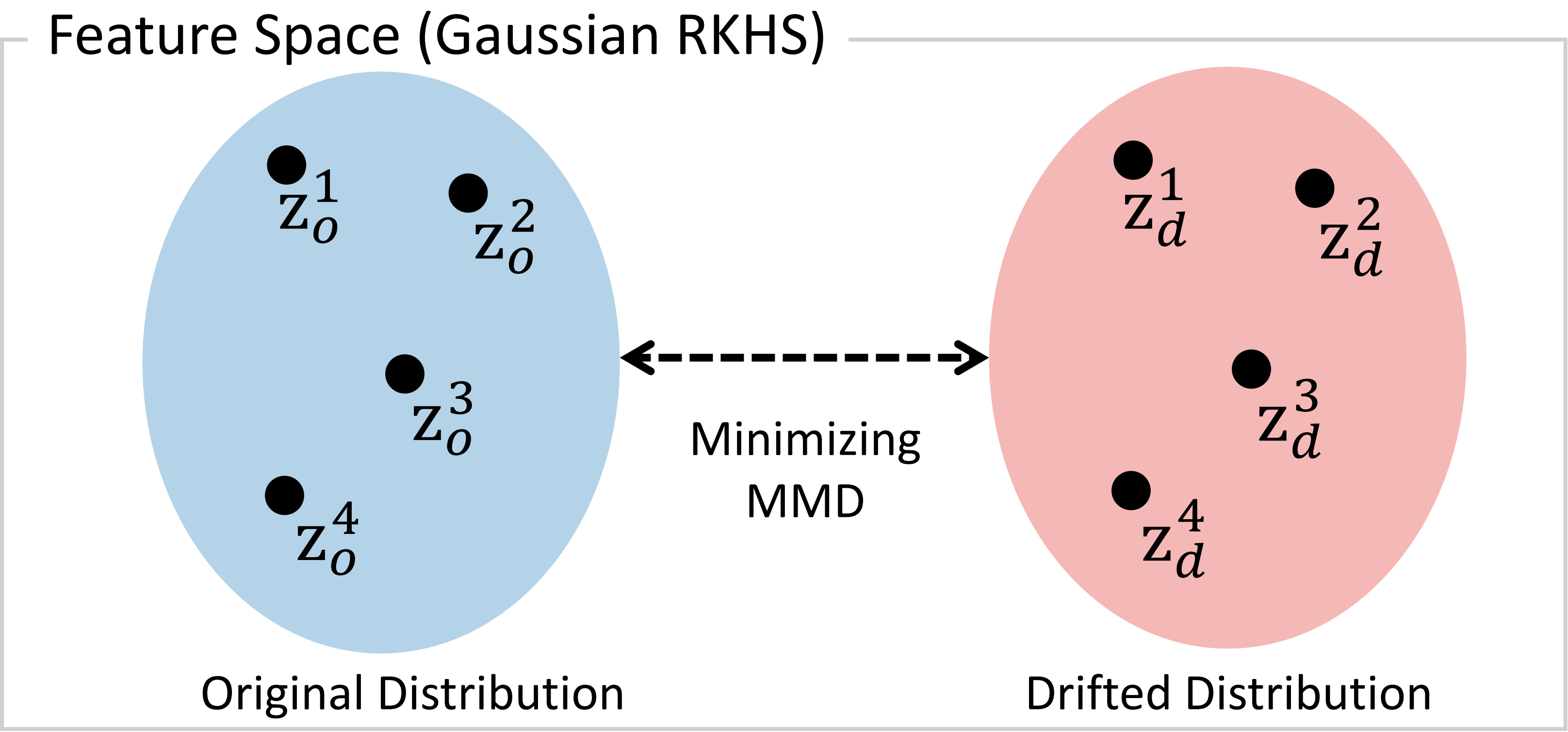
- Relationships between two feature distributions are **complex**
- Causes of traffic drift in the real world are **diverse**



Stage1: Feature Distribution Alignment

Solution: Minimizing squared **MMD** in **Gaussian RKHS** (Reproducing Kernel Hilbert Space)

- **Gaussian kernel** makes complex features more **linearly separable** in RKHS
- **Adaptive bandwidth** enhances **adaptability** to diverse traffic drift scenarios



