Automating Cookie Consent and GDPR Violation Detection

Dino Bollinger, Karel Kubicek, Carlos Cotrini, David Basin
31st USENIX Security Symposium (August 11, 2022)
Cookie consent

- Solomos et al. (2019): 90% of websites use tracking cookies
- EU law: Websites must notify users, gather consent
- Consent notices to comply with regulations
ePrivacy Directive and General Data Protection Regulation (GDPR)

**ePrivacy Directive:**

- All but strictly necessary data processing requires consent

**GDPR Consent:**

- Freely-given
- Unambiguous
- Specific
- Informed
- Purpose-limited

This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

[Cookie declaration](#)
Non-compliance is widespread

• Empirical studies:
  – Non-compliance in up to 80% of websites (e.g. cookies set before consent)
    (Utz 2019, Trevisan 2019, Matte 2020, Nouwens 2020, Kampanos 2021, Santos 2021, etc.)
  – Websites do not respect user choices
    (Libert 2018, Trevisan 2019, Matte 2020, Nouwens 2020, etc.)

• Usability: dark patterns successfully trick users
  (Bösch 2016, Grassl 2020, Hasner 2021, Sanchez-Rola 2019, Htut Soe 2020, etc.)
Goal: Enforce cookie consent on client-side while browsing.
Our solution: CookieBlock

• Browser extension to predict purposes for cookies using machine learning
• Enforce cookie consent on client-side

Implementation:

1. Crawl web to gather training data (ground truth)
2. Extract features from cookies
3. Train classifier model and evaluate
4. Apply the model in the browser extension
Data collection: selecting data sources

- Source for training data:

<table>
<thead>
<tr>
<th>Cookie declaration</th>
<th>About</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Provider</td>
</tr>
<tr>
<td>_ga</td>
<td>Google</td>
</tr>
<tr>
<td>_gat</td>
<td>Google</td>
</tr>
<tr>
<td>__qca</td>
<td>Quantcast</td>
</tr>
</tbody>
</table>

- Consent Management Platform (CMP):
Data collection: web crawlers

- CMP presence crawler:
  - Input: 6 million domains, sourced from Tranco list
  - Output: ~37.5k domains with confirmed CMP

- Cookie consent crawler:
  - Browse websites, gather cookie declarations + observed cookies
  - Based on OpenWPM – visit subpages, move cursor, etc.
  - Successful for ~30k domains
Data collection: results of OpenWPM crawl

- Declared: 2.2M
- Observed:
  - Matched: 323k
  - 602k

Bar chart:
- Necessary: 286,113 (30.0%)
- Functional: 186,863 (10.1%)
- Analytics: 247,807 (35.6%)
- Advertising: 1,319,868 (8.5%)
Feature extraction from textual cookies

Ex 1: Shannon entropy
- Higher entropy, more randomness
- Indicator for unique identifiers

\[ H(X) = - \sum_{i=1}^{n} p_i \log_2 p_i \]

Ex 2: Content encodings
- JSON, CSV, Base64, etc.

52 types in total, including:
- Name patterns, content size, timestamps, language strings, content encoding, cookie flags, third-party status, expiry, etc.
XGBoost classifier and baseline

- XGBoost used for training our model
- Cookiepedia
  - Repository of cookies labeled by experts
  - Same cookie purposes, 70% of our dataset

![Largest Database of Pre-Categorized Cookies](image)

XGBoost
Prediction: Advertising

Session?
- True
- False

entropy > 0.8
- 3
- 1
+ 3
Classifier evaluation

Cookiepedia bal. accuracy
84.7% ± 0.3%

XGBoost bal. accuracy
84.2% ± 0.27%
CookieBlock browser extension

- User defines consent preferences when installed
- Classifies cookies and deletes those with rejected purpose
- Available for Firefox, Chrome, Edge and Opera (8k users)

Empirical evaluation:
- No broken functionality on 85 out of 100 pages
- Authentication issues on 7 websites
- 7 broken consent popups, 1 language setting problem
Potential violations: per type

- **Wrong Purpose (GA)**: 2402 websites (8.2%)
- **Outlier from Majority**: 9094 websites (30.9%)
- **Multiple Purposes**: 674 websites (2.3%)
- **Unclassified**: 7460 websites (25.4%)
- **Undeclared**: 24248 websites (82.5%)
- **Incorrect Expiry**: 3983 websites (13.5%)
- **Implicit Consent**: 20498 websites (69.7%)
- **Ignored Choices**: 6274 websites (21.3%)
Potential violations: wrong purpose and undeclared cookies

"Google Analytics" cookie with wrong purpose
- Detected on 8.2% of all websites
- 2.7% misclassified as necessary

Decision from Planet49 case

Undeclared cookies
- Cookies that were not listed in the consent
- 82.5% of websites
- 40.2% of cookies in total not declared

