Realistic Website Fingerprinting Attacks
John Mikos and Eric Chan-Tin

Problem & Motivation
- Website fingerprinting [1] attacks allow an adversary to identify the website visited even if the victim utilizes anonymity network, e.g. Tor [2] and HTTPS.
- However, real users visit multiple websites (overlapping website visits) at the same time.
- Visiting multiple websites add enough noise to mitigate website fingerprinting attacks.

Goals

<table>
<thead>
<tr>
<th>Overlapping Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website 1</td>
</tr>
<tr>
<td>Overlapped Part</td>
</tr>
<tr>
<td>Website 2</td>
</tr>
</tbody>
</table>

Datasets
- January to February 2019
- 118 unique websites from Alexa [3] top 200 sites
- In between 120 and 150 visits (instances) to each website in a given directory
- Overlap: 10 seconds and 20 seconds
  - First website is randomly chosen from the 118 sites
  - Second website is randomly chosen from the remaining 117 sites
  - 1,000 instances each

Proposed Algorithms

Key idea: create a reference list that holds the non-unique outgoing packet sizes

Create Reference List
- Loop through all websites and count the number of occurrences of each packetsize.
- Choose a threshold value
- Loop through the packetsizes present in the list and remove any packetsizes whose number of occurrences >= that threshold value

Detect Overlap
- Read through all the websites dataset
- For each instance of the websites
  - Building off the reference list, take the timestamp:packetsize data and add it to a unique set S only if the packet size is not present in the reference list
  - Loop through every packet size in set S to find the packetsize with the earliest timestamp
  - If a packetsize was found, record it and use the timestamp information to calculate average time, median time, standard deviation, and percent correct.
- Percent correct is the percentage of websites with a predicted average time ±1 second from the real overlap time (10 seconds or 20 seconds)

Graphs

The number of points are the datapoints chosen from a whole collection of recordings. The graphs represent the most important changes throughout all data.

[20v10] Percent Correct vs. Occurrence (First 1265 Points)

Fig 1. The 20s directory as reference against the 10s directory data to see percent correct achieved.

[20vTrain] Average Time vs. Occurrence (First 960 points)

Fig 2. The 20s directory as reference against the Training directory data to see the average time achieved.

[20v10] Average Time vs. Occurrence (First 1250 Points)

Fig 3. The 20s directory as reference against 10s directory data to see the average time achieved.

[20v20] Average Time vs. Occurrence (First 500 Points)

Fig 4 + 5. The 20s directory as reference against itself. Two different views are presented to show how the results behave as number of occurrences increases.

[20v20] Average Time vs. Occurrence (First 1250 Points)

Acknowledgements
- LUROP: Provost Fellowship for funding the continuation of this research.
- We thank Tao Chen for collecting and providing us the dataset.

References
2. Tor Project, [https://www.torproject.org](https://www.torproject.org)