Rethinking Encrypted Traffic Classification: A Multi-Attribute Associated Fingerprint Approach

Yige Chen∗†, Tianning Zang∗†, Yongzheng Zhang∗†,
Yuan Zhou‡ , Yipeng Wang∗†

∗ Institute of Information Engineering, Chinese Academy of Sciences

†School of Cyber Security, University of Chinese Academy of Sciences

‡National Computer Network Emergency Response Technical Team/Coordination Center of China

The 27th IEEE International Conference on Network Protocols
Chicago, Illinois, USA, October 7-10, 2019

What is Encrypted Traffic Classification?

- Most mobile applications adopt SSL/TLS to ensure their communication security
- Classify mixed mobile encrypted traffic into applications
- Provide fundamental support to application QoS and application-level firewall
- The existing approaches lack a balance between accuracy and computational complexity

Key Insight -- Domain Name

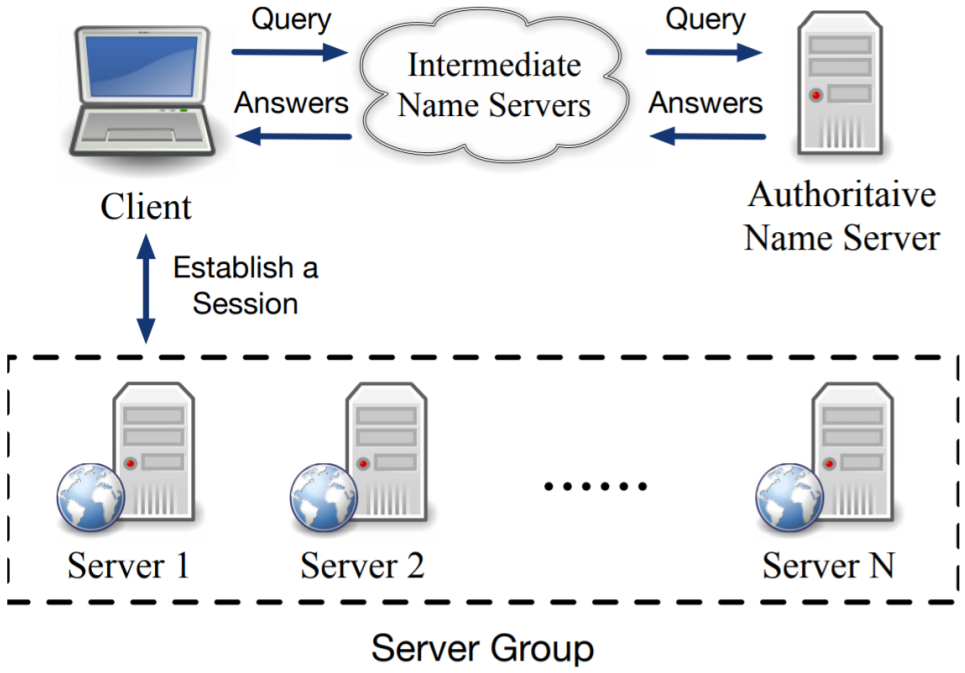
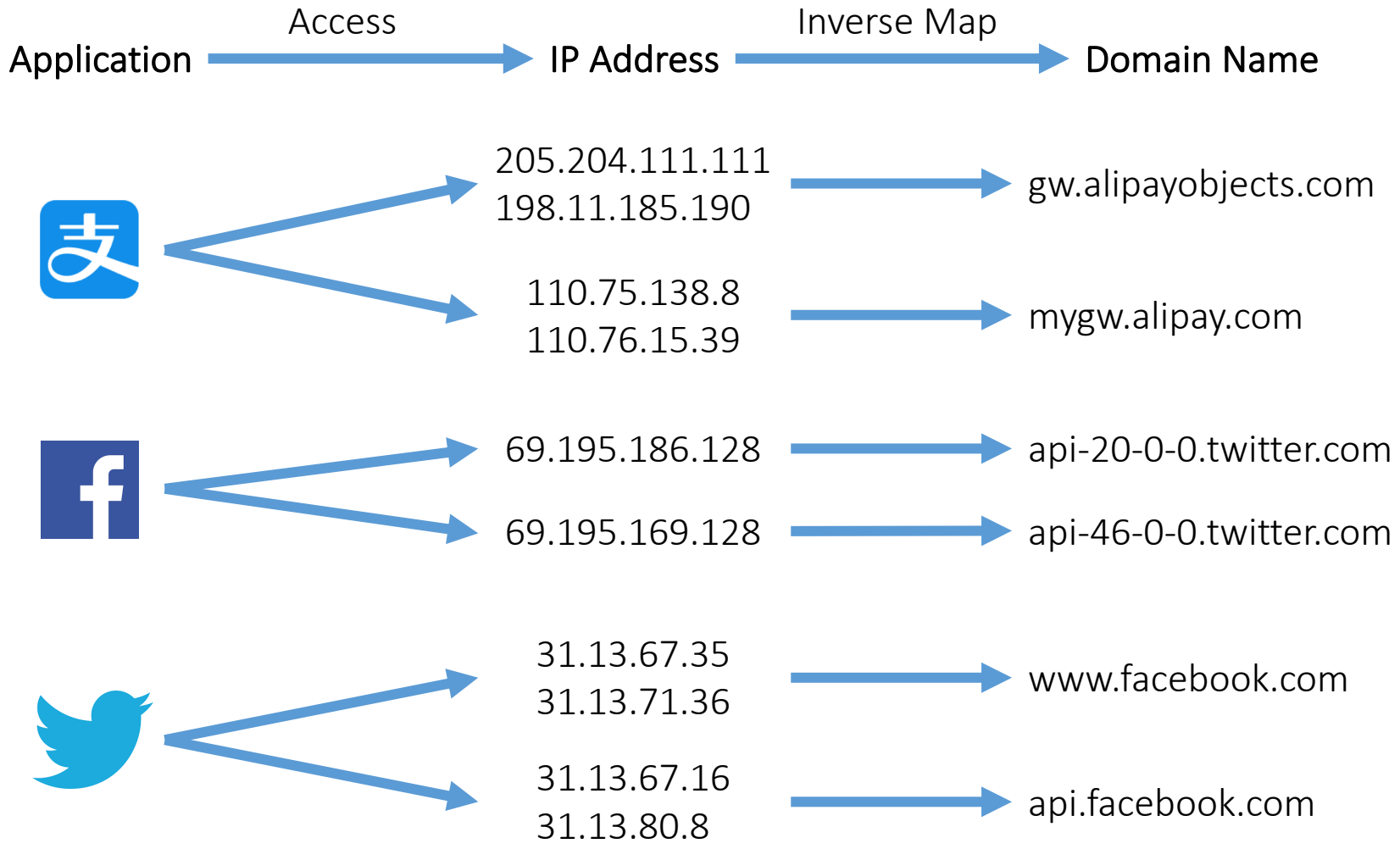