Stage2: Model Confidence Enhancement

Goal: Improving confidence of prediction results

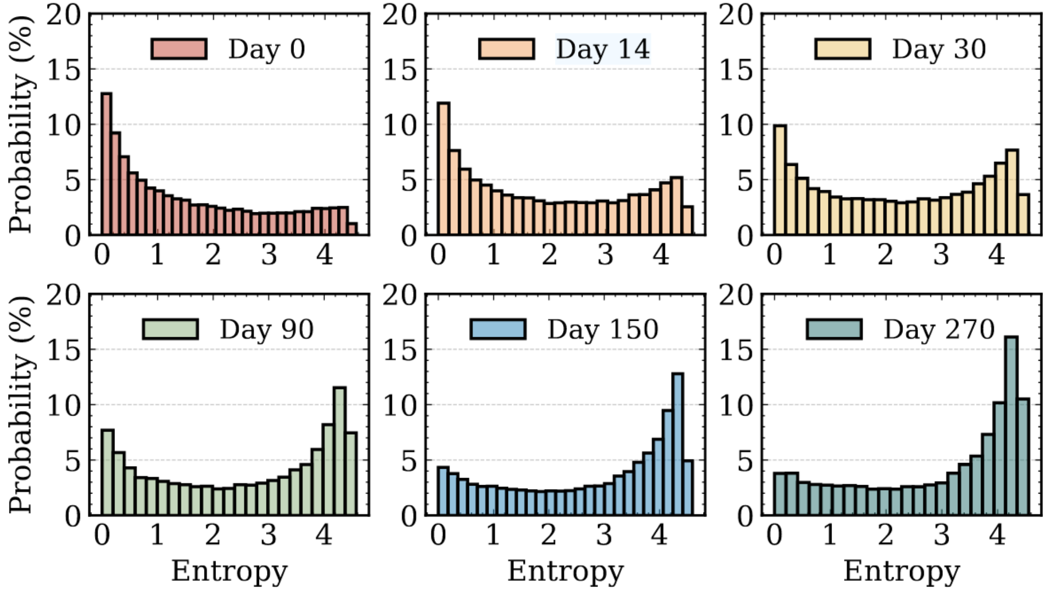
Challenges:

- Drift traffic is **unlabeled**
- Simply maximizing softmax probabilities **biases** toward the dominant class

Solution: Optimizing **Entropy Distribution**

- **Batch-based** optimization constraints
- Minimizing entropy for each sample
- Maximizing entropy for all prediction results within a batch

Shannon entropy measures uncertainty



Model uncertainty raises while drift increases, leading to misclassification.

Yixiang Zhang, Tsinghua University

Stage3: Adaptive Pseudo-Labeling

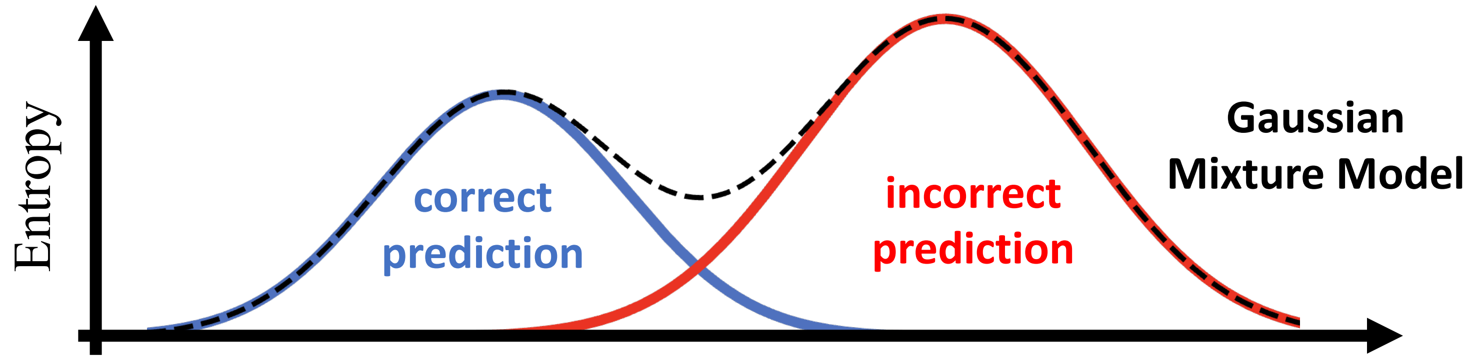
Goal: Generating reliable pseudo labels

Challenges:

- Tor traffic drift is **dynamic** and **diverse**
- Simply selecting high-confidence prediction suffers from **severe noise**

