The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem

IETF 103, Bangkok, 2018

Quirin Scheitle (TUM), Oliver Gasser (TUM), Theodor Nolte (HAW Hamburg), Johanna Amann (ICSI/Corelight/LBNL), Lexi Brent (The University of Sydney), Georg Carle (TUM), Ralph Holz (The University of Sydney), Thomas C. Schmidt (HAW Hamburg), Matthias Währisch (FU Berlin)
Owner of a name gets a certificate
3rd party incorrectly gets a certificate
Name owner cannot verify existence
Certificate Transparency (CT) in a nutshell
Certificate Transparency (CT) in a nutshell

Goal
Provide transparency into issued certificates to detect certificate mis-issuances

Approach
Uses public, append-only logs to record certificates
Additional advantages and new challenges

Does CT introduce new dependencies?
Log servers are operated by multiple companies
Concentration on few log operators should be prevented

What do we lose (or gain) by exposing domain names?
Logs provide data to easily search for names
Might help to identify malicious domain names but might also help attackers to find victims
The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem
How did the log volume change over time?
How did the log volume change over time?

- Large increase of log entries before CT deadline
- Let’s Encrypt dominates
How did the log volume change over time?

- Large increase of log entries before CT deadline
- Let’s Encrypt dominates
- Strong rise
Are CAs distributing certificates over many CT logs?
Are CAs distributing certificates over many CT logs?

- System overly relies on few log servers
- Almost all CAs use few logs for their certificate
Are CAs distributing certificates over many CT logs?

- System overly relies on few log servers
- Almost all CAs use few logs for their certificate
- No, CAs use few logs which limits reliability
The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem
Can CT be used to find malicious domains?
Can CT be used to find malicious domains?

Method
• Inspect domains with similarities to domains from
  • Apple
  • Paypal
  • Hotmail
  • Google
  • Ebay
• Example: appleid.apple.com-7etr6eti.gq
Can CT be used to find malicious domains?

Method

• Inspect domains with similarities to domains from
  • Apple
  • Paypal
  • Hotmail
  • Google
  • Ebay

• Example: appleid.apple.com-7etr6eti.gq

Results 126k potential phishing domains

<table>
<thead>
<tr>
<th>Service</th>
<th>Count</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>63k</td>
<td>appleid.apple.com-7etr6eti.gq</td>
</tr>
<tr>
<td>PayPal</td>
<td>58k</td>
<td>paypal.com-account-security.money</td>
</tr>
<tr>
<td>Microsoft</td>
<td>4k</td>
<td>www-hotmail-login.live</td>
</tr>
<tr>
<td>Google</td>
<td>1k</td>
<td>accounts.google.co.am</td>
</tr>
<tr>
<td>eBay</td>
<td>&lt;1k</td>
<td><a href="http://www.ebay.co.uk.dll7.bid">www.ebay.co.uk.dll7.bid</a></td>
</tr>
</tbody>
</table>
Can CT be used to find malicious domains?

Method

• Inspect domains with similarities to domains from
  • Apple
  • Paypal
  • Hotmail
  • Google
  • Ebay
  • Example: appleid.apple.com-7etr6eti.gq

Results 126k potential phishing domains

<table>
<thead>
<tr>
<th>Service</th>
<th>Count</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>63k</td>
<td>appleid.apple.com-7etr6eti.gq</td>
</tr>
<tr>
<td>PayPal</td>
<td>58k</td>
<td>paypal.com-account-security.money</td>
</tr>
<tr>
<td>Microsoft</td>
<td>4k</td>
<td>www-hotmail-login.live</td>
</tr>
<tr>
<td>Google</td>
<td>1k</td>
<td>accounts.google.co.am</td>
</tr>
<tr>
<td>eBay</td>
<td>&lt;1k</td>
<td><a href="http://www.ebay.co.uk.dll7.bid">www.ebay.co.uk.dll7.bid</a></td>
</tr>
</tbody>
</table>

• CERT confirmed that a subset was used to host malicious content
Can CT be used to find malicious domains?

Method

• Inspect domains with similarities to domains from
  • Apple
  • Paypal
  • Hotmail
  • Google
  • Ebay

• Example: appleid.apple.com-7etr6eti.gq

Results

126k potential phishing domains

<table>
<thead>
<tr>
<th>Service</th>
<th>Count</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>63k</td>
<td>appleid.apple.com-7etr6eti.gq</td>
</tr>
<tr>
<td>PayPal</td>
<td>58k</td>
<td>paypal.com-account-security.money</td>
</tr>
<tr>
<td>Microsoft</td>
<td>4k</td>
<td>www-hotmail-login.live</td>
</tr>
<tr>
<td>Google</td>
<td>1k</td>
<td>accounts.google.co.am</td>
</tr>
<tr>
<td>eBay</td>
<td>&lt;1k</td>
<td><a href="http://www.ebay.co.uk.dll7.bid">www.ebay.co.uk.dll7.bid</a></td>
</tr>
</tbody>
</table>

• CERT confirmed that a subset was used to host malicious content

• **Yes**, CT can be used to find malicious (i.e., phishing) domains
Does CT help attackers to find new domains?
Does CT help attackers to find new domains?

Method

- Extract subdomain labels from all CT logged certificates
  - dev for .io
- Generate new FQDNs with most common subdomain labels
  - dev.foureyes.io
- Ignore .com, .net, .org
- Request A records
Does CT help attackers to find new domains?

Method
• Extract subdomain labels from all CT logged certificates
  • dev for .io
• Generate new FQDNs with most common subdomain labels
  • dev.foureyes.io
• Ignore .com, .net, .org
• Request A records

Results
• 18.8M new FQDNs found
  17.7M more FQDNS compared to other public lists
Does CT help attackers to find new domains?

