DNS-based User Tracking
(Attacks and Defenses)

Presenter: Zhou Li

Data-driven Security and Privacy (DSP) Lab
EECS department, University of California, Irvine
DINR’23, 02/22/2023
Outline

Background
Threat Model
Attack: DSCorr
Defense: LDPResolve
Conclusion

Background: DNS-based User Tracking

- Users send DNS queries before almost every network activity.
- Different users have different preferences.
- **Can we track a user by their DNS queries?**
  - Privacy violation

```
play.google.com
ieeexplore.ieee.org
telegram.org
...

www.apple.com
www.icloud.com
www.netflix.com
...
```
Goal: track users based on their DNS queries
- E.g., public recursive resolvers

Challenge: a user’s identifier, aka, source IP keeps changing
- E.g., DHCP, moving from one access point to another, cellular network

This is an inference/classification problem
- Attacker’s input: Session, DNS queries from one source IP in a time window
- Attacker’s output: user ID (real or pseudo)
Threat Model

- Formalization of DNS-based user tracking
  - Link different sessions of a same user from different source IPs.
Existing Attacks

- Supervised, semi-supervised or unsupervised learning
  - Feature extraction from DNS queries
  - Bayesian classifier, KNN, Dirichlet multinomial mixture
  - Fixed threshold
- All assuming a closed-world setting
  - The attacker already knows the set of users before tracking
- How about open-world setting?
  - Unknown user can be encountered during tracking

Our Attack: DSCorr

**STEP 1** Convert domains to domain embedding vectors.

**STEP 2** Build user profiles (clusters of sessions) from a labeled session set.

**STEP 3** Given an unlabeled session $s$, identify $k$ nearest labeled clusters through a data-sketching process.

**STEP 4** Compute the nearest distance between session $s$ and $k$ clusters under user-centric threshold for open-world setting.
Domain Embedding

- **Domain distance**: 0 or 1 by previous works
  - Too coarse-grained
- **Fine-grained domain distance** based on domain context
  - Domains usually visited together should have small distance
- **Use Word2Vec (NLP)** to build domain embedding vectors
  - Domain -> Word
  - DNS session -> Context

**Text Corpus**

```
Window Size = 2

The  quick  brown  fox  jumps over the red dog
The  quick  brown  fox
The  quick  brown  fox  jumps over the red dog
The  quick  brown  fox  jumps over the red dog
```

**Training Samples**

```
( The , quick )
( The , brown )
( quick , the )
( quick , brown )
( quick , fox )
( brown , the )
( brown , quick )
( brown , fox )
( brown , jumps )
( fox , quick )
( fox , brown )
( fox , jumps )
( fox , over )
```

SkipGram of Word2Vec
Evaluation of DSCorr

- Different tracking methods: Jaccard/Cosine/Bayesian Classifier/DSCorr
- Different feature: unigram & bi-gram
- Different number of sessions in labeled set for each user

<table>
<thead>
<tr>
<th>#</th>
<th>jac</th>
<th>cos</th>
<th>bay</th>
<th>ja-bi</th>
<th>co-bi</th>
<th>ba-bi</th>
<th>DSCorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42.2</td>
<td>40.7</td>
<td>37.4</td>
<td>45.4</td>
<td>40.1</td>
<td>36.5</td>
<td>52.6</td>
</tr>
<tr>
<td>2</td>
<td>56.0</td>
<td>52.8</td>
<td>54.8</td>
<td>59.2</td>
<td>52.8</td>
<td>54.3</td>
<td>67.5</td>
</tr>
<tr>
<td>3</td>
<td>67.2</td>
<td>60.3</td>
<td>65.7</td>
<td>67.2</td>
<td>60.3</td>
<td>65.8</td>
<td>74.4</td>
</tr>
<tr>
<td>5</td>
<td>74.8</td>
<td>69.3</td>
<td>76.3</td>
<td>74.8</td>
<td>69.3</td>
<td>76.8</td>
<td>80.5</td>
</tr>
<tr>
<td>10</td>
<td>78.8</td>
<td>78.0</td>
<td>86.2</td>
<td>82.7</td>
<td>77.6</td>
<td>87.3</td>
<td>87.4</td>
</tr>
</tbody>
</table>

Table. Tracking accuracy under closed-world setting.

Findings:

- DSCorr is more effective under closed-world setting, especially when there’s less labeled data.
- Auto-threshold works. It allows DSCorr to work under open-world setting.
- Popular domains affect user tracking.

Fig. Tracking accuracy under open-world setting.
Defense: Local Differential Privacy (LDP)

- The data collector is untrustworthy
- Noises added to the clients' data before collection
- LDP guarantees the information leakage after noises are bounded by $\epsilon$
- Used by Apple to collect emoji usage …

**Definition 1** ($\epsilon$-Local Differential Privacy [89]). An algorithm $A$ satisfies $\epsilon$-local differential privacy ($\epsilon$-LDP), where $\epsilon > 0$, if and only if for any pair of input $x_1$ and $x_2$, we have

$$\forall y \in \text{Range}(A) : \frac{\Pr[A(x_1) = y]}{\Pr[A(x_2) = y]} \leq e^\epsilon$$  \hspace{1cm} (1)

where $\text{Range}(A)$ denotes the set of all possible output results of an algorithm $A$. 
Our Defense Method: LDPResolve

Fig. Design of ULDP

Fig. Modified URR

$\epsilon$-LDP

$X_S$, $\epsilon_1$, $\epsilon_2$)-Utility-optimized Randomized Response

Popular domains

Privacy budget

Primary Resolver

Alternative Resolvers

Design of LDPResolve

google.com
facebook.com
npop0.com
npop1.com

Alternative Resolvers

Primary Resolver
Design of LDPResolve

LDPResolve

- google.com → google.com
- facebook.com → bbc.com
- npop0.com → cnn.com
- npop1.com → npop1.com

Primary Resolver

Alternative Resolvers

Design of LDPResolve

LDPPResolve

google.com → google.com
facebook.com → bbc.com
npop0.com → cnn.com
npop1.com → npop1.com

Primary Resolver

google.com
bbc.com
cnn.com
npop1.com

Alternative Resolvers

facebook.com
npop0.com

Evaluation of LDPResolve: Privacy

Fig. Tracking Accuracy given different sensitive set size (i.e., 2k and 10k)
Key Terms

Ns: Size of sensitive set

\( \varepsilon_1 \): Overall privacy Budget

\( \varepsilon_2 \): Privacy Budget for sensitive domains. \( \varepsilon_2 \leq \varepsilon_1 \)
Conclusion

- DNS-based user tracking is a real privacy concern
  - Existing works are effective under closed-world setting.
  - Our attack DSCorr is effective in both closed-world and open-world settings.

- Popular domain is the key to DNS-based user tracking.

- LDPRReslove could be effective in terms of defeating tracking.
  - LDP ensures the privacy leakage is bounded regardless of the attack methods
Sponsors

- NSF
- DARPA
- AMP
- EAGER SaTC
- IMR
- CAREER

Team (DSP Lab)

Zhou Li  Joann Chen  Jiacen Xu  Qifan Zhang  Xuesong Bai  Xiang Li