Realistic Website Fingerprinting Attacks

John Mikos and Eric Chan-Tin

Preparing people to lead extraordinary lives

Preparing people to lead extraordinary lives

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

- Create an algorithm to identify where the start of the overlap (start of second website visit) is
- This allows to split a multiple-websites-data into two websites where typical website fingerprinting can be applied

Datasets

- January to February 2019
- 118 unique websites from Alexa [3] top 200 sites
- 1,000 visits (instances) to each website (JOHN: check how many lines per file and replace 1,000 with that number)
- 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.

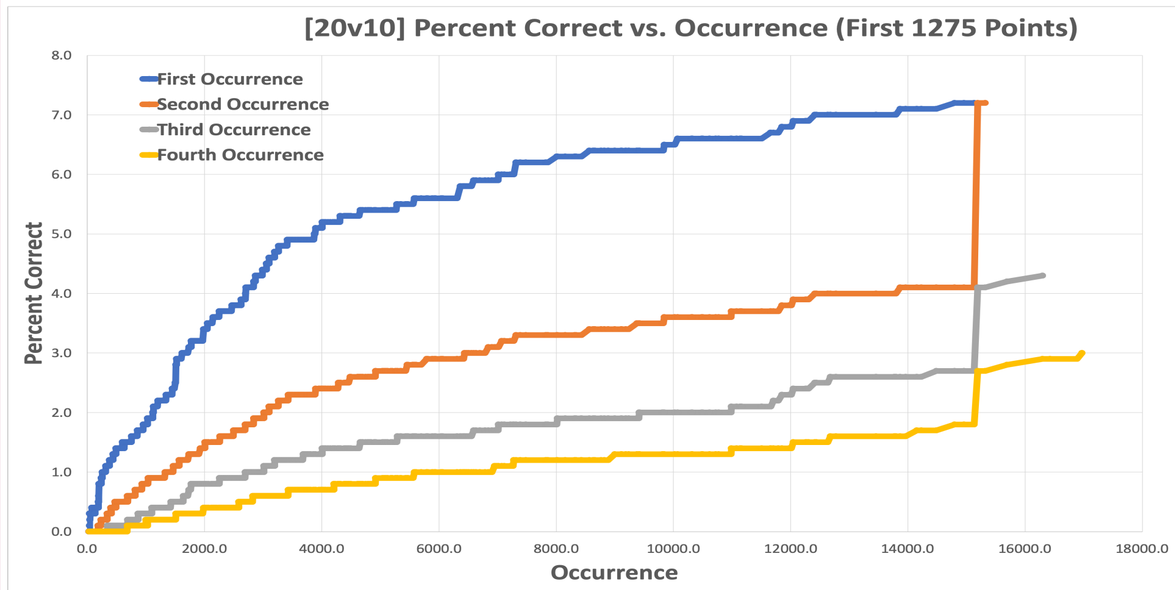

Fig 1. The 20s directory as reference against the 10s directory data to see percent correct achieved.

