Net-track: Generic Web Tracking Detection Using Packet Metadata

Dongkeun Lee[*], Minwoo Joo[†], and Wonjun Lee[*]

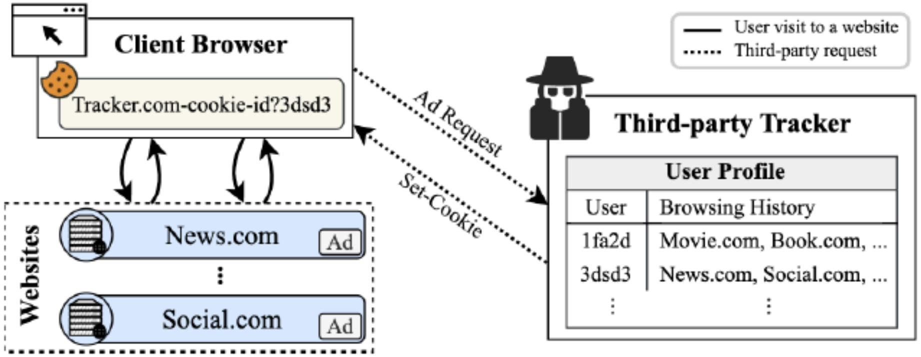
[*] Korea University, [†] Samsung Research

The ACM Web Conference 2023, Austin, TX, USA, April-May 2023

Dongkeun Lee
Network and Security Research Lab. (**NetLab**)
School of Cybersecurity
Korea University, Seoul, Korea
dklee98 AT korea.ac.kr
https://netlab.korea.ac.kr

Threats of Web Tracking

- **Third-party trackers breach users' privacy**
  - Collect information such as user's location or browsing history
  - 22 trackers per site on average, with more than 81,000 of them in total
- **COTS products are also equipping privacy-protecting features to combat trackers**
  - e.g., Mozilla Firefox, Apple Safari, Brave



User	Browsing History
1fa2d	Movie.com, Book.com, ...
3dsd3	News.com, Social.com, ...
⋮	⋮

Limitations of Existing Solutions

- Coarse-grained or platform-dependent
  - Require an instrumented browser for dynamic feature analysis

- Deep packet inspection (DPI)-based solutions are ineffective against encrypted traffic
  - 79.8% of all websites use HTTPS as a default

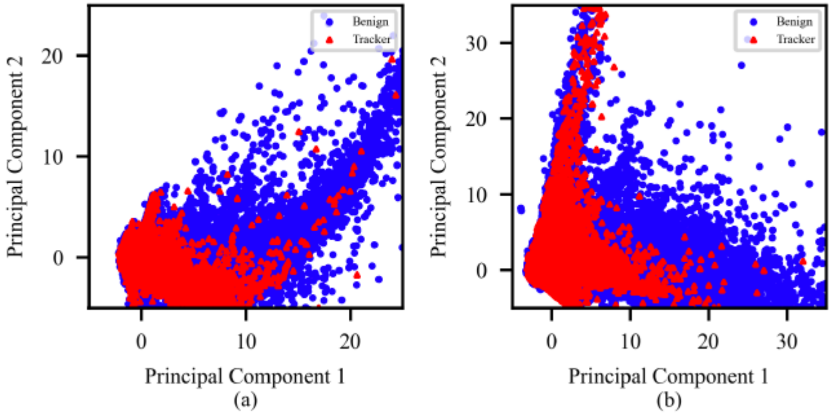
Our Motivation

- Key observation
  - Trackers' intrinsic functionalities generate distinctive traffic patterns
    - i.e., collecting and sending user data
- Collecting and analyzing real-world traffic
  - Visit top-20k Alexa websites
  - Divide the captured traffic in terms of connection
    - Capture each client-server interaction with diverse third parties as well as with the host
  - Label each trace as tracker or benign based on filter lists
    - EasyList and EasyPrivacy
  - 222,009 benign traffic traces and 126,664 tracker traffic traces