Solution: **Probability-based** pseudo-labeling

- Using **Gaussian Mixture Model** to fit the entropy distribution
- Selecting predictions with **higher posterior probability** to be a correct prediction



Entropy of in- and correctly predictictions ***fits Gaussian distribution, separately***

Experiment Setup

Baselines

- WF baseline
  - AWF [NDSS,18], DF [CCS,18], TikTok [PETS,19], Var-CNN [PETS,19], BAPM [ACSAC,21], ARES [S&P,23], RF [Security,23] and NetCLR [CCS,23]
- Drift adaptation baseline
  - Holmes [CCS,24] and UAF [SecureComm,23]

Implementation of Proteus

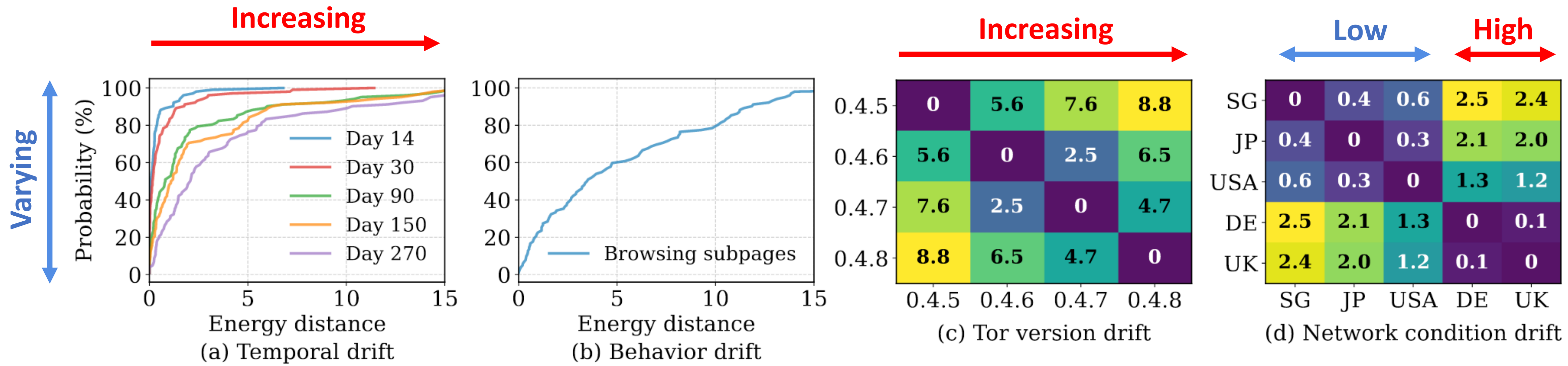
- Default WF model: RF
- Proteus can also integrate with other DL-based WF attacks

Experiment Setup

Dataset: 6 scenarios, 102 constant monitored websites, 350,000 real-world traces

Scenario	Trace Number	Description
Temporal Drift	140,000	Collected at 6 checkpoints over 9 months
Tor Version Drift	100,000	Across Tor 0.4.8, 0.4.7, 0.4.6, 0.4.5
Network Condition Drift	34,000	From 5 countries with varying network conditions
Browsing Behavior Drift	20,000	17,739 subpages of 102 monitored websites
Open-World Dataset	160,000	20,000 unmonitored + 102 monitored websites
Dataset with WF Defense	200,000	Obfuscated by WTF-PAD, obfs4, and Front

Drift Quantification Analysis



- Traffic drift **varies significantly** across websites
- The degree of drift **increases with drift factors** in specific scenarios
  - Temporal drift increases progressively over time.
  - The larger differences between Tor versions, the more severe traffic drift.
- Different scenarios with various factors result in **various magnitude** of traffic drift
  - User behaviors involving subpage visits lead to more pronounced drift.
  - The magnitude of network condition drift is lower compared to the other scenarios.

Evaluation under Diverse Traffic Drift

Proteus achieves the **best performance** under diverse traffic drift.