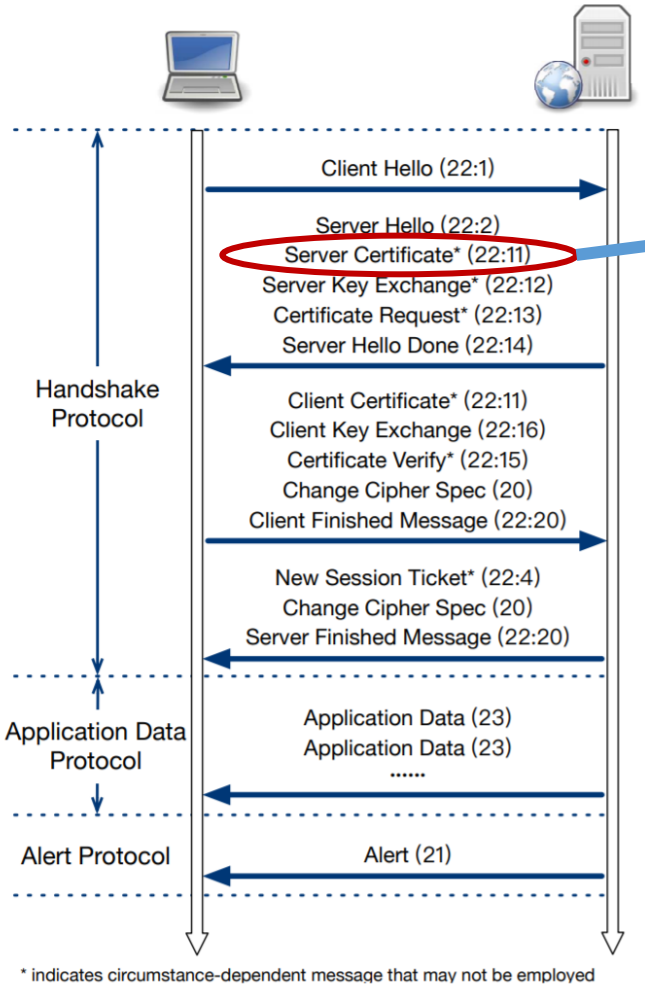
Fig. 2. The Topology of DNS-based Load Balancing

- The DNS-based load balancing topology is popularly employed by applications
- The topology balances the workload distribution of the application servers
- The Domain name is a unified entrance of the application servers.

Key Insight -- Domain Name (Example)



Key Insight -- Certificate



Extract Subjects

```
item: (id-at-countryName=CN)
item: (id-at-stateOrProvinceName=Shanghai)
item: (id-at-localityName=Shanghai)
item: (id-at-organizationName=Alipay.com Co.,Ltd)
item: (id-at-organizationalUnitName=Secure Web Server)
item: (id-at-commonName=*.alipay.com)
```

Organization Name (OID = 2.5.4.10)

Common Name (OID = 2.5.4.3)

- The server certificate in the SSL/TLS session contains essential subjects
- Common Name -- the glob domain name of the servers
- Organization Name -- the organization of the servers

Fig. 1. An Example of the SSL/TLS Protocol Communication Session

Key Insight – Certificate (Example)

Extract Subjects

item: (id-at-organizationName=Alipay.com Co.,Ltd)

item: (id-at-commonName=*.alipay.com)
item: (id-at-commonName=entphz.alipay.com)
item: (id-at-commonName=*.alipayobjects.com)

Extract Subjects

item: (id-at-organizationName=Facebook, Inc.)

item: (id-at-commonName=*.facebook.com)

Extract Subjects

item: (id-at-organizationName=Twitter, Inc.)

item: (id-at-commonName=twitter.com)
item: (id-at-commonName=mobile.twitter.com)
item: (id-at-commonName=api.twitter.com)

