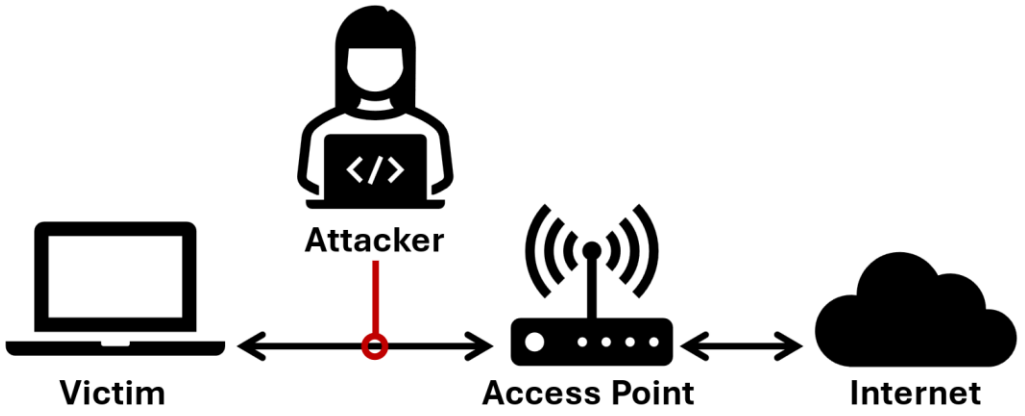
Website Fingerprinting

Josh Honig and Nathan Ferrell

Loyola University Chicago

What is Website Fingerprinting?

- Form of Network Traffic Analysis
- Without breaking encryption
- Requires trivial network visibility
- Create a fingerprint for each webpage:
  - Packet direction, frequency/pattern, size
  - aka packet metadata
- Use ✨ Machine Learning ✨
  - Non-resource-intensive model (CPU)
- Not limited to webpages
- Website Fingerprinting ≠ Browser/Device Fingerprinting



LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Goals

Based only on encrypted network traffic metadata, can we...

Wikipedia

guess which Wikipedia article a user is visiting in the browser?

NY Times

guess which NY Times article a user is visiting in the browser?

Google

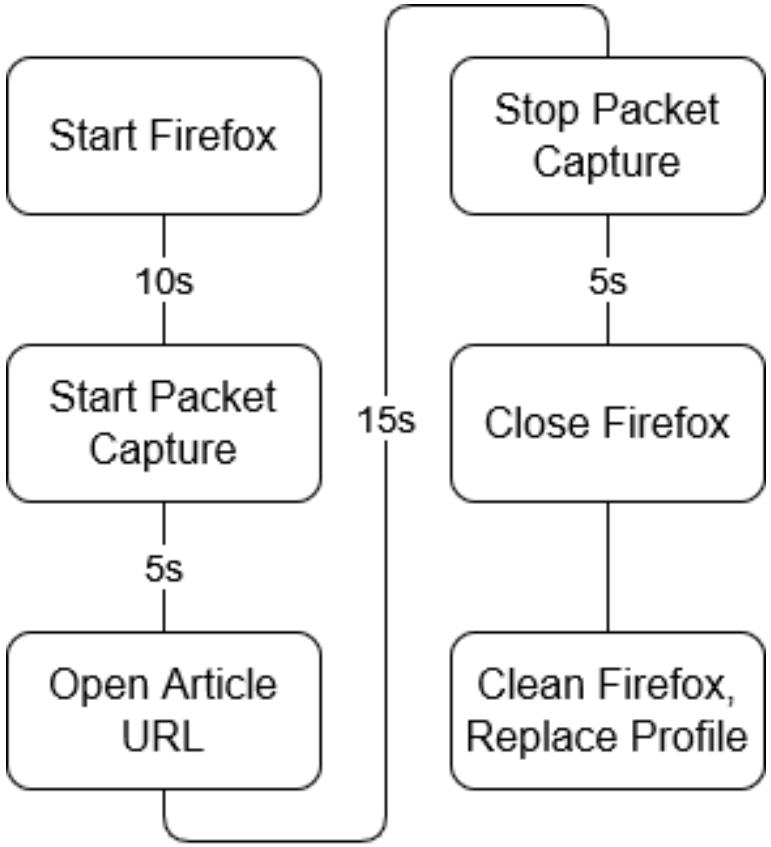
guess what a user is typing in the search box; guess their query?

MetaQuest VR

guess which actions a user takes on a MetaQuest VR headset?

Data Collection: Wikipedia, NYTimes

- Firefox
- Packet Capture
- 1,000 URLs
- Wikipedia
- Issues with New York Times
  - Storage requirements
  - User login
  - Bot detection
  - Time to collect samples



LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Data Collection: Google

- Google Trends to collect ~1,000 queries
- Emulate user searching from home page
  - Similar to Wikipedia, NYTimes
  - Captured Autocomplete, then Search data
- Selenium, Firefox
- Limitations
  - Bot detection
  - Collection timeframe
  - Dynamic results

Data Collection: MetaQuest VR

- Manually Collected Data
- Wi-Fi Hotspot
- Wireshark
- Example User Actions
    - Logging into Google
    - Searching “football” on Google
    - Logging into Netflix
    - Searching Netflix for “Spiderman”
    - Joining Beatsaber Party
    - Playing Game of Beatsaber

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Data Pre-Processing

Outgoing
39 bytes

Outgoing
50 bytes

Empty

[39, 50, -382, 0, -14]

Incoming
382 bytes

Incoming
14 bytes

Machine Learning: Feature Engineering

- Training on just raw data produced poor results
  - Introduced limitations on number of features
  - <1% accuracy 😟
- 34 Machine Learning features
  - Number of [Incoming/Outgoing/All] Packets
  - Average [Incoming/Outgoing/All] Packet Size
  - Largest/Smallest Packet Size
  - # Packets in Highest [Incoming/Outgoing] Streak
  - Cumulative Sum (Panchenko et al., 2016)
- 100+ packet size features