Difference in Traffic Patterns

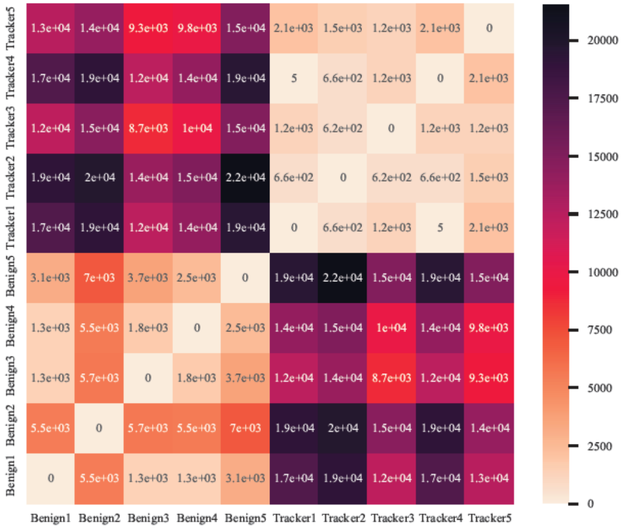
- Statistics from traffic traces
  - Principal component analysis (PCA) on 62 statistical features

- Similarity between packet sequences
  - Dynamic time warping (DTW) between random traces



(a)

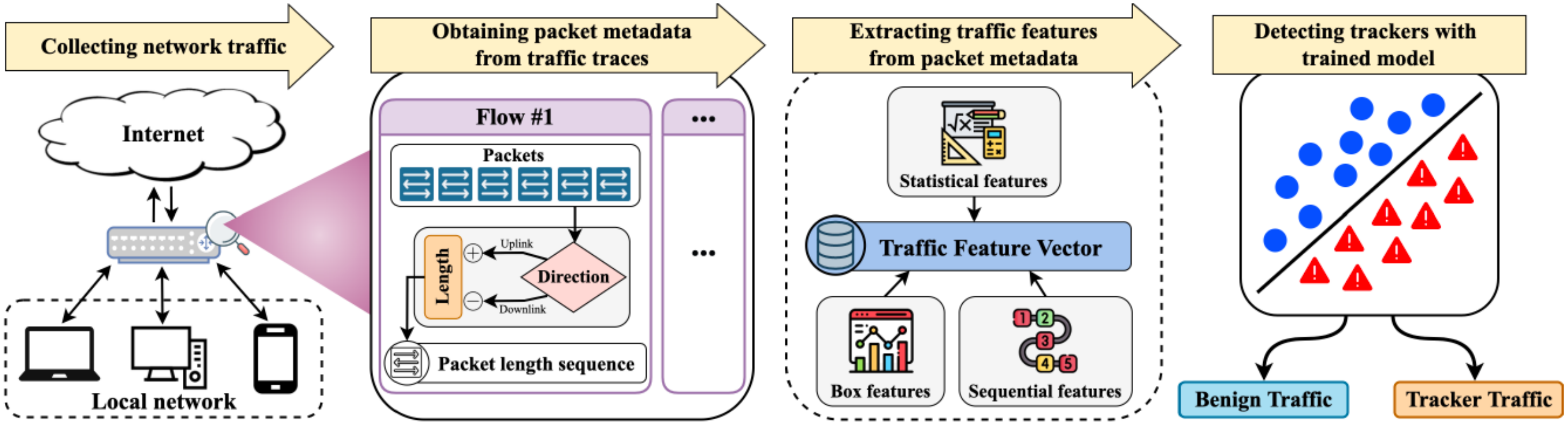
(b)