Multi-Attribute Associated Fingerprint

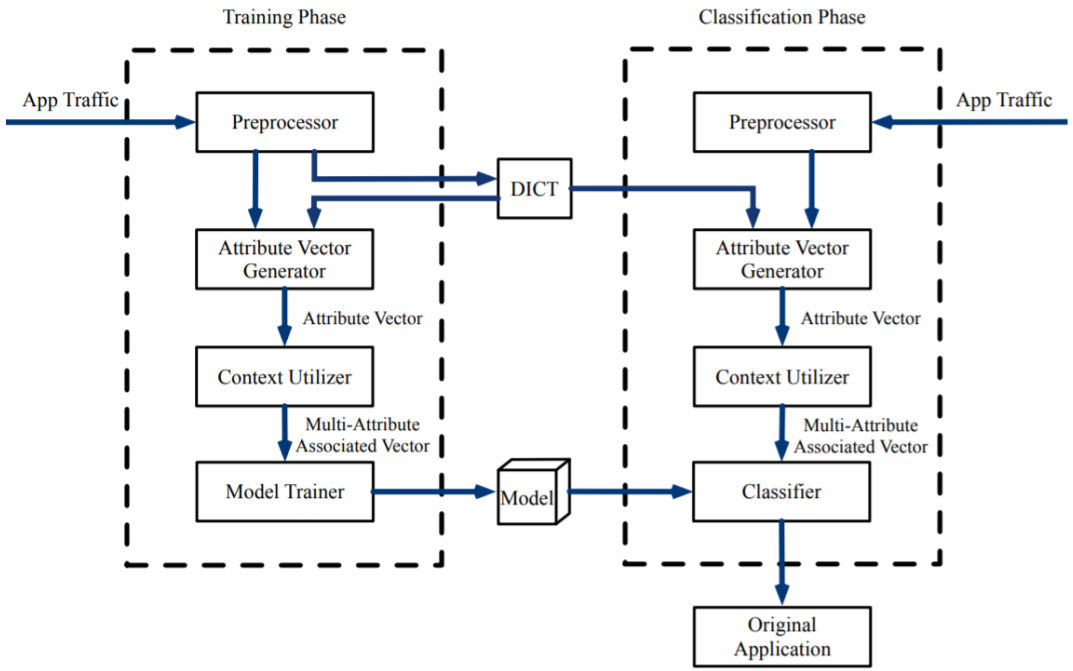
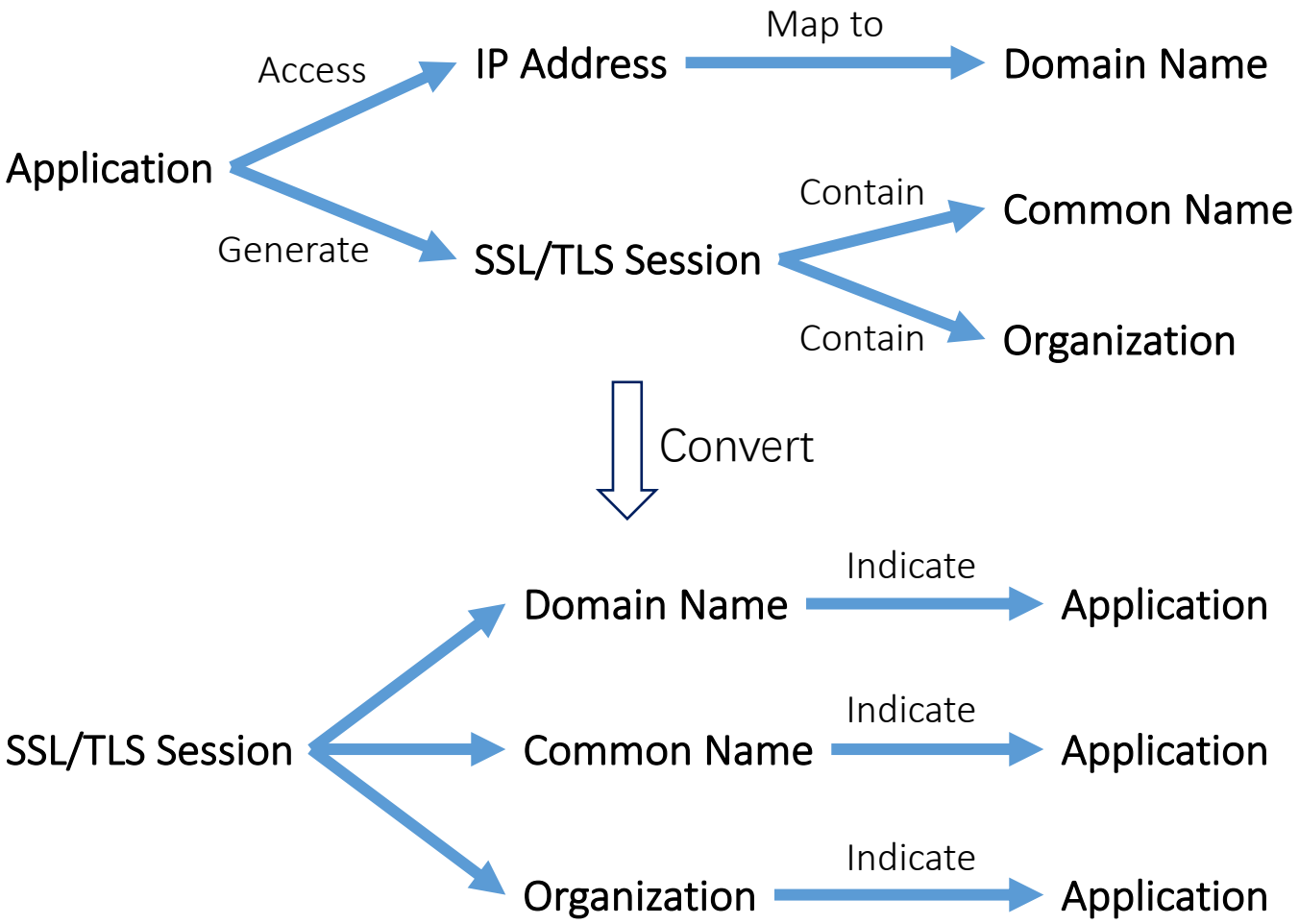


