Research Track Responsible Web

Automating Website Registration for Studying GDPR Compliance

Karel Kubicek, Jakob Merane, Ahmed Bouhoula, and David Basin (ETH Zurich)

1 Motivation

Privacy (compliance) research on websites is limited to index pages or randomly visited pages [1]. Many websites require authentication, yet privacy in such scenarios is understudied, and analyses of non-authenticated pages fail to capture real users' behavior. We developed a crawler automating website registration and demonstrated its usefulness by studying the GDPR and ePrivacy Directive compliance of the registration process and subsequent sending of marketing emails.

2 Crawler

Our crawler automates website registration and newsletter subscription with **>20% success rate** for websites with such forms, significantly outperforming the state of the art [2]. We achieved that by multiple improvements: our crawler supports 37 European languages, it employs a multitude of bot-evasion techniques, and it can navigate complex multi-stage forms.

Every registration is performed with a unique email address, and accounts are activated automatically (*double-opt-in* mail).

We crawled 660'202 websites and received emails from 33'899 of them.

Fig. 2 summarizes security violations detected by our crawler. The GDPR requires websites to follow best security practices, such as securing registrations using TLS and not sending the user-provided password via insecure email.

3 ML detection of potential privacy violations

In the EU, websites are required to collect user consent before sending marketing emails. In our prior work [3], we translated the legal requirements into a decision tree depicted in Fig 1. This decision tree takes as input legal properties of the received emails and the website form, such as whether the form contains a marketing checkbox.

We use 1k forms and 5k emails annotated in [3] to train ML models predicting these properties. As features, we use Universal Sentence Encoder embeddings and further numerical and categorical properties crafted with domain knowledge. The best-performing models were XGBoost.

Fig 3. summarizes the form consent violations, and Fig. 5 presents classification of the first email, which should collect email recipient's consent (double-opt-in procedure).

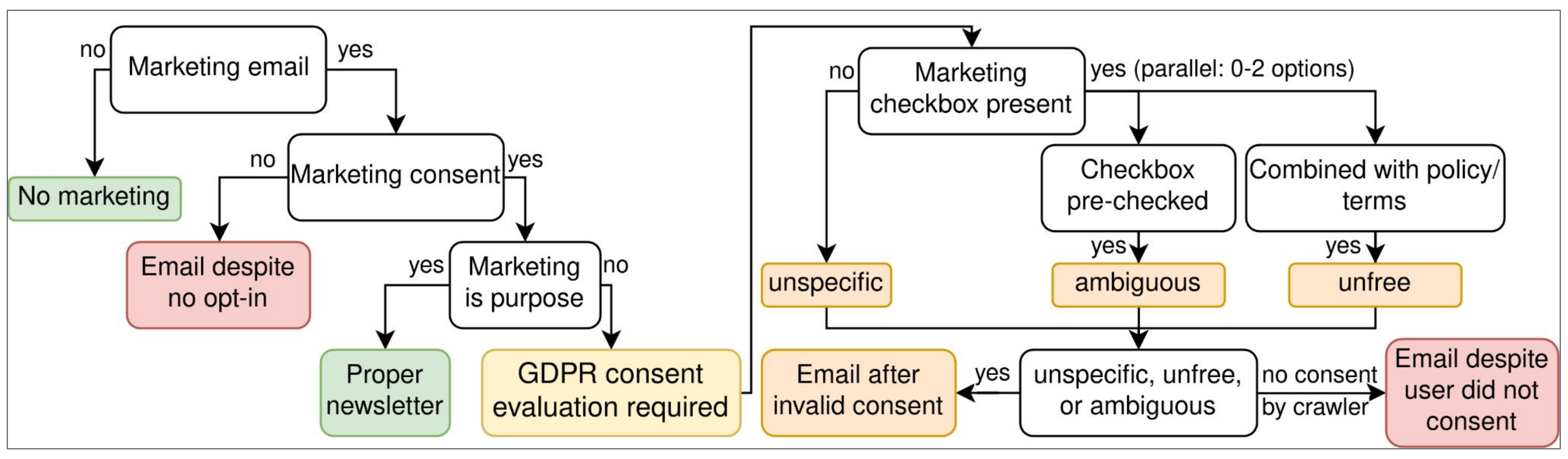
no
Marketing email
yes
No marketing
no
Marketing consent
yes
Email despite no opt-in
yes
Marketing is purpose
no
Proper newsletter
GDPR consent evaluation required
no
Marketing checkbox present
yes (parallel: 0-2 options)
Checkbox pre-checked
Combined with policy/ terms
yes
yes
unspecific
ambiguous
unfree
Email after invalid consent
yes
unspecific, unfree, or ambiguous
no consent by crawler
Email despite user did not consent

Fig. 1: Decision tree predicting potential violations of marketing consent in registration and newsletter subscription forms.