GDPR informed consent requirement

The Twemoji cookie image is licensed under the Creative Commons Attribution 4.0 International license, and has been altered from its original form. We claim no ownership of the image.
Potential violations: implicit and ignored consent

Cookies set prior to user’s consent

- Found on 69.7% of all domains
  - Nouwens et al. (32.5%)
  - Matte et al. (9.9%)

Article 5(3) of the ePrivacy Directive

Cookies set despite negative consent

- Found on 21.3% of all domains
  - Differs from Matte et al. (5.3%)

Article 5(3) of the ePrivacy Directive

Potential violations: histogram

Number of websites

Number of violation types

- 0: 1548 (5.3%)
- 1: 5123 (17.4%)
- 2: 8166 (27.8%)
- 3: 7336 (25.0%)
- 4: 5320 (18.1%)
- 5: 1550 (5.3%)
- 6: 345 (1.2%)
- 7: 10 (0.01%)
Conclusion

- Consent notices are broken
- Crawled ground truth + extracted features
- Trained XGBoost model to predict purposes for cookies
- CookieBlock enforces user consent preferences
- Detected 8 potential violation types on ~95% of websites

● Thank you for your attention!

Questions?

Dino Bollinger, Karel Kubicek, Dr. Carlos Cotrini, Prof. David Basin

More info, source, extension links:
https://karelkubicek.github.io/post/cookieblock
Backup Slides
## Consent management platforms: market share & analysis

<table>
<thead>
<tr>
<th>CMP</th>
<th>Market share</th>
<th>Remote</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osano</td>
<td>2.25%</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Cookie Notice</td>
<td>1.29%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>OneTrust</td>
<td>1.17%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OptAnon</td>
<td>1.08%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cookie Law Info</td>
<td>0.95%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Cookiebot</td>
<td>0.77%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quantcast CMP</td>
<td>0.68%</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>UK Cookie Consent</td>
<td>0.33%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TrustArc</td>
<td>0.26%</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>WP GDPR Comp.</td>
<td>0.20%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Moove GDPR Comp.</td>
<td>0.18%</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CMP</th>
<th>Market share</th>
<th>Remote</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>tarteaucitron.js</td>
<td>0.16%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Usercentrics</td>
<td>0.16%</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>CookiePro</td>
<td>0.15%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Borlabs Cookie</td>
<td>0.12%</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>EU Cookie Law</td>
<td>0.12%</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>PrimeBox CookieBar</td>
<td>0.09%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Cookie Script</td>
<td>0.07%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cookie Information</td>
<td>0.06%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Termly</td>
<td>0.05%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cookie Info Script</td>
<td>0.05%</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Easy GDPR</td>
<td>0.04%</td>
<td>✓</td>
<td>X</td>
</tr>
</tbody>
</table>
Feature importance

How many times is the feature used
High weight $\rightarrow$ feature used close to the leaf

How many cookies were influenced by the feature
High importance $\rightarrow$ feature close to the root
# Model precision and recall

<table>
<thead>
<tr>
<th></th>
<th>Strictly necessary</th>
<th>Functionality</th>
<th>Performance/analytics</th>
<th>Tracking/advertising</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>88.5%</td>
<td>78.7%</td>
<td>93.0%</td>
<td>79.0%</td>
</tr>
<tr>
<td>Cookiepedia</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td></td>
<td>81.7%</td>
<td>87.3%</td>
<td>89.7%</td>
<td>89.8%</td>
</tr>
<tr>
<td>XGBoost</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>94.5%</td>
<td>38.1%</td>
<td>84.2%</td>
<td>94.9%</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td></td>
<td>Recall</td>
<td></td>
</tr>
<tr>
<td></td>
<td>87.3%</td>
<td>52.9%</td>
<td>89.8%</td>
<td>93.6%</td>
</tr>
</tbody>
</table>

**Cookiepedia accuracy** 86.1% ± 0.1%  
**XGBoost accuracy** 87.2% ± 0.23%
CookieBlock: manual evaluation

- 7 websites with authentication issues
  - authorstream.com
  - walmart.com
  - dafont.com
  - tpsl-india.in
  - sage.com
  - eventbrite.co.uk
  - formstack.com

- 8 websites with functional issues
  - tandf.co.uk: can't change region
  - sherdog.com: CMP issues
  - martindale.com: CMP issues
  -taboola.com: CMP issues
  - thegatewaypundit.com: CMP issues
  - philips.com: CMP issues
  - windowsupdate.com: CMP issues
  - prweb.com: CMP issues
Related work – extensions

- Rachel's GDPR Consent Manager
- CookieEnforcer
- Consent-O-Matic
- I don’t care about cookies
- uBlock Origin with Easylist cookies
Related work – comparable approaches

- CCCC: Corralling Cookies into Categories with CookieMonster [Hu et al., 2021]
  
  ![](image1)

  **Figure 2:** Confusion matrix of Multinomial Naive Bayes (MNB). Majority of the misclassification happened due to Targeting/Advertising cookies.