Fig. 3. The System Structure of MAAF

- Two Phases -- training and classification
- Preprocessor -- extract attributes
- Attribute Vector Generator -- convert flows into computable vectors
- Context Utilizer -- extend contextual information
- Classifier -- classify vectors into applications

MAAF -- Preprocessor



- The amount of the correlated traffic indicates the correlation between the attribute and the applications

MAAF – Preprocess (Example)

Flow Attribute Table

Flow	App	Domain Name	Common Name	Organization
1	Alipay	gw.alipayobjects.com	*.alipayobjects.com	Alipay.com Co.,Ltd
2	Alipay	mygw.alipay.com	*.alipay.com	Alipay.com Co.,Ltd
3	Facebook	www.facebook.com	*.facebook.com	Facebook, Inc.
4	Facebook	api.facebook.com	*.facebook.com	Facebook, Inc.
5	Twitter	api-20-0-0.twitter.com	*.twitter.com	Twitter, Inc.
6	Twitter	api-20-0-0.twitter.com	*.twitter.com	Twitter, Inc.

Convert

Domain Name -- Application

Domain Name	Alipay	Facebook	Twitter
gw.alipayobjects.com	1	0	0
mygw.alipay.com	1	0	0
www.facebook.com	0	1	0
api.facebook.com	0	1	0
api-20-0-0.twitter.com	0	0	2

Convert

Organization -- Application

Organization	Alipay	Facebook	Twitter
Alipay.com Co.,Ltd	2	0	0
Facebook, Inc.	0	2	0
Twitter, Inc.	0	0	2

Convert

Common Name -- Application

Common Name	Alipay	Facebook	Twitter
*.alipayobjects.com	1	0	0
*.alipay.com	1	0	0
*.facebook.com	0	2	0
*.twitter.com	0	0	2

MAAF -- Attribute Vector Generator

Domain Name -- Application

Domain Name	Alipay	Facebook	Twitter
gw.alipayobjects.com	1	0	0
mygw.alipay.com	1	0	0
www.facebook.com	0	1	0
api.facebook.com	0	1	0
api-20-0-0.twitter.com	0	0	2

Common Name -- Application

Common Name	Alipay	Facebook	Twitter
*.alipayobjects.com	1	0	0
*.alipay.com	1	0	0
*.facebook.com	0	2	0
*.twitter.com	0	0	2

Organization -- Application

Organization	Alipay	Facebook	Twitter
Alipay.com Co.,Ltd	2	0	0
Facebook, Inc.	0	2	0
Twitter, Inc.	0	0	2

Flow Example	
Domain Name	gw.alipayobjects.com
Common Name	*.alipayobjects.com
Organization	Alipay.com Co.,Ltd
Application Data Length	48, 64, 39

Attribute Vector = [1, 0, 0, 1, 0, 0, 2, 0, 0, 48, 64, 39]