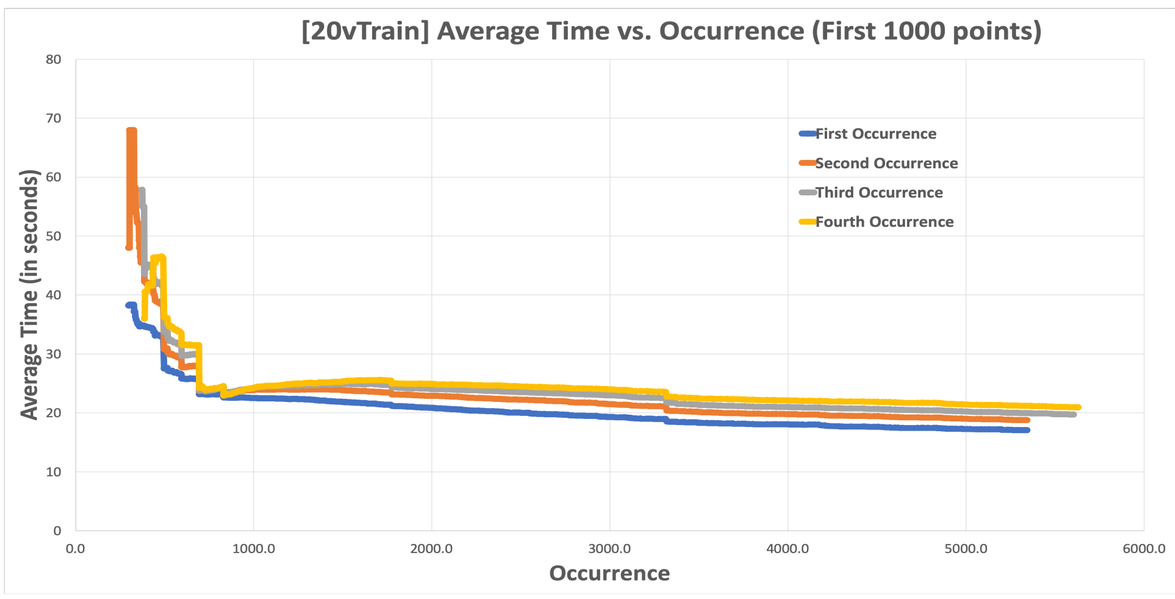

Fig 2. The 20s directory as reference against the Training directory data to see the average time achieved.

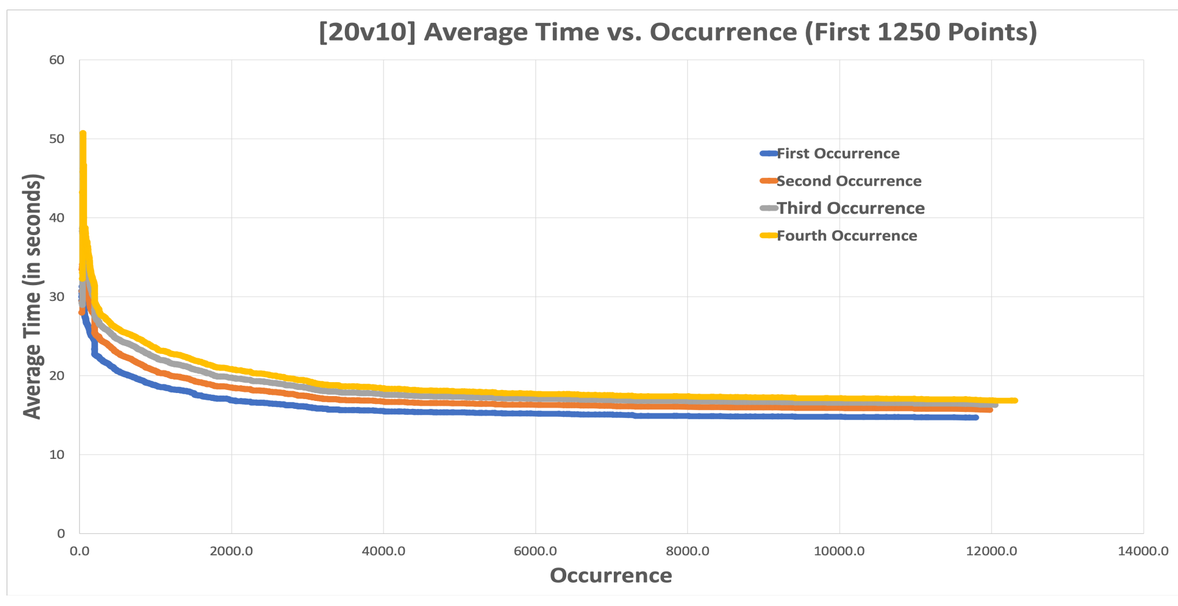

Fig 3. The 20s directory as reference against 10s directory data to see the average time achieved.