Method
- Extract subdomain labels from all CT logged certificates
  - dev.for.io
- Generate new FQDNs with most common subdomain labels
  - dev.foureyes.io
- Ignore .com, .net, .org
- Request A records

Results
- 18.8M new FQDNs found
  - 17.7M more FQDNS compared to other public lists
- Yes, CT helps attackers find previously unknown domains
Does CT leak private data to attackers?

Top 20 subdomain labels in CT-logged certificates

<table>
<thead>
<tr>
<th>SDL</th>
<th>Count</th>
<th>SDL</th>
<th>Count</th>
<th>SDL</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>www</td>
<td>61.1M</td>
<td>shop</td>
<td>303k</td>
<td>secure</td>
<td>176k</td>
</tr>
<tr>
<td>mail</td>
<td>14.4M</td>
<td>whm</td>
<td>280k</td>
<td>admin</td>
<td>158k</td>
</tr>
<tr>
<td>webdisk</td>
<td>8.7M</td>
<td>dev</td>
<td>256k</td>
<td>mobile</td>
<td>156k</td>
</tr>
<tr>
<td>webmail</td>
<td>8.6M</td>
<td>remote</td>
<td>253k</td>
<td>server</td>
<td>146k</td>
</tr>
<tr>
<td>cpanel</td>
<td>8.2M</td>
<td>test</td>
<td>249k</td>
<td>cloud</td>
<td>141k</td>
</tr>
<tr>
<td>autodiscover</td>
<td>3.6M</td>
<td>api</td>
<td>239k</td>
<td>smtp</td>
<td>140k</td>
</tr>
<tr>
<td>m</td>
<td>310k</td>
<td>blog</td>
<td>235k</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Is CT actively being misused to find victims?
Is CT actively being misused to find victims?

Method

• Deploy CT honeypot for scanners

• Leak existence of pseudorandom subdomains only via CT logs

• Check DNS logs and check requests on IP addresses belonging to A/AAAA records

• Use EDNS field to reveal locations of stub resolvers
Is CT actively being misused to find victims?

Method
• Deploy CT honeypot for scanners
• Leak existence of pseudorandom subdomains only via CT logs
• Check DNS logs and check requests on IP addresses belonging to A/AAAA records
• Use EDNS field to reveal locations of stub resolvers

Results
• First DNS lookups after 1 minute, HTTP(S) access after 1 hour
• Most scanners without info in rDNS, WHOIS, or on website
Is CT actively being misused to find victims?

Method
• Deploy CT honeypot for scanners
• Leak existence of pseudorandom subdomains only via CT logs
• Check DNS logs and check requests on IP addresses belonging to A/AAAA records
• Use EDNS field to reveal locations of stub resolvers

Results
• First DNS lookups after 1 minute, HTTP(S) access after 1 hour
• Most scanners without info in rDNS, WHOIS, or on website
• One scanner requested A/AAAA records fast and scanned 30 ports
Is CT actively being misused to find victims?

Method
• Deploy CT honeypot for scanners
• Leak existence of pseudorandom subdomains only via CT logs
• Check DNS logs and check requests on IP addresses belonging to A/AAAA records
• Use EDNS field to reveal locations of stub resolvers

Results
• First DNS lookups after 1 minute, HTTP(S) access after 1 hour
• Most scanners without info in rDNS, WHOIS, or on website
• One scanner requested A/AAAA records fast and scanned 30 ports
• Yes, CT is being misused by actors with undeclared intent
Take-Aways

CT ecosystem dominated by few stakeholders
Majority of logging volume from few CAs to few logs

CT helps in finding phishing domains
Enables near-time detection and reaction

CT helps attackers
Find previously unknown domains
Scans from dubious actors within minutes

Data and code available ✓
The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem

Quirin Scheitle¹, Oliver Gasser¹, Theodor Nolte², Johanna Amann³, Lexi Brent⁴, Georg Carle¹, Ralph Holz¹, Thomas C. Schmidt⁵, Matthias Wählsch⁶
¹TUM, ²HAW Hamburg, ³ICSI/Corelight/LBNL, ⁴The University of Sydney, ⁵FU Berlin

ABSTRACT

In this paper, we analyze the evolution of Certificate Transparency (CT) over time and explore the implications of exposing certificate DNS names from the perspective of security and privacy. We find that certificates in CT logs have seen exponential growth. Website support for CT has also constantly increased, with now 33% of established connections supporting CT. With the increasing deployment of CT, there are also concerns of information leakage due to all certificates being visible in CT logs. To understand this threat, we introduce a CT honeypot and show that data from CT logs is being used to identify targets for scanning campaigns only minutes after certificate issuance. We present and evaluate a methodology to learn and validate new subdomains from the vast number of domains extracted from CT logged certificates.

In this paper, we contribute to a better understanding of CT rollout and related security and privacy implications:

- **CA and CT Log Evolution** (§ 2): Using data of all CT log servers deployed, we investigate the evolution of CT logs over time and the dependency of Certificate Authorities (CAs) on CT log operators.
- **Server CT Deployment** (§ 3): Using passive and active measurements, we quantify the evolution of CT adoption among server operators and show positive effects.
- **DNS Information Leakage** (§ 4): We investigate the mass leakage of Fully Qualified Domain Names (FQDNs), and use subdomain data to construct and query new FQDNs.
- **Detecting Phishing Domains** (§ 5): We show that CT logs can be used to detect and study phishing domains.
- **CT Honeypot** (§ 6): We introduce a CT honeypot to show that third parties monitor CT logs to initiate likely malicious scans.

https://doi.org/10.1145/3278532.3278562