MAAF -- Context Utilizer

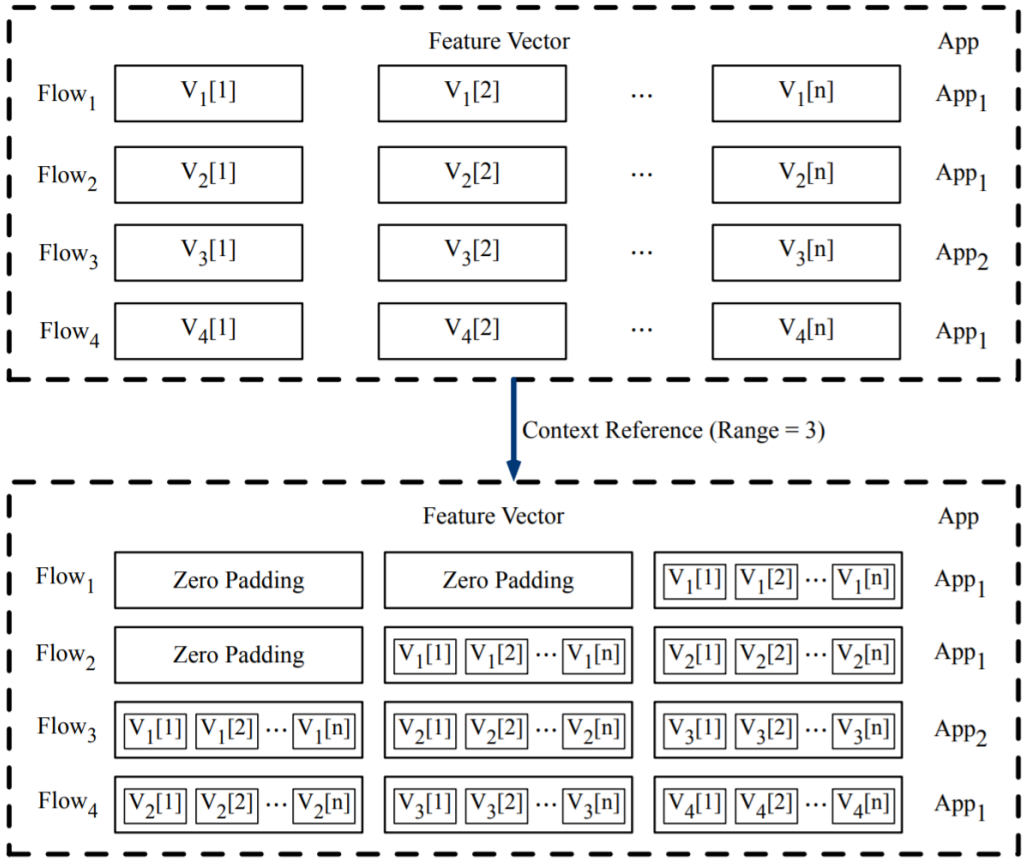


Fig. 4. The Process of Contextual Flow Utilization

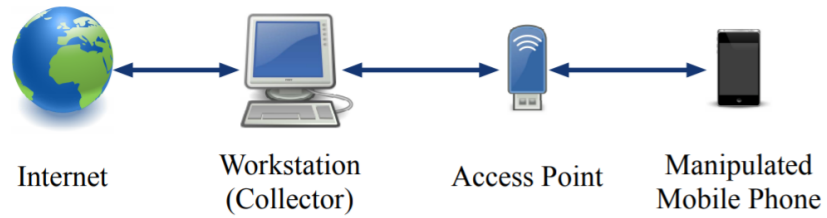
- Extend contextual information
- Attribute may be lacking
- Applications' flows cluster in the timeline

MAAF -- Classifier

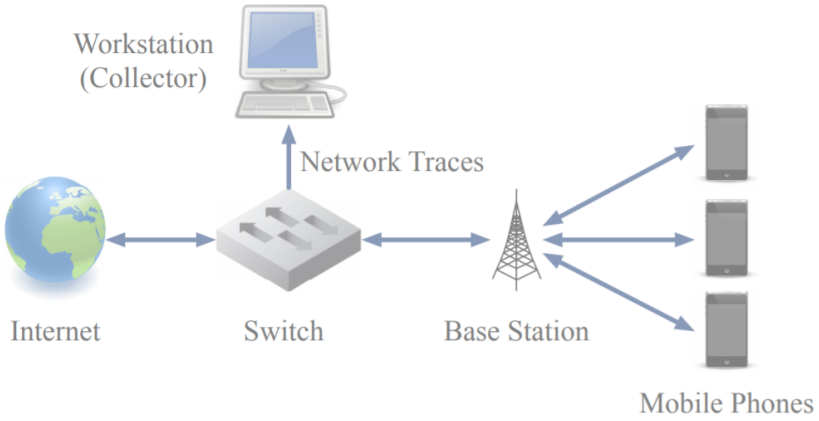
- Supervised classifier

- C4.5
- Random Forest
- XGBoost

Dataset Collection



(a) Active Scheme



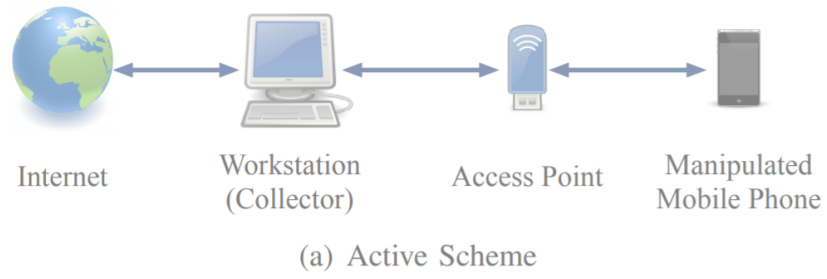
(b) Passive Scheme

Fig. 5. Mobile Application Traceset Collection Schemes.

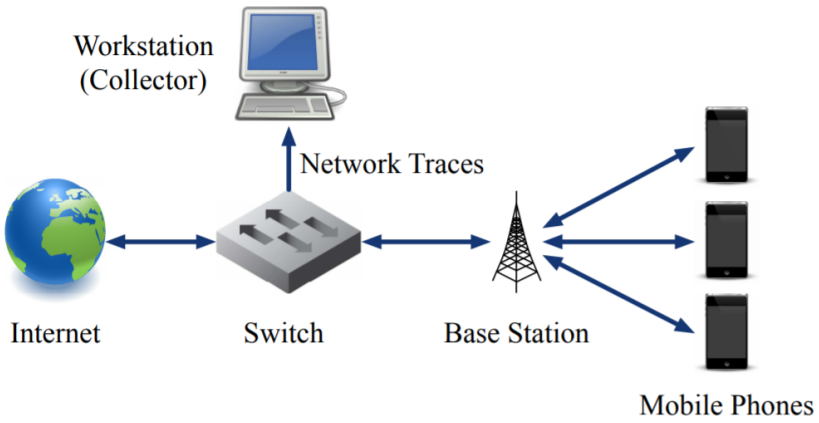
Active Scheme

- The phone is linked to the workstation via an access point
- The phone runs an application
- The workstation records the traces