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Machine Learning: Model

- Random Forest Classifier
- Unique model for each dataset
- Closed set
- Techniques to increase accuracy:
  - Grid Search to determine best hyperparameters
  - 5-fold cross-validation
  - Scaling
  - Normalization
- Machine Learning != Deep Learning

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Results: Wikipedia

- Wikipedia
  - 1,000 articles, visited 20 times each (20,000 samples)
- Significant Features (195 total)
  - Standard Deviation of Outgoing Packets (9.549%)
  - Total Incoming Packet Size (7.343%)
  - Total Outgoing Packet Size (4.812%)
- Maximum Accuracy of **69.85%**
  - Base case accuracy: 0.001%
- Conclusions
  - Lack of dynamic content makes for easy fingerprinting

LOYOLA UNIVERSITY CHICAGO 1870 AD · MAJOREM · DEI · GLORIAM

Results: New York Times

- New York Times
  - 1,000 articles, visited 20 times each (20,000 samples)
- Significant Features (244 total)
  - Total Number of Packets, Excluding 0-byte Packets (4.301%)
  - Average Outgoing Packet Size (2.390%)
  - 7/10 top features concerned Outgoing Packets
- Maximum Accuracy of **48.0%**
  - Base case accuracy: 0.001%
- Conclusions
  - Multimedia & dynamic content impact analysis

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Results: Google Autocomplete

- Google Autocomplete
  - 1,143 queries, performed 25 times each (~28,000 samples)
- Significant Features (186 total)
  - Incoming Packet Sizes, 25th Percentile (1.230%)
  - Lack of consistency in direction among significant features
  - Low weights among all features
- Maximum Accuracy of **22.70%**
  - Base case accuracy: 0.0875%
- Conclusions
  - We knew this would be difficult

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Results: Google Search

- Google Search
  - 1,143 queries, performed 25 times each (~28,000 samples)
- Significant Features (311 total)
  - Average Outgoing Packet Size & Outgoing Packet Size, 90th Percentile (0.691%)
  - First 5 features concerned outgoing packets
  - Again, low weights among all features
- Maximum Accuracy of **15.14%**
  - Base case accuracy: 0.0875%
- Conclusions
  - Limitations may have affected accuracy
  - Page optimization

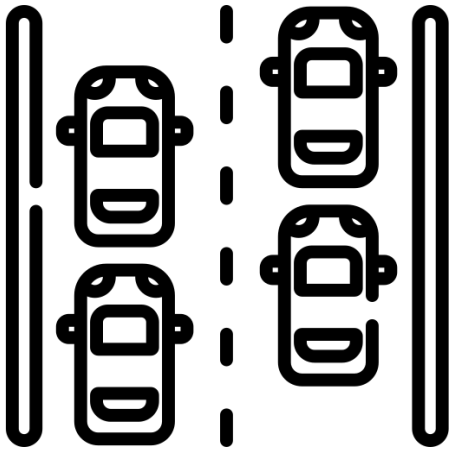
LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Results: MetaQuest VR

- MetaQuest Virtual Reality Headset
  - 14 actions, performed 5 times each (70 samples)
- Some traffic unencrypted, but not identifying
- Significant Features (634 total)
  - Standard Deviation of Incoming Packets (4.786%)
  - Total Incoming Packet Size (4.179%)
  - 4/5 top features concerned Incoming Packets
- Maximum Accuracy of **90.91%**
  - Base case accuracy: 7.14%
- Conclusions
  - Applications using HTTPS vulnerable to fingerprinting

LOYOLA UNIVERSITY CHICAGO
1870
AD · MAJOREM · DEI · GLORIAM

Future of Website Fingerprinting

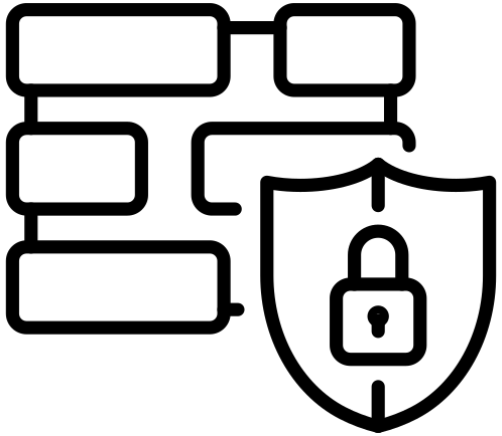


Developers should consider their network traffic patterns

Sensitive application?

Too uniform? Predictable? Noisy?

Known issue

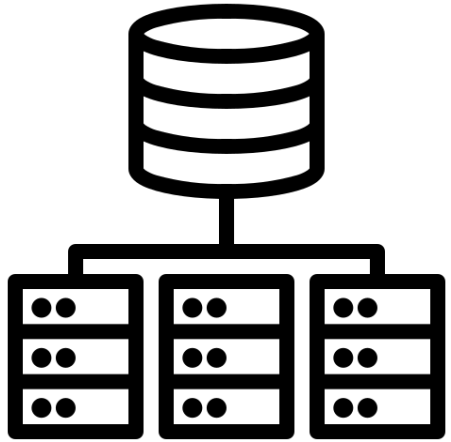


Defending against network traffic analysis attacks

Add padding, noise

Decreases efficiency,

Increases data use



Future work to improve attack effectiveness

Larger datasets

Live model training

Different mediums

LOYOLA UNIVERSITY CHICAGO

1870

AD · MAJOREM · DEI · GLORIAM