Understanding the Privacy Practices of Political Campaigns: A Perspective from the 2020 US Election Websites

Kaushal Kafle*, Prianka Mandal*, Benjamin Andow†§, Kapil Singh‡, and Adwait Nadkarni*
*William & Mary, Williamsburg, VA, USA; kkafle@, pmandal@, apnadkarni@wm.edu
†Google; andow@google.com
‡IBM T.J. Watson Research Center; kapil@us.ibm.com

Abstract—**Political campaigns are known to collect private user data, whether for building voter profiles, engaging with volunteers, or for soliciting donations. However, as such campaigns are classified as nonprofit in the United States (U.S.), their *privacy practices* have not received the same level of scrutiny as those of for-profit enterprises. This paper presents the Polityzer framework to evaluate the privacy posture of *political campaign websites*, and uses it to analyze 2060 campaign websites active during the U.S. election of November 2020. Our analysis leads to *20* key findings that demonstrate gaps in the privacy postures of political campaigns. For instance, we find that campaigns collect extensive private data they are not required to by the Federal Election Commission (FEC), and a vast majority do not provide any form of privacy disclosure. When disclosures are provided, they are often incomplete. We also found that campaigns may be inadvertently sharing data with other campaigns through common fundraising platforms, without disclosing such sharing. Reporting the lack of privacy disclosure to the respective campaigns yields further insights into the rationale behind their security posture. Finally, we discuss ways in which our results could enable future research, inform emerging privacy regulations, and transform user behavior regarding data privacy in this critical context.**

1. Introduction

Political campaigns are increasingly relying on their online presence, *i.e.*, social media, campaign websites and mobile apps, to engage with potential voters. Campaigns have been observed to leverage their web presence as one of the primary means of gathering information on voters, which is often combined with publicly and commercial sources to create accurate profiles of individual voters [8], [88]. Such information is often personal, *e.g.*, email, phone number, and salary, and highly private in some cases, *e.g.*, citizenship, partner's name and contact information, with serious privacy implications [57]. To our knowledge, while the use and impact of social media on election campaigns has been previously studied [30] [47] [68] [92], the privacy posture of *campaign websites* is yet unexplored at a large scale.

The privacy practices of campaign websites must be systematically studied for four reasons. First, political campaigns are (at least in the U.S.) generally classed as "non-profit organizations", and hence, data privacy regulations such as California Privacy Rights Act (CPRA) do not apply to them [4]. This gap in regulation may mean a lack of incentive in following privacy best practices. Second, while U.S. political campaigns are required by the Federal Election Commission (FEC) to collect donor names, mailing addresses, occupation, and employer [42], they may collect significant additional private data. Third, prior work shows that campaigns often share data [81], and deploy aggressive tactics to get users to submit information [100] or interact with their political emails [70]. Fourth, campaign websites are ephemeral in nature, and hence, it is unclear what happens to user data after the election. Thus, the user may lose any agency over their data to prevent future misuse, in which case, transparent disclosure of collection and sharing practices by the website is the only recourse for users. These factors, along with the increased transparency users desire regarding the use of their political data [93], and the governmental interest in regulating this space [39], [49], [59], [102], motivate us to empirically understand the privacy posture of campaign websites.

This paper describes Polityzer, a semi-automatic framework for a systematic, large-scale analysis of the privacy practices of political campaign websites. The design of Polityzer leverages the fact that political campaigns generally interact with potential voters, volunteers, and donors through the *campaign websites*, and thus the privacy implications of political campaigns can be approximated through a comprehensive analysis of campaign websites. We use Polityzer to analyze the websites of 2060 campaigns established for the 2020 US Presidential, Senate, and House elections, to answer four fundamental ***research questions (RQs)***:

RQ$_1$ (***Collection***) – What data do campaigns collect from their websites?

RQ$_2$ (***Disclosure***) – Do campaigns properly disclose the collection, sharing and retention of this data to users?

RQ$_3$ (***Conflict***) – Does the collection and sharing of campaign data conflict with their privacy disclosures?

RQ$_4$ (***Risk***) – Do campaign websites expose users to privacy or security risks such as malware or trackers?

Polityzer addresses these questions through a semi-automated methodology that combines text and website

§. This work was completed when the author was at IBM Research.

analysis: First, it extracts unique types of privacy-sensitive data collected by election campaigns, through an analysis of forms contained in campaign websites (**RQ**$_1$). Second, it compares the collected data types against the campaign website's privacy policy to assess whether campaigns properly disclose the collection to users (**RQ**$_2$). Third, it facilitates a study to measure the conflicts in a campaign's privacy disclosure (**RQ**$_3$) by examining the privacy policies of campaigns and fundraising platforms to understand potential/indirect collection (and sharing) not disclosed in a campaign's policy, but may occur due to the use of the fundraising platform. Finally, it leverages popular security tools (VirusTotal [13], ApiVoid [12]) to assess the general security and privacy-hygiene of campaign websites in terms of malware, hosting, and SSL/TLS misuse (**RQ**$_4$). The contributions of this paper are summarized as follows:

- **Polityzer:** We design and implement Polityzer to enable large-scale analysis of the privacy practices of political campaign websites. Polityzer is highly precise in