Dataset Collection



(a) Active Scheme



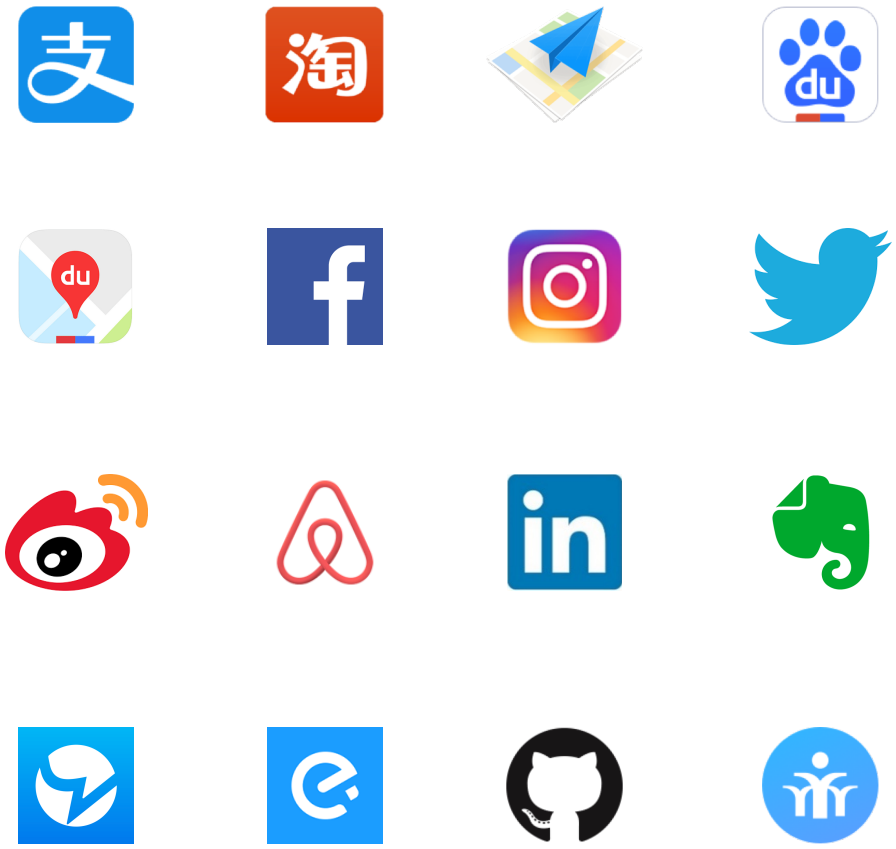
(b) Passive Scheme

Fig. 5. Mobile Application Traceset Collection Schemes.

Passive Scheme

- The workstation collects unlabeled traces through the port mirroring switch
- Label these unlabeled trace by matching application-related domain names

Dataset Introduction



16 Mainstream Applications

- Adopt active collection scheme
- Refer the application list in [1]
- Add three applications to study the classification of the applications from the same developer

[1] M. Shen, M. Wei, L. Zhu, and M. Wang, "Classification of encrypted traffic with second-order markov chains and application attribute bigrams," IEEE Transactions on Information Forensics and Security, vol. 12, no. 8, pp. 1830–1843, 2017.

Dataset Introduction

TABLE II
THE STATISTIC OF 16 APPLICATION TRACESETS

Developer	Application	Manually Collected Traceset				
		Flows	Packets	Domain	Cert	Both[1]
Alibaba	Alipay	5201	315234	16.4%	96.3%	97.3%
	Taobao[2]	3231	291348	93.9%	96.8%	99.4%
	AMap[2]	3624	114513	91.7%	98.8%	99.4%
Baidu	Baidu Search	4732	181971	52.5%	90.3%	94.3%
	Baidu Map[2]	5544	215920	40.0%	89.2%	93.8%
Facebook	Facebook	4148	526289	46.3%	82.2%	87.4%
	Instagram	4379	343809	27.0%	5.8%	31.8%
Twitter	Twitter	4463	167166	45.6%	89.7%	93.9%
Sina	Weibo	3817	127057	95.4%	95.2%	99.6%
Airbnb	Airbnb	5843	875837	76.0%	67.7%	82.2%
Linkedin	Linkedin	4203	160614	91.4%	91.8%	98.5%
Evernote	Evernote	7504	202557	98.4%	48.1%	98.5%
Blued	Blued	4833	478467	73.4%	55.6%	73.8%
Ele	Ele	6740	99193	98.9%	98.5%	99.9%
Github	Github	4431	151355	98.6%	96.4%	98.8%
Yirendai	Yirendai	4585	61356	98.1%	97.5%	99.2%
Total		77278	4312686	71.7%	79.9%	90.7%

- About 72% flows start from a domain name
- About 80% flows contain a certificate
- About 10% flows are not related with any domain name and certificate