  **Figure 3:** Recall, Precision and F-score of for different classification models to categorise cookies. MNB and MLP achieved more than 94% average F1-score.

- CookieEnforcer: Automated Cookie Notice Analysis and Enforcement [Khandelwal et al., 2022]
  
  - Requires honest implementation of consent
  - Works only for English (EN) websites
  - ML not in browser, but in crawler - limited to crawled websites with "recipes"
Violation detection: outliers, conflicting purposes

Outlier purpose from majority opinion

- Majority label for third-party cookie, find declarations that do not match
- Outliers found on 30.9% of all websites

Lower bound, indicates misbehavior

Conflicting purposes

- 2+ purposes for same cookie
- Cookie is enabled by consent to at least one purpose
- Found on 2.3% of all websites
  - 0.7% of all sites use “Necessary” and another class

Non-ambiguous requirement of GDPR
Violation detection: unclassified cookies, incorrect expiry

**Unclassified cookies**
- Unclassified in the declaration on 25.4% of all websites
- ~4% of declarations were unclassified
- Cannot be rejected in Cookiebot notice

Informed consent requirement of GDPR

**Incorrect expiry**
- Cookie expiry is 1.5 times longer than declared
- Also includes cases of session cookies being persistent
- On 13.5% of all domains

Violation in Planet49 case
Violation statistics, repeated results after 1 year

May 2021 crawl (29’206 websites)
Cookiebot: 45.8%, OneTrust: 52.1%, Termly: 2.2%

- Wrong Purpose (GA): 8.2% (2402)
- Outlier from Majority: 30.9% (9094)
- Multiple Purposes: 2.3% (674)
- Unclassified: 25.4% (7460)
- Undeclared: 82.5% (24248)
- Incorrect Expiry: 13.5% (3983)
- Implicit Consent: 69.7% (20498)
- Ignored Choices: 21.3% (6274)

July 2022 crawl (52’162 websites)
Cookiebot: 57.9%, OneTrust: 39.6%, Termly: 2.6%

- Wrong Purpose (GA): 6.3% (3293)
- Outlier from Majority: 35.9% (18731)
- Multiple Purposes: 3.4% (1751)
- Unclassified: 24.9% (13014)
- Undeclared: 84.2% (43929)
- Incorrect Expiry: 11.8% (6142)
- Implicit Consent: Evaluation not performed for cost reasons
- Ignored Choices: Evaluation not performed for cost reasons
Violation statistics, grouped by CMP

May 2021 crawl

- Wrong Purpose (GA): 8.2% (All CMPS), 3.3% (Cookiebot), 12.9% (Onetrust), 0.3% (Termly)
- Outlier from Majority: 31.1% (All CMPS), 24.1% (Cookiebot), 36.4% (Onetrust), 54.2% (Termly)
- Multiple Purposes: 2.3% (All CMPS), 0.0% (Cookiebot), 4.4% (Onetrust), 0.0% (Termly)
- Unclassified: 5.5% (All CMPS), 25.5% (Cookiebot), 51.9% (Onetrust), 79.5% (Termly)
- Undeclared: 13.6% (All CMPS), 85.9% (Cookiebot), 88.7% (Onetrust), 79.5% (Termly)
- Incorrect Expiry: 5.8% (All CMPS), 20.9% (Cookiebot), 3.5% (Onetrust), 2.7% (Termly)

July 2022 crawl

- Wrong Purpose (GA): 6.3% (All CMPS), 2.7% (Cookiebot), 12.0% (Onetrust), 1.0% (Termly)
- Outlier from Majority: 35.9% (All CMPS), 36.0% (Cookiebot), 35.8% (Onetrust), 36.6% (Termly)
- Multiple Purposes: 3.4% (All CMPS), 0.0% (Cookiebot), 8.5% (Onetrust), 0.0% (Termly)
- Unclassified: 0.3% (All CMPS), 24.9% (Cookiebot), 39.7% (Onetrust), 72.0% (Termly)
- Undeclared: 11.8% (All CMPS), 88.6% (Cookiebot), 82.2% (Onetrust), 81.3% (Termly)
- Incorrect Expiry: 5.4% (All CMPS), 21.7% (Cookiebot), 3.7% (Onetrust), 2.7% (Termly)
Presentation Authors:
Dino Bollinger, dino.bollinger@gmail.com
Karel Kubicek, karel.kubicek@inf.ethz.ch

Team:
Dino Bollinger, Karel Kubicek, Dr. Carlos Cotrini, Prof. David Basin

ETH Zurich
D-INFK
Institute of Information Security
https://informationsecurity.ethz.ch/