Lightweight Fingerprint Attack and Encrypted Traffic Analysis on News Articles

David Hasselquist, Linköping University & Sectra Communications, Sweden
Martin Lindblom, Linköping University, Sweden
Niklas Carlsson, Linköping University, Sweden

Proc. IFIP Networking, Catania, Italy, June 2022
Motivation

» Most of our news obtained online today

» The news we read can reveal much about us

» Users should be able to obtain independent news without adversary monitoring or control

» An adversary capable of extracting small fraction of our obtained news presents a privacy threat
Example: news bias
Examples: political misinformation

Political ads during the 2020 presidential election cycle collected personal information and spread misleading information

Sarah McQuate and Rebecca Gourley

UW News

POSTED UNDER: ENGINEERING, INTERACTIVE, NEWS RELEASES, POLITICS AND GOVERNMENT, RESEARCH, TECHNOLOGY
Examples: political misinformation
Examples: news filtering

China blocks 23 per cent of 215 accredited foreign news sites, watchdog says

- Citizens denied access to 31 per cent of news organisations that publish primarily in English, according to research by GreatFire.org and Foreign Correspondents’ Club of China
- Digital blockade ‘runs counter to the ethos of internet openness’, FCCC says
Examples: news filtering

China reportedly blocks access to US news sites

The Great Firewall of China has taken down access to The Guardian, The Intercept, NBC News and HuffPost, a report says.
Examples: news filtering

China reportedly blocks access to US news

The Great Firewall of China has taken action to block access to The Guardian, The International News and HuffPost, a report says.

Kazakhstan government is intercepting HTTPS traffic in its capital

This marks the third time since 2015 that the Kazakh government is mandating the installation of a root certificate on its citizens’ devices.
Examples: news filtering

Russia blocks access to BBC and Voice of America websites

March 4, 2022
12:34 PM GMT+1
Last Updated 3 months ago

Media & Telecom

The Kazakh government is mandating the installation of a root certificate on its citizens' devices.
Contributions

» Design and evaluation of lightweight framework
  » Identify individual browsed news articles (internal pages) despite encryption
  » Separate between articles delivered over same infrastructure (e.g., CDN)

» Demonstrate that naive use of HTTPS is not enough to protect users’ privacy
  » X.509 certificate size (encrypted with TLS 1.3)
  » Web document size

» Provide insights into why websites are more/less resilient to the attack

» Real-world scenario using Twitter

» Provide insights for websites and users to better protect their privacy
System overview
TLS record size extraction

<table>
<thead>
<tr>
<th>TCP</th>
<th>header</th>
<th>payload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
TLS record size extraction
TLS record size extraction

» Handshake: [0x16, 0x03, m]

» Application data: [0x17, 0x03, m_a]

\[ m \in \{0x00, 0x01, 0x02, 0x03\} \]

### TLS Record

<table>
<thead>
<tr>
<th>Byte</th>
<th>0</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1..4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5..n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n..</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Content type
- Version
- Length
- Payload

TCP

<table>
<thead>
<tr>
<th>TCP</th>
<th>header</th>
<th>payload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TLS stream

<table>
<thead>
<tr>
<th>TCP</th>
<th>header</th>
<th>payload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ m \in \{0x00, 0x01, 0x02, 0x03\} \]
TLS record size extraction

» Handshake: [0x16, 0x03, \(m\)]

» Application data: [0x17, 0x03, \(m_a\)]

\[m \in \{0x00, 0x01, 0x02, 0x03\}\]

<table>
<thead>
<tr>
<th>Byte</th>
<th>+0</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Content type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1..4</td>
<td>Version</td>
<td>Length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5..n</td>
<td>Payload</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n..</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TLS Record

TCP
header | payload

TLS stream

\[m \in \{0x00, 0x01, 0x02, 0x03\}\]
For each repeated connection, the certificate is delivered

