Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission

Asuman Senol
imec-COSIC, KU Leuven
asuman.senol@esat.kuleuven.be

Gunes Acar
Radboud University
g.acar@cs.ru.nl

Mathias Humbert
University of Lausanne
mathias.humbert@unil.ch

Frederik Zuiderveen Borgesius
Radboud University
frederikzb@cs.ru.nl
Background

- Websites use advertising and marketing for monetization
  - built-in anti-tracking countermeasures
  - potential third-party cookie phase-out
- Tracking by email addresses
  - persistent, cross-site, cross-platform
Motivation

- PII collection before form submission on a mortgage calculator website (Gizmodo, 2017)
- A 2018 survey (n=502):
  - 81% abandoned forms at least once
  - 59% abandoned a form in the last month
Study Objectives

• Measure email and password collection prior to form submission
  • effect of location: EU vs. US
  • effect of consent
  • mobile vs. desktop
Method – Web Crawler

- Built on Tracker Radar Collector (developed by DuckDuckGo)
## Crawls (May, June 2021)

<table>
<thead>
<tr>
<th>Crawl Option</th>
<th>EU</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no-action</td>
<td>accept-all</td>
</tr>
<tr>
<td>Crawled URLs</td>
<td>100K</td>
<td>7,720</td>
</tr>
<tr>
<td>Visited websites</td>
<td>99,380</td>
<td>7,716</td>
</tr>
<tr>
<td>Crawled pages</td>
<td>625,143</td>
<td>44,752</td>
</tr>
</tbody>
</table>

> 2.8 million pages
## Results

<table>
<thead>
<tr>
<th></th>
<th>Email</th>
<th>Password</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EU</td>
<td>US</td>
</tr>
<tr>
<td>Websites where we filled</td>
<td>52,055</td>
<td>53,038</td>
</tr>
<tr>
<td>Leaks to 1st party</td>
<td>4,395</td>
<td>5,518</td>
</tr>
<tr>
<td>Leaks to 3rd party</td>
<td>2,633</td>
<td>3,790</td>
</tr>
<tr>
<td>Leaks to trackers</td>
<td><strong>1,844</strong></td>
<td><strong>2,950</strong></td>
</tr>
</tbody>
</table>

**How did we label them?**

- By using blocklists such as Easylist, EasyPrivacy
- + Manual labeling: Discovered **41** unlisted tracker domains
# Prominent Tracker Domains

<table>
<thead>
<tr>
<th>EU</th>
<th></th>
<th>US</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Domain</strong></td>
<td><strong>Num Sites</strong></td>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Taboola</td>
<td>taboola.com</td>
<td>327</td>
<td>LiveRamp</td>
</tr>
<tr>
<td>FullStory</td>
<td>fullstory.com</td>
<td>182</td>
<td>Taboola</td>
</tr>
<tr>
<td>Adobe</td>
<td>bizible.com</td>
<td>160</td>
<td>Adobe</td>
</tr>
<tr>
<td>Yandex</td>
<td>yandex.com</td>
<td>121</td>
<td>BounceX</td>
</tr>
<tr>
<td>Awin</td>
<td>awin1.com</td>
<td>113</td>
<td>Awin</td>
</tr>
<tr>
<td></td>
<td>zenaps.com</td>
<td>112</td>
<td></td>
</tr>
</tbody>
</table>
Lookalike Targeting

Look-alike models are used to increase scale by finding new people likely to be interested in your business because they resemble existing customers.

Use your CRM data to create lookalike audiences on Taboola! You can upload either a customer list of hashed email addresses, mobile device IDs, or 5 digit US zip codes and Taboola's predictive engine will find similar users that are more likely to convert based on the assumption that these users will be “like” your current customers in your database.

Visit our Advertiser Data Use Policy here.
If you selected a method that includes On-page detection, use the Start Detecting Identifier on dropdown to choose the listener event type for when ATS needs to actually detect the identifier on the website:

- **Click event**: Click event will fire off whenever a specified element is clicked (enter these elements in the Trigger Elements field).

- **Submit Event**: Submit event will fire off whenever a specified form is submitted (enter these elements in the Trigger Elements field).

**Note**
Trigger Elements are CSS selectors to define elements on which the event will be triggered. For examples: #button-id-click or #form-id. As shown in the examples, the given value should start with a hash #. In order to configure Trigger Elements it is recommended to add a CSS ID of HTML elements to your forms.

- **Blur Event**: Blur event will fire off whenever a specified input field loses focus for example when a user clicks outside of the input field.