- Proteus: 0.8227 F1-score on day-270 traffic
- Other attacks: < 0.6 F1-score

	Day 14			Day 30			Day 90			Day 150			Day 270		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AWF	50.36	49.29	48.84	46.42	46.05	45.36	38.28	39.09	37.49	33.12	33.63	32.21	29.40	30.07	28.64
BAPM	64.62	62.83	62.08	58.89	57.33	56.28	48.48	48.78	46.75	44.28	43.92	41.89	38.38	38.30	35.98
ARES	71.36	69.07	68.82	66.74	64.72	64.17	56.53	55.96	54.05	49.06	49.21	47.43	46.04	45.51	43.48
DF	73.30	72.10	71.94	67.12	66.49	65.79	55.91	57.17	55.24	50.11	50.49	48.91	45.47	46.69	44.48
NetCLR	73.78	72.93	72.71	68.27	67.47	66.92	56.99	57.94	55.81	49.29	50.82	49.03	44.72	46.80	43.92
Tik-Tok	78.98	78.35	77.89	73.46	73.25	72.36	62.01	63.08	60.87	53.89	54.90	52.56	48.70	49.47	46.58
Var-CNN	81.23	79.93	79.77	76.20	74.76	74.30	65.03	65.43	63.14	57.43	58.14	55.68	52.79	53.49	50.84
RF	88.46	87.99	87.63	82.92	82.15	81.62	73.83	73.85	72.02	68.07	68.13	66.25	61.00	62.81	59.57
Proteus	**92.53**	**92.59**	**92.53**	**91.21**	**91.23**	**91.18**	**90.67**	**90.77**	**90.65**	**86.15**	**86.82**	**86.32**	**81.90**	**83.21**	**82.27**

Evaluation under Temporal Drift

Evaluation under Diverse Traffic Drift

Proteus achieves the **best performance** under diverse traffic drift.

- Proteus: 0.5524 F1-score on different unknown subpages
- Other attacks: < 0.45 F1-score

	Homepage			Subpages		
	P	R	F1	P	R	F1
AWF	54.74	53.74	53.04	22.18	22.87	21.68
BAPM	70.29	65.68	65.38	29.98	27.67	25.96
ARES	75.36	73.28	72.77	31.87	31.96	29.87
DF	77.73	77.11	76.72	32.12	33.56	31.36
NetCLR	78.66	77.58	77.27	34.10	35.50	33.58
Tik-Tok	83.48	82.00	81.82	36.07	36.66	34.56
Var-CNN	85.08	83.60	83.59	39.58	39.19	37.24
RF	89.85	89.21	89.00	48.79	46.74	44.76
Proteus	**91.28**	**91.40**	**91.18**	**55.32**	**56.32**	**55.24**

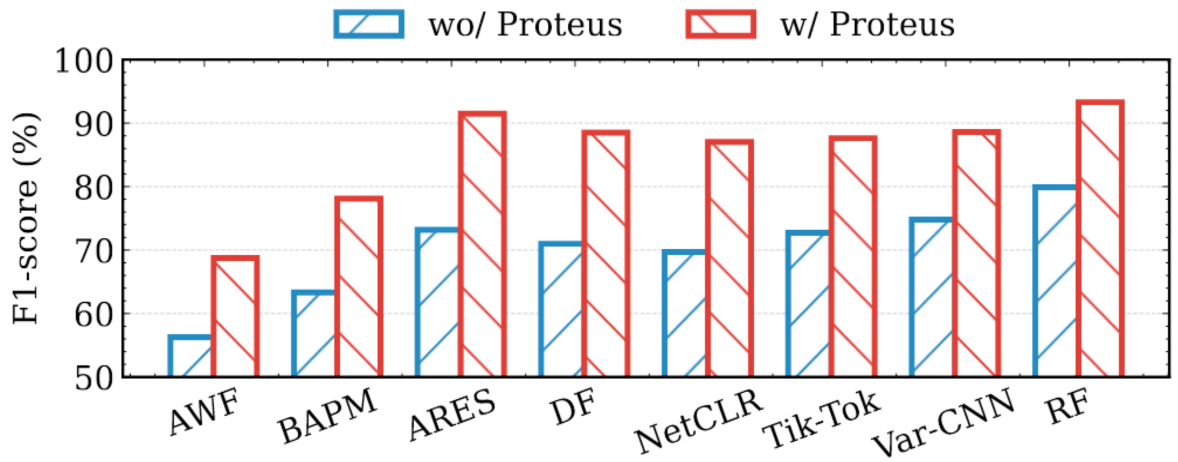
↑

Evaluation under Browsing Behavior Drift

Evaluation under Diverse Traffic Drift

Proteus integrates with WF attacks and improves their performance.