- in similar TLS record index
- with similar TLS record size

<table>
<thead>
<tr>
<th>Domain</th>
<th>Certificate size</th>
<th>Certificate index</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York Times</td>
<td>$C_s \in {5176}$</td>
<td>$C_i \in {1, 2}$</td>
</tr>
<tr>
<td>Yahoo</td>
<td>$C_s \in {5253, 4774}$</td>
<td>$C_i \in {2, 4}$</td>
</tr>
<tr>
<td>Fox News</td>
<td>$C_s \in {2933, 2934, 2935}$</td>
<td>$C_i \in {2, 4}$</td>
</tr>
<tr>
<td>MSN</td>
<td>$C_s \in {5558, 5562}$</td>
<td>$C_i \in {0}$</td>
</tr>
<tr>
<td>BBC</td>
<td>$C_s \in {5390, 5310}$</td>
<td>$C_i \in {2, 4}$</td>
</tr>
<tr>
<td>NBC News</td>
<td>$C_s \in {2772}$</td>
<td>$C_i \in {1, 3}$</td>
</tr>
<tr>
<td>Forbes</td>
<td>$C_s \in {2715, 2720}$</td>
<td>$C_i \in {1}$</td>
</tr>
<tr>
<td>Buzzfeed</td>
<td>$C_s \in {3028}$</td>
<td>$C_i \in {1, 4}$</td>
</tr>
<tr>
<td>Reuters</td>
<td>$C_s \in {6280}$</td>
<td>$C_i \in {2, 4}$</td>
</tr>
<tr>
<td>New York Post</td>
<td>$C_s \in {4563}$</td>
<td>$C_i \in {2, 4}$</td>
</tr>
</tbody>
</table>
Document size extraction

» Predictable patterns to reconstruct transfer size of main document

» Domain specific reconstruction process

» Sequence based
  » Unbroken TLS records of size $D_i \in D$

» Anchor based
  » Anchor records $T_s$ and $T_e$
Document size extraction

» Predictable patterns to reconstruct transfer size of main document

» Domain specific reconstruction process

» Sequence based
  » Unbroken TLS records of size $D_i \in D$

» Anchor based
  » Anchor records $T_s$ and $T_e$

Examples:
New York Times: $D = \{1395, 1055, 202, 40\}$
MSN: $T_s = 33 \quad T_e = 33$
NBC News: $T_s \in \{72, 2907\} \quad T_e \in \{843, ..., 744\}$
Identification: voting group system

\[ C = \prod_{i=2}^{n} \left( 1 - \frac{A_1}{A_1 + A_i} \right) \]
Performance testing

» Single-factor experiments

» Data extraction parameters
  » Pages per domain
  » Time window
  » Score deviation

» Identification parameters
  » Voting group size
  » Confidence threshold
  » Score threshold
Performance testing

» Single-factor experiments

» Data extraction parameters
  » Pages per domain
  » Time window
  » Score deviation

» Identification parameters
  » Voting group size
  » Confidence threshold
  » Score threshold
Example results: pages per domain

- **BBC**:
  - F1: 0.924
  - Recall: 0.966
  - Precision: 0.901

- **Buzzfeed**:
  - F1: 0.827
  - Recall: 0.724
  - Precision: 1

- **Forbes**:
  - F1: 0.977
  - Recall: 0.976
  - Precision: 0.98

Legend:
- **Max value**
- **Min value**
- **Optimistic setting**
- **Conservative setting**
- **Stability around default parameter**
Example results: pages per domain

Attacks performs well
• Only small drops
• High stability

High metrics
• High stability
• Attack scales well
Example results: pages per domain

- Performance starts well
- Quickly drops
- Precision near 1

<table>
<thead>
<tr>
<th>Buzzfeed</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.827</td>
<td>0.724</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forbes</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.977</td>
<td>0.976</td>
<td>0.98</td>
</tr>
</tbody>
</table>

- Poor performance
- No clear TLS record size pattern
- Difficult to extract encrypted sizes
Example results: pages per domain

- In general high results
- Decrease to ~0.5 for all 3 metrics

- For domains where attack worked we see similar but better results
Example results: voting group size

- Tradeoff between F1/recall and precision
- Stability for F1/recall smaller than for precision