**Warning**
The 'Blur Event' method doesn't require human interaction for identifiers to be obtained, while other methods require users to click on a button such as "Submit" or "Ok". To your users, this may give the perception that malicious activities are happening in the background, which is not the case because ATS.js will only start detection with proper consent in place.

Blur Event detection also leaves room for incorrect identifiers because it will not wait for actions from the user like clicking on a login button. For these reasons, we recommend using On Click or On Submit method instead.
## Top ten websites

<table>
<thead>
<tr>
<th>Rank</th>
<th>Website</th>
<th>3rd-party</th>
<th>Rank</th>
<th>Website</th>
<th>3rd-party</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>usatoday.com*</td>
<td>taboola.com</td>
<td>95</td>
<td>issue.com</td>
<td>taboola.com</td>
</tr>
<tr>
<td>242</td>
<td>trello.com*</td>
<td>bizible.com</td>
<td>128</td>
<td>businessinsider.com</td>
<td>taboola.com</td>
</tr>
<tr>
<td>243</td>
<td>independent.co.uk*</td>
<td>taboola.com</td>
<td>154</td>
<td>usatoday.com</td>
<td>taboola.com</td>
</tr>
<tr>
<td>300</td>
<td>shopify.com</td>
<td>bizible.com</td>
<td>191</td>
<td>time.com</td>
<td>bouncex.net</td>
</tr>
<tr>
<td>328</td>
<td>marriott.com</td>
<td>glassboxdigital.io</td>
<td>196</td>
<td>udeemy.com</td>
<td>awin1.com</td>
</tr>
<tr>
<td>567</td>
<td>newsweek.com*</td>
<td>rlcdn.com</td>
<td></td>
<td>udeemy.com</td>
<td>zenaps.com</td>
</tr>
<tr>
<td>705</td>
<td>prezi.com</td>
<td>taboola.com</td>
<td>217</td>
<td>healthline.com</td>
<td>rlcdn.com</td>
</tr>
<tr>
<td>754</td>
<td>branch.io*</td>
<td>bizible.com</td>
<td>34</td>
<td>foxnews.com</td>
<td>rlcdn.com</td>
</tr>
<tr>
<td>1,153</td>
<td>prothomalo.com</td>
<td>facebook.com</td>
<td>242</td>
<td>trello.com*</td>
<td>bizible.com</td>
</tr>
<tr>
<td>1,311</td>
<td>codecademy.com</td>
<td>fullstory.com</td>
<td>278</td>
<td>theverge.com</td>
<td>rlcdn.com</td>
</tr>
<tr>
<td>1,543</td>
<td>azcentral.com*</td>
<td>taboola.com</td>
<td>288</td>
<td>webmd.com</td>
<td>rlcdn.com</td>
</tr>
</tbody>
</table>

*: Not reproducible anymore as of February 2022.
## Website Categories

<table>
<thead>
<tr>
<th>Categories</th>
<th>EU/US</th>
<th>EU</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sites</td>
<td>Filled sites</td>
<td>Leaky sites</td>
</tr>
<tr>
<td>Fashion/Beauty</td>
<td>1,669</td>
<td>1,176</td>
<td>131 (11.1%)</td>
</tr>
<tr>
<td>Online Shopping</td>
<td>5,395</td>
<td>3,658</td>
<td>345 (9.%)</td>
</tr>
<tr>
<td>General News</td>
<td>7,390</td>
<td>3,579</td>
<td>235 (6.6%)</td>
</tr>
<tr>
<td>Software/Hardware</td>
<td>4,933</td>
<td>2,834</td>
<td>138 (4.9%)</td>
</tr>
<tr>
<td>Business</td>
<td>13,462</td>
<td>7,805</td>
<td>377 (4.8%)</td>
</tr>
<tr>
<td>Gov’t/Military</td>
<td>3,754</td>
<td>939</td>
<td>3 (0.5%)</td>
</tr>
<tr>
<td>Pornography</td>
<td>1,388</td>
<td>528</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>
# EU vs US

<table>
<thead>
<tr>
<th></th>
<th>EU</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visited websites</td>
<td>99,380</td>
<td>99,437</td>
</tr>
<tr>
<td>Websites where we filled</td>
<td>52,055</td>
<td>53,038</td>
</tr>
<tr>
<td>Emails sent to 1st party</td>
<td>4,395</td>
<td>5,518</td>
</tr>
<tr>
<td>Emails sent to 3rd party</td>
<td>2,633</td>
<td>3,790</td>
</tr>
<tr>
<td>Emails sent to trackers</td>
<td>1,844</td>
<td>2,950</td>
</tr>
</tbody>
</table>

60% difference

addthis.com, yahoo.com, doubleclick.net and criteo.com

Only appear in the US crawl
Received Emails

- 290 emails from 88 distinct sites
  - Offer a discount, or
  - Invite us back to their site

Email from: diabetes.org.uk
Tracker domain: freshaddress.biz
Received Emails

 Searching for products that actually work?  

Email from: mypillow.com
Tracker domain: listrakbi.com

Thanks For Stopping By

When I started MyPillow, my passion was to help people get the best sleep of their life! What a blessing it has been to see that dream become a reality!

To help you best care for your MyPillow, please read our product care recommendations. If you have any questions, please don’t hesitate to contact us!

¡Se despide Hot Days! 18 MSI + BONIFICACIÓN

Email from: walmart.com.mx
Tracker domain: veinteractive.com
Password Leaks

- Incidental collection on 52 sites by
  - Yandex Metrica: due to React framework (50 websites)
  - Mixpanel: due to outdated SDK usage (1 website)
  - LogRocket: No response (1 website)
- Fixed thanks to our disclosures
МОЯ ТОЙОТА

Войти в личный кабинет

Электронная почта/Имя пользователя

Пароль

mypassword12345678

У меня нет учетной записи
Восстановить пароль
Outreach Efforts

First parties: 30/58 replied
- Were not aware & removed
  - fivethirtyeight.com (via Walt Disney’s DPO)
  - trello.com (Atlassian)
- Marriott: Glassbox is used for **customer care, technical support, and fraud prevention**

Third parties: 15/28 replied
- Adobe and Yandex: Referred to corresponding first parties
- Taboola: ad & content personalization, CMP misconfiguration

0/33 first parties replied (Websites in the US crawl)
- No response from these 33 websites.
Leaks to Facebook & TikTok

- Closer look to Facebook
- Due to Automatic Advanced Matching feature of Facebook/Tiktok Pixel
  (scrapes personal information from forms)
LeakyForms - USENIX Security '22

New to 9GAG?
Sign up now to see more content!

Log in
- Continue with Facebook
- Continue with Google
- Continue with Apple

Forgot password?

www.facebook.com/tv
How it Works

With automatic advanced matching, we can capture the hashed customer data (ex: email addresses) you collect from your website during processes like checkout, account sign-in, or registration. Hashing is the process we use to transform data for security reasons. We can then use hashed identifiers to better match people visiting your website with people on Facebook, which can lead to more attributed conversions for your Facebook campaigns and a larger size of your custom audiences.

1. A visitor fills out a form on your website, such as during checkout, account sign-in, or registration.

2. After the visitor hits the Pixel's Submit JavaScript code automatically detects and passes the relevant form fields to Facebook. Sensitive data, such as passwords or financial data, is never shared with Facebook.

3. The form field data is hashed in the visitor's browser before it is sent to Facebook.

4. Facebook takes the form data and the action that was taken (for example, a purchase), and matches it to a Facebook user.
LeakyForms - USENIX Security '22

Leaks to Facebook & TikTok

- Triggered when the user clicks any link or button on the page

<table>
<thead>
<tr>
<th></th>
<th>EU</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>7,379</td>
<td>8,438</td>
</tr>
<tr>
<td>TikTok</td>
<td>147</td>
<td>154</td>
</tr>
</tbody>
</table>
Countermeasures

- Adblockers that block requests to tracker domains
- Private email relay services that hide users’ emails
  - Apple, Mozilla, DuckDuckGo
  - e.g. testuser@duck.com-> testuser@gmail.com
- NO tool for detection and prevention of sniff & exfiltration on online forms
**LEAKINSPECTOR**

- Proof-of-concept browser add-on
  - [https://github.com/leaky-forms/leak-inspector](https://github.com/leaky-forms/leak-inspector)
- Detects sniff attempts
- Blocks leaky requests
LEAKINSPECTOR
Summary

- Email leaks on 1,844 (EU), 2,950 (US) websites
- Password leaks on 52 websites due to session replay scripts
- Uncovered 41 unlisted tracking domains
- Developed a transparency browser add-on that detects and blocks personal data exfiltration from online forms
Any Questions?

- Source code: [https://github.com/leaky-forms/leaky-forms](https://github.com/leaky-forms/leaky-forms)
- Press coverage: Wired, Fortune, Le Monde, Die Zeit