➢ F1-score improved **significantly** and **consistently** for WF attacks under Tor Version Drift and Network Condition Drift



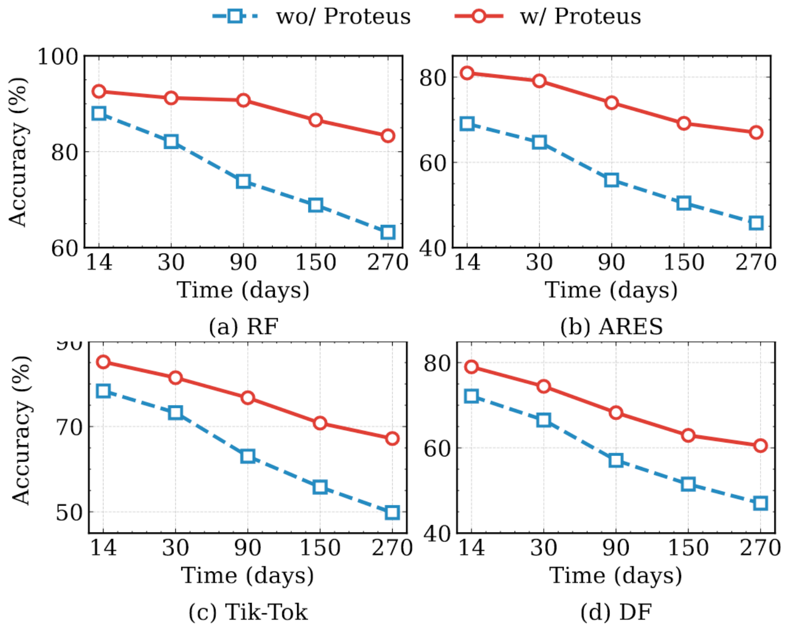
		SG	JP	USA	DE	UK
AWF	w/o Proteus	52.15	43.78	43.21	30.63	32.59
	w/ Proteus	**54.25**	**46.87**	**44.96**	**31.59**	**35.87**
BAPM	w/o Proteus	66.01	57.10	53.23	38.26	39.61
	w/ Proteus	**68.64**	**58.48**	**56.29**	**44.98**	**48.82**
ARES	w/o Proteus	72.96	66.48	62.87	43.75	46.19
	w/ Proteus	**80.27**	**74.25**	**70.95**	**58.83**	**65.35**
DF	w/o Proteus	76.24	68.58	63.94	46.67	50.14
	w/ Proteus	**78.10**	**73.03**	**69.22**	**56.18**	**61.87**
NetCLR	w/o Proteus	78.25	71.76	68.13	46.66	49.76
	w/ Proteus	**78.91**	**73.07**	**68.76**	**57.58**	**60.23**
Tik-Tok	w/o Proteus	81.36	74.74	74.61	51.34	50.25
	w/ Proteus	**84.49**	**81.95**	**79.40**	**67.40**	**70.08**
Var-CNN	w/o Proteus	78.59	74.42	73.00	45.80	47.11
	w/ Proteus	**84.38**	**81.57**	**81.69**	**66.10**	**69.89**
RF	w/o Proteus	88.87	87.45	83.10	20.85	17.38
	w/ Proteus	**90.63**	**90.98**	**92.64**	**81.66**	**84.60**