Evaluation Preliminary

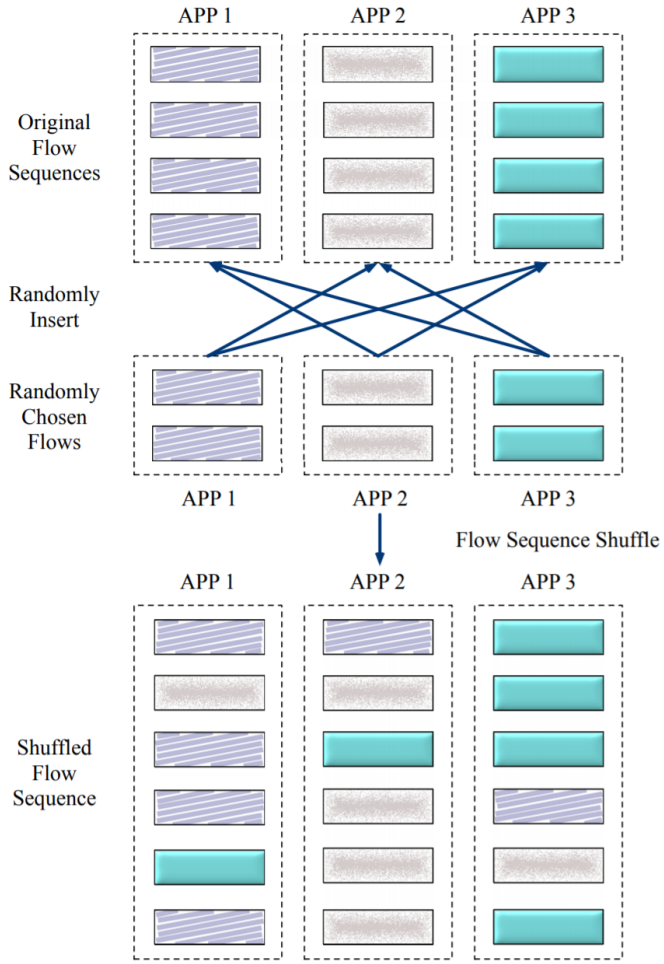
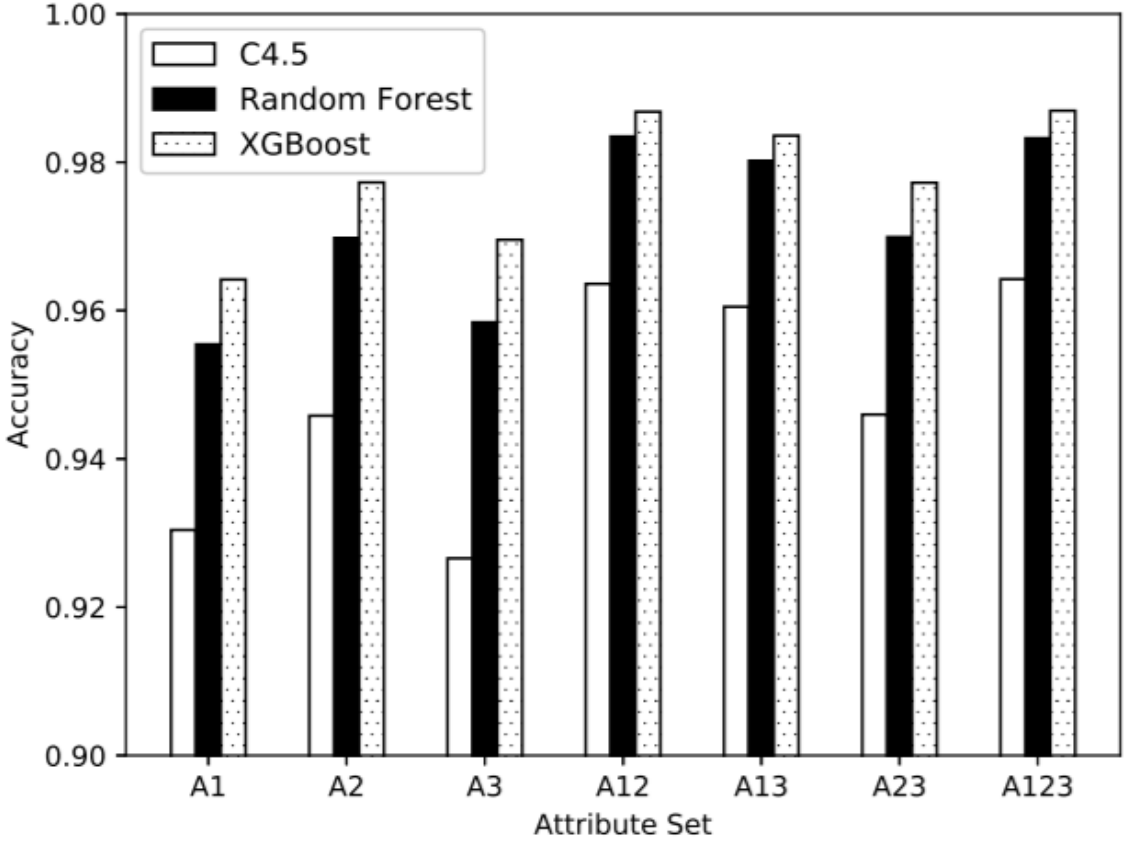