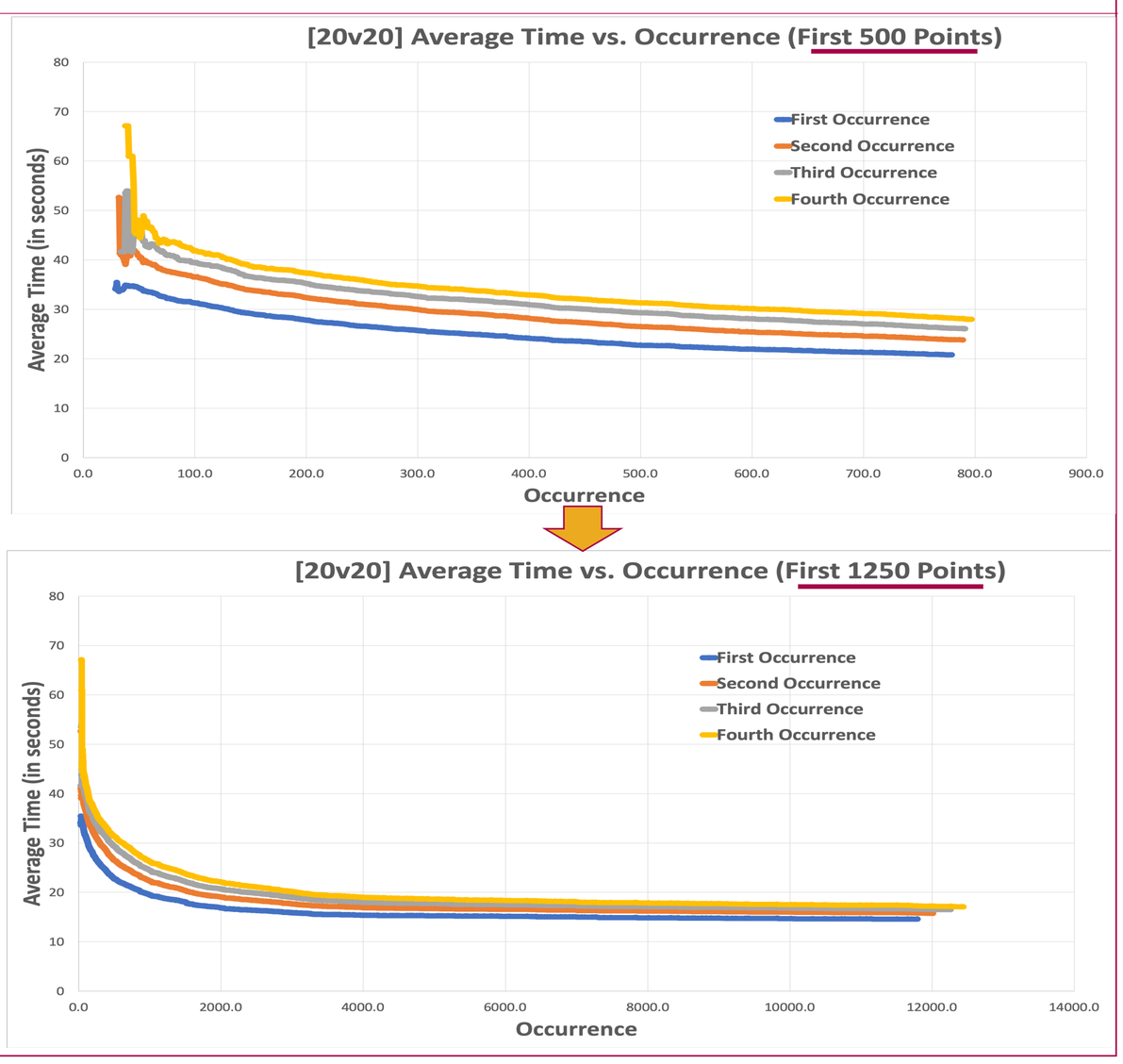

Fig 4 + 5. The 20s directory as reference against itself. Two different views are presented to show how the results behave as number of occurrences increases.

Acknowledgements

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

References

1. Panchenko, A., Lanze, F., Zinnen, A., Henze, M., Pennekamp, J., Wehrle, K. and Engel, T. (2019). Website Fingerprinting at Internet Scale. NDSS, 2019
2. Tor Project, https://www.torproject.org
3. Alexa top sites, https://www.alexa.com/topsites