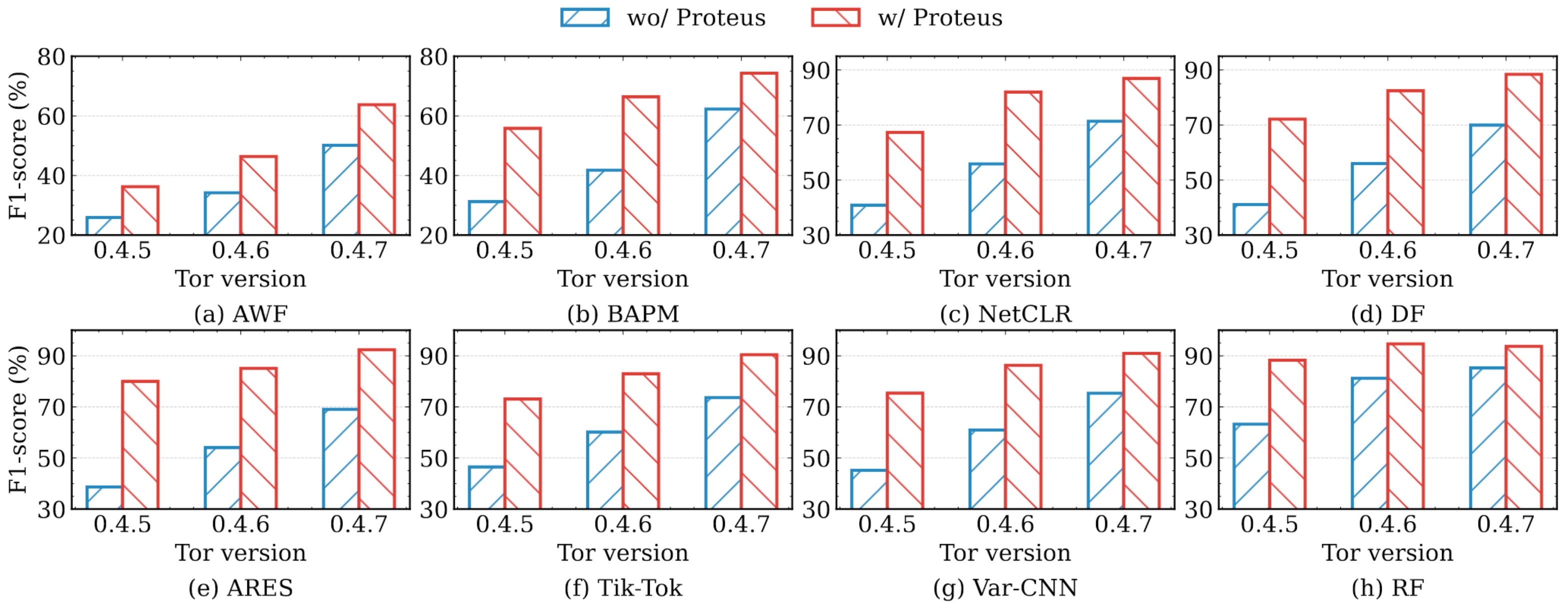
Evaluation under Tor Version Drift and Network Condition Drift

Evaluation under Diverse Traffic Drift

The advantage of **Proteus** increases as the degree of drift increases.

➤ Advantage increases as **drift time** & **version difference** increases

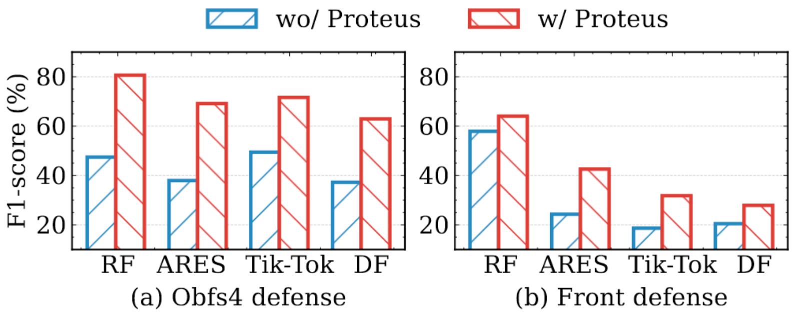




Evaluation under Temporal Drift and Tor Version Drift

Evaluation under Diverse Traffic Drift

- **Proteus** achieves the **best performance** under diverse traffic drift.
- **Proteus integrates with WF attacks** and improves their performance.
- The advantage of **Proteus increases as the degree of drift increases**.



Traffic Obfuscation

	Day 0→270	Tor 0.4.8→0.4.5	SG→DE	Homepage→Subpage
UAF	50.31	51.15	57.33	36.50
Holmes	54.31	55.62	29.43	38.82
Proteus	**82.27**	**88.28**	**81.66**	**55.24**

Comparison with Drift Adaptation Baseline

Conclusion

- We propose **Proteus**
  - a WF attack framework that adapts to real-world traffic drift by fine-tuning models with traffic without ground-truth labels
- Proteus accomplishes this by
  - aligning feature distributions and optimizing entropy distributions to improve the consistency of WF model predictions
  - generating reliable pseudo-labels that enable supervised finetuning to improve the model's adaptability
- We conduct extensive real world evaluations on large scale Tor traces to demonstrate its effectiveness

Enhancing Website Fingerprinting Attacks against Traffic Drift

Xinhao Deng, **Yixiang Zhang**, Qi Li, Zhuotao Liu, Ke Xu

Tsinghua University, Beijing, China

Ant Group, Hangzhou, China

Zhongguancun Laboratory , Beijing, China

Q&A Form