	Benign1	Benign2	Benign3	Benign4	Benign5	Tracker1	Tracker2	Tracker3	Tracker4	Tracker5
Tracker5	1.3e+04	1.4e+04	9.3e+03	9.8e+03	1.5e+04	2.1e+03	1.5e+03	1.2e+03	2.1e+03	0
Tracker4	1.7e+04	1.9e+04	1.2e+04	1.4e+04	1.9e+04	5	6.6e+02	1.2e+03	0	2.1e+03
Tracker3	1.2e+04	1.5e+04	8.7e+03	1e+04	1.5e+04	1.2e+03	6.2e+02	0	1.2e+03	1.2e+03
Tracker2	1.9e+04	2e+04	1.4e+04	1.5e+04	2.2e+04	6.6e+02	0	6.2e+02	6.6e+02	1.5e+03
Tracker1	1.7e+04	1.9e+04	1.2e+04	1.4e+04	1.9e+04	0	6.6e+02	1.2e+03	5	2.1e+03
Benign5	3.1e+03	7e+03	3.7e+03	2.5e+03	0	1.9e+04	2.2e+04	1.5e+04	1.9e+04	1.5e+04
Benign4	1.3e+03	5.5e+03	1.8e+03	0	2.5e+03	1.4e+04	1.5e+04	1e+04	1.4e+04	9.8e+03
Benign3	1.3e+03	5.7e+03	0	1.8e+03	3.7e+03	1.2e+04	1.4e+04	8.7e+03	1.2e+04	9.3e+03
Benign2	5.5e+03	0	5.7e+03	5.5e+03	7e+03	1.9e+04	2e+04	1.5e+04	1.9e+04	1.4e+04
Benign1	0	5.5e+03	1.3e+03	1.3e+03	3.1e+03	1.7e+04	1.9e+04	1.2e+04	1.7e+04	1.3e+04

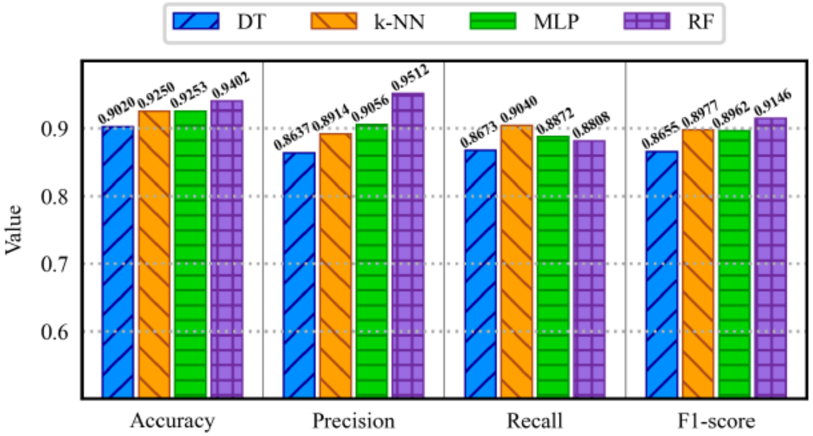
Design of Net-track

- Net-track utilizes packet metadata from network traffic
  - Platform-independent and encryption-agnostic
- Three types of features from packet length sequence
  - Statistical features, box features, and sequential features



Evaluation

- **Detection performance of Net-track**
  - Net-track with random forest is the highest both in accuracy (94.02%) and precision (95.12%)
  - Performance attained only with side-channel data from network traffic
    - Net-track does not require analyzing resources loaded at the application layer nor inspecting contents in the HTTP payloads



	DT	*k*-NN	MLP	RF
Training Time (s)	27.7080	0.0516	1638.53	73.3319
Inference Time (ms)	0.0011	12.813	0.0209	0.0163

Evaluation

- Discovering new trackers
  - Case study on 200 samples of randomly selected *false positives*
    - i.e., Net-track classified as tracker though labeled as benign
  - 34.5% of these ‘detection errors’ were new, unknown trackers
    - Domain changes
      - e.g., *mc.yandex.ru* → *mc.yandex.com*
    - Cookie syncing
      - e.g., *x.dlx.addthis.com*
    - Tracking script from first-party domain
      - e.g., *afterpay-1.x.js* on afterpay.com
  - Manually curated filter lists fail to adapt to trackers’ evasions
    - 37.68% of these newly found trackers are still unenrolled
      - Newest version of the filter lists (10+ months after data collection)

Conclusions and Future Work

- Net-track enables encryption-agnostic, platform-independent detection of trackers
  - 94.02% accuracy using only packet metadata
- Net-track can discover many new trackers unrecognized by existing filter lists
  - 34.5% of false positives were indeed trackers that have not yet been discovered
- We aim to apply Net-track as a source of information that feeds other systems
  - Net-track can update firewall rules or tracking domain lists to block subsequent flows in the network

Thank you!

- For more information
  - Network and Security Research Lab. (NetLab), Korea University, Seoul, Korea https://netlab.korea.ac.kr
  - Prof. Wonjun Lee wlee@korea.ac.kr
  - Dongkeun Lee dklee98@korea.ac.kr