- Stable regardless of voting group size
- High stability for all 3 metrics
Example results: voting group size

- Again, poor performance
- Difficult to extract encrypted sizes

- High performance only with small group size

![Graphs showing F1, Recall, and Precision for different voting group sizes. The graphs are divided into three categories: 'New York Post', 'All', and 'Top Performing'. Each category has three subplots for F1, Recall, and Precision, respectively. The x-axis represents the voting group size, and the y-axis represents the performance metrics.](image-url)
Example results: voting group size

- No significant performance gain when increasing group size
- Size near default value 10 performs well
Transfer size analysis

- BBC: Overlap 59.3%
- BuzzFeed: Overlap 57.5%
- Forbes: Overlap 63.5%
- Fox News: Overlap 72.3%
- MSN: Overlap 80.3%
- NBC News: Overlap 63.9%
- New York Post: Overlap 8.7%
- New York Times: Overlap 85.3%
- Reuters: Overlap 84.6%
- Yahoo: Overlap 89.3%
Discussion: example attack

» High correlation between retweets and reads

» Reads at news websites are heavily skewed
  » Top-10 of links account for 37% of reads/retweets
  » Top-50 for 67%
  » Top-100 for 78%

» News cycle typically changes daily
Discussion: example attack

- Conservative results of precision $P_K$ and recall $R_K$ when fingerprinting the top-$K$ news articles

- Recall $R$ on full set of articles observed is same as $R_K$

- $P_{LB} = q_K P_K$
  - $q_K$ is fraction of requests to the top-$K$ articles
  - E.g., for a specific domain:
    - $q_{10} = 0.37$  $q_{50} = 0.67$
Discussion: example attack

» Conservative results of precision $P_K$ and recall $R_K$ when fingerprinting the top-K news articles

» Recall $R$ on full set of articles observed is same as $R_K$

» $P_{LB} = q_K P_K$
  » $q_K$ is fraction of requests to the top-K articles
  » E.g., for a specific domain:
    
    $q_{10} = 0.37$    $q_{50} = 0.67$

» $F_{1LB} = \frac{2R_K q_K P_K}{R_K + q_K P_K}$
Discussion: example attack

- Conservative results of precision $P_K$ and recall $R_K$ when fingerprinting the top-K news articles
- Recall $R$ on full set of articles observed is same as $R_K$

\[
P_{LB} = q_K P_K
\]
- $q_K$ is fraction of requests to the top-K articles
- E.g., for a specific domain:
  \[q_{10} = 0.37 \quad q_{50} = 0.67\]

\[
F_{1LB} = \frac{2R_K q_K P_K}{R_K + q_K P_K}
\]

<table>
<thead>
<tr>
<th>Domain</th>
<th>$K=10$</th>
<th></th>
<th>$K=50$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>$P_{LB}$</td>
<td>$F_{1LB}$</td>
<td>$R$</td>
</tr>
<tr>
<td>BBC</td>
<td>0.97</td>
<td>0.48</td>
<td>0.64</td>
<td>0.83</td>
</tr>
<tr>
<td>Buzzfeed</td>
<td>0.64</td>
<td>0.34</td>
<td>0.44</td>
<td>0.30</td>
</tr>
<tr>
<td>Forbes</td>
<td>0.98</td>
<td>0.38</td>
<td>0.54</td>
<td>0.96</td>
</tr>
<tr>
<td>Fox News</td>
<td>0.96</td>
<td>0.41</td>
<td>0.58</td>
<td>0.60</td>
</tr>
<tr>
<td>MSN</td>
<td>0.39</td>
<td>0.10</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td>NBC News</td>
<td>0.99</td>
<td>0.36</td>
<td>0.52</td>
<td>0.73</td>
</tr>
<tr>
<td>New York Post</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>New York Times</td>
<td>0.99</td>
<td>0.33</td>
<td>0.49</td>
<td>0.89</td>
</tr>
<tr>
<td>Reuters</td>
<td>0.91</td>
<td>0.27</td>
<td>0.42</td>
<td>0.68</td>
</tr>
<tr>
<td>Yahoo</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Discussion: example attack