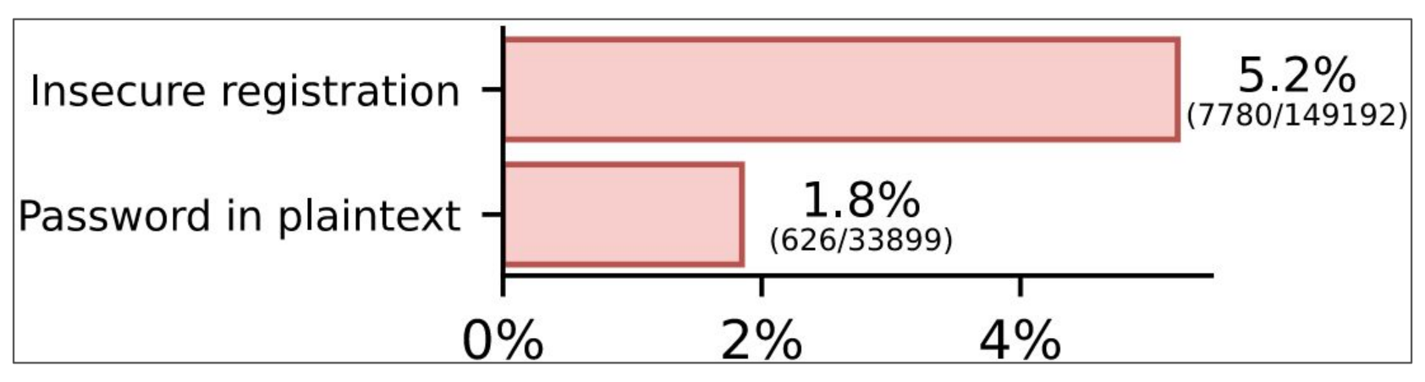

Fig. 2: Security threats of website registration.