Fig. 9. The Process of Flow Sequence Shuffle

Flow Sequence Shuffle

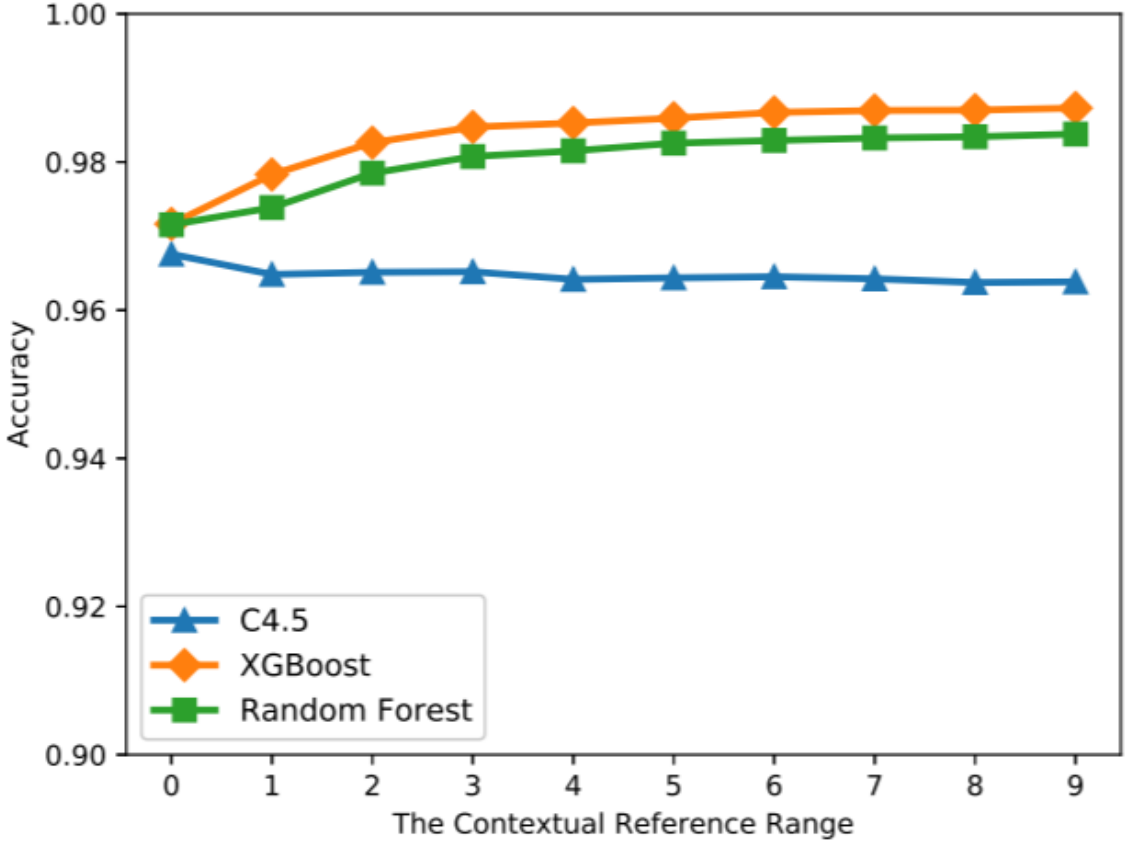
- Simulate contextual flows
- Select part of the trace for random insertion

Evaluation of MAAF



- Classification accuracy under different attribute sets
- A1 – Domain Name
- A2 – Common Name
- A3 – Organization

Evaluation of MAAF



- Classification accuracy under different contextual reference ranges

Compared with Other Approaches

TABLE IV
EXPERIMENTAL RESULTS OF DIFFERENT APPROACHES

Application	SOB		MaMPF		FS-Net		MAAF	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Alipay	0.7196	0.8062	0.8209	0.9449	0.9707	0.9650	**0.9859**	**0.9915**
Taobao	0.5221	0.6078	0.4739	0.8183	0.8727	0.8718	**0.9671**	**0.9696**
Amap	0.6910	0.7374	0.9321	0.9040	0.9502	0.9484	**0.9830**	**0.9767**
Baidu Search	0.6502	0.4556	0.8205	0.5515	0.8797	0.8809	**0.9742**	**0.9441**
Baidu Map	0.6505	0.7278	0.8427	0.6010	0.9006	0.9151	**0.9517**	**0.9817**
Facebook	0.8319	0.7992	0.8571	0.8205	0.9706	0.9623	**0.9891**	**0.9891**
Instagram	0.9127	0.8445	0.9272	0.8825	0.9819	0.9737	**0.9922**	**0.9888**
Twitter	0.8728	0.9159	0.9703	0.9005	0.9792	0.9729	**0.9903**	**0.9900**
Weibo	0.7807	0.8598	0.7565	0.9020	0.9532	0.9459	**0.9959**	**0.9907**
Airbnb	0.7115	0.5883	0.6598	0.9420	0.9536	0.9646	**0.9816**	**0.9905**
Linkedin	0.6423	0.7575	0.9719	0.7144	0.9648	0.9691	**0.9973**	**0.9970**
Evernote	0.9124	0.8722	0.9486	0.9782	0.9750	**0.9971**	**0.9974**	0.9970
Blued	0.7402	0.8883	0.8508	0.9415	0.9852	0.9645	**0.9938**	**0.9910**
Ele	0.9386	0.8298	0.9569	0.8789	0.9753	0.9606	**0.9951**	**0.9897**
Github	0.7801	0.7925	0.9782	0.7972	0.9871	0.9832	**0.9962**	**0.9964**
Yirendai	0.7464	0.5949	0.9780	0.8043	0.9683	0.9767	**0.9958**	**0.9954**
Acc/F1	0.7601	0.7519	0.8422	0.8370	0.9561	0.9537	**0.9869**	**0.9864**

- Accuracy – 98.69%
- F1 Score – 98.64%
- Outperforms state-of-the-art approaches

- For more details, please contact chenyige@iie.ac.cn

  or visit https://ychen.info

- Questions & Answers