- F1-score > 0.5 for half of domains even with conservative estimates
Discussion: example attack

» F1-score > 0.5 for half of domains even with conservative estimates

» Top-50 to increase precision

<table>
<thead>
<tr>
<th>Domain</th>
<th>$K=10$</th>
<th>$K=50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>$P_{LB}$</td>
</tr>
<tr>
<td>BBC</td>
<td>0.97</td>
<td>0.48</td>
</tr>
<tr>
<td>Buzzfeed</td>
<td>0.64</td>
<td>0.34</td>
</tr>
<tr>
<td>Forbes</td>
<td>0.98</td>
<td>0.38</td>
</tr>
<tr>
<td>Fox News</td>
<td>0.96</td>
<td>0.41</td>
</tr>
<tr>
<td>MSN</td>
<td>0.39</td>
<td>0.10</td>
</tr>
<tr>
<td>NBC News</td>
<td>0.99</td>
<td>0.36</td>
</tr>
<tr>
<td>New York Post</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>New York Times</td>
<td>0.99</td>
<td>0.33</td>
</tr>
<tr>
<td>Reuters</td>
<td>0.91</td>
<td>0.27</td>
</tr>
<tr>
<td>Yahoo</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Discussion: example attack

» F1-score > 0.5 for half of domains even with conservative estimates

» Top-50 to increase precision

» Top-10 to increase recall
  » Recall > 0.9 for 6 domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>$K=10$</th>
<th></th>
<th></th>
<th>$K=50$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>$P_{LB}$</td>
<td>$F^1_{LB}$</td>
<td>$R$</td>
<td>$P_{LB}$</td>
<td>$F^1_{LB}$</td>
</tr>
<tr>
<td>BBC</td>
<td>0.97</td>
<td>0.48</td>
<td>0.64</td>
<td>0.83</td>
<td>0.61</td>
<td>0.70</td>
</tr>
<tr>
<td>Buzzfeed</td>
<td>0.64</td>
<td>0.34</td>
<td>0.44</td>
<td>0.30</td>
<td>0.72</td>
<td>0.43</td>
</tr>
<tr>
<td>Forbes</td>
<td>0.98</td>
<td>0.38</td>
<td>0.54</td>
<td>0.96</td>
<td>0.63</td>
<td>0.76</td>
</tr>
<tr>
<td>Fox News</td>
<td>0.96</td>
<td>0.41</td>
<td>0.58</td>
<td>0.60</td>
<td>0.49</td>
<td>0.54</td>
</tr>
<tr>
<td>MSN</td>
<td>0.39</td>
<td>0.10</td>
<td>0.15</td>
<td>0.21</td>
<td>0.29</td>
<td>0.24</td>
</tr>
<tr>
<td>NBC News</td>
<td>0.99</td>
<td>0.36</td>
<td>0.52</td>
<td>0.73</td>
<td>0.56</td>
<td>0.63</td>
</tr>
<tr>
<td>New York Post</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>New York Times</td>
<td>0.99</td>
<td>0.33</td>
<td>0.49</td>
<td>0.89</td>
<td>0.51</td>
<td>0.65</td>
</tr>
<tr>
<td>Reuters</td>
<td>0.91</td>
<td>0.27</td>
<td>0.42</td>
<td>0.68</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Yahoo</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.03</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Conclusions

» Design and evaluation of lightweight framework
  » Identify individual browsed news articles (internal pages) despite encryption
  » Separate between articles delivered over same infrastructure (e.g., CDN)

» Demonstrate that naive use of HTTPS is not enough to protect users’ privacy
  » X.509 certificate size (encrypted with TLS 1.3)
  » Web document size

» Provide insights into why websites are more/less resilient to the attack

» Real-world scenario using Twitter

» Provide insights for websites and users to better protect their privacy
Lightweight Fingerprint Attack and Encrypted Traffic Analysis on News Articles

David Hasselquist
Martin Lindblom
Niklas Carlsson

David Hasselquist (david.hasselquist@liu.se)