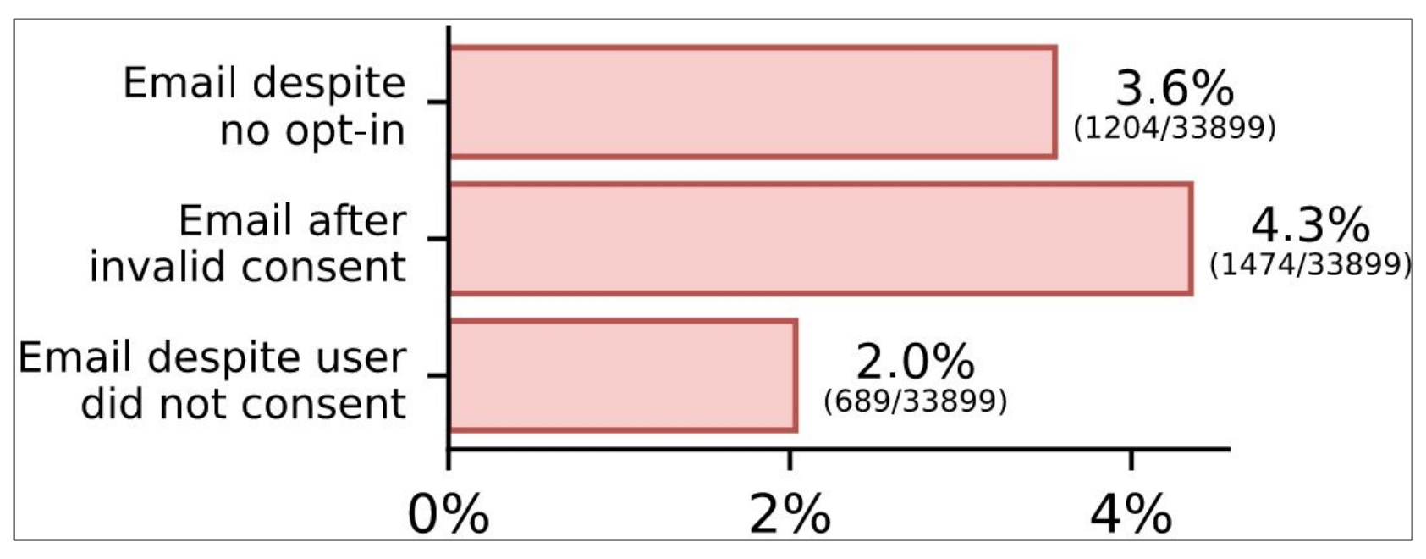

Fig. 3: Consent violations according to Fig. 1.

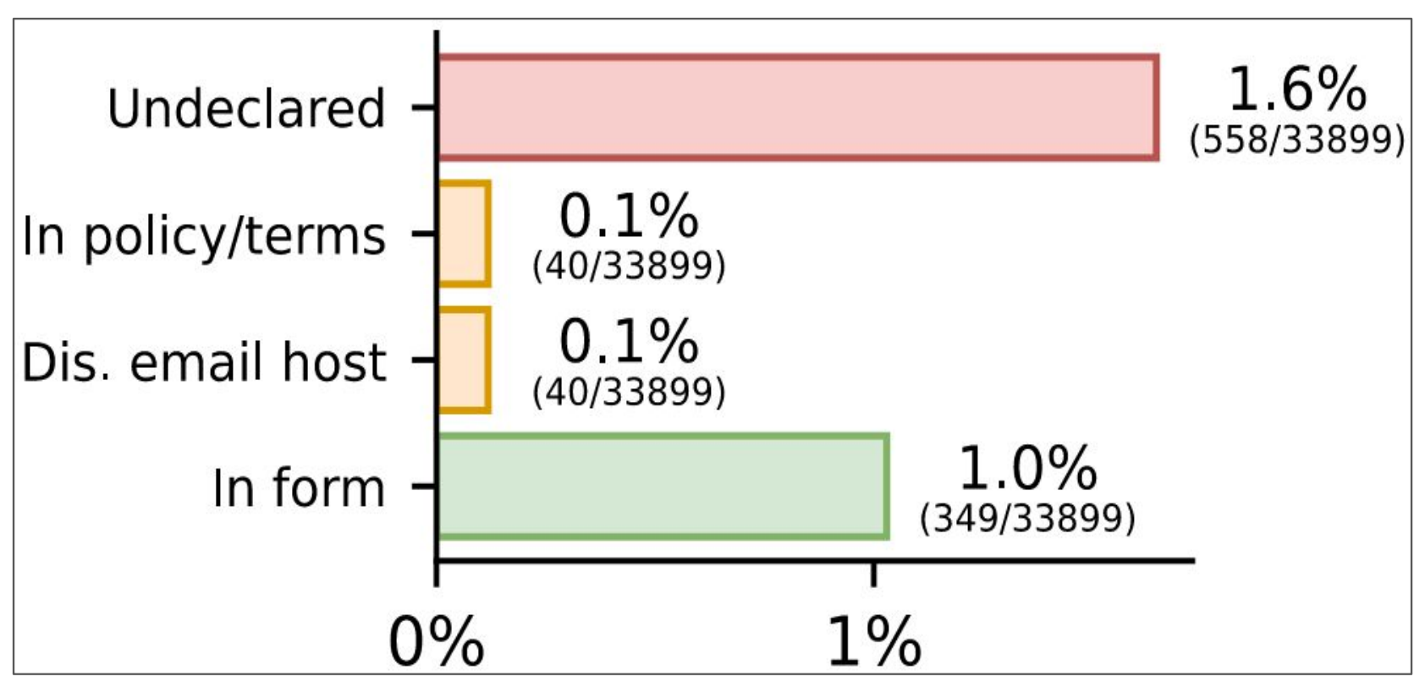

Fig. 4: Email sharing classification.

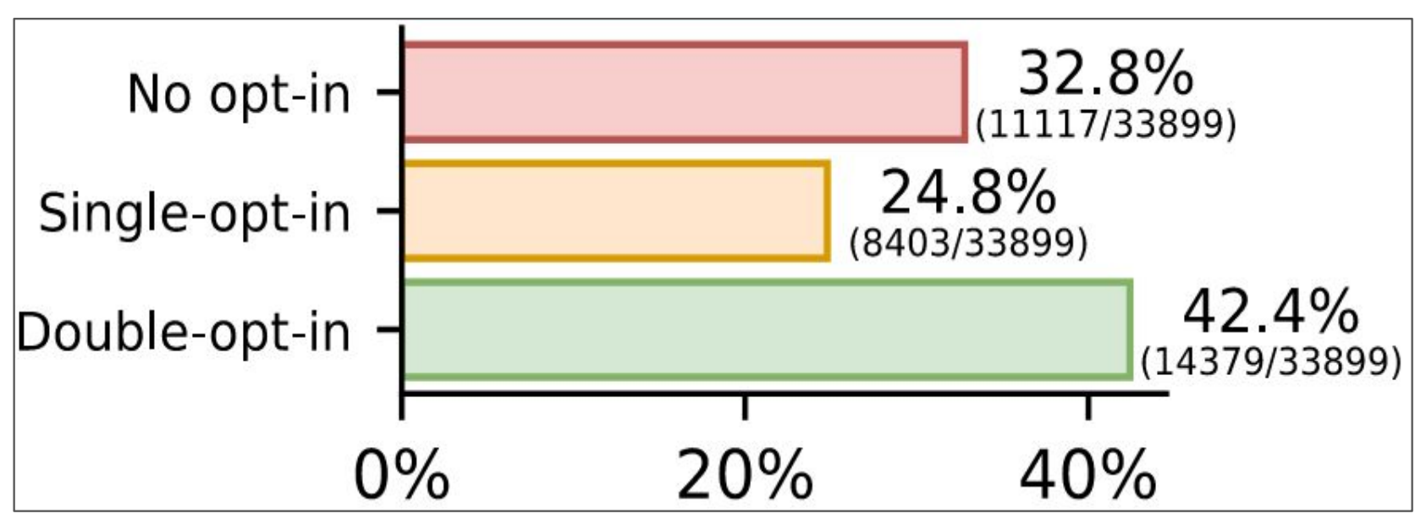

Fig. 5: Double-opt-in requirement violations.

4 Disclosure of email sharing

Each registration was performed with a unique email address. This allows us to inspect whether the website shares our email address to undisclosed third parties. We designed a heuristic that classifies third-party senders according to disclosure of their domain within registration. Fig. 4 summarizes this classification. Overall, the incidence of <2% is significantly lower than in the US election campaigns [4]. This suggests that EU regulations foster privacy protection.

5 Conclusion and future work

Overall, **37.2% of senders potentially violated consent requirements**. Further research should inspect incidences in other jurisdictions to determine the real impact of existing laws. In collaboration with legal researchers, we are working on an in-depth legal analysis of marketing emails violations depending on the website's popularity, category, and location.

We plan to explore the application of our infrastructure for regulatory enforcement. However, our violation detection is prone to too many false positives, which makes its application unsuitable for enforcement agencies. Extending the training data and use of advanced ML can help address those.

Our crawler fosters various kinds of research, such as inspecting security and privacy of of registration pages or pages requiring prior authentication. Studies of marketing emails can be more representative when inspecting a large dataset of emails. Finally, we intend to analyze the process of unsubscribing from the newsletter from both legal and security perspectives.

For more details, visit the publication page, QR code above or https://karelkubicek.github.io/post/reg-www.

References

1. Urban, Tobias, et al. "Beyond the front page: Measuring third party dynamics in the field." *WWW 2020.* https://doi.org/10.1145/3366423.3380203
2. Drakonakis, Kostas, et al. "The cookie hunter: Automated black-box auditing for web authentication and authorization flaws." *CCS 2020.* https://doi.org/10.1145/3372297.3417869
3. Kubicek, Karel, et al. "Checking Websites' GDPR Consent Compliance for Marketing Emails." Proceedings on Privacy Enhancing Technologies 2022.2 (2022): 282-303. https://doi.org/10.2478/popets-2022-0046
4. Mathur, Arunesh, et al. "Manipulative tactics are the norm in political emails: Evidence from 300K emails from the 2020 US election cycle." *Big Data & Society* (2023). https://doi.org/10.1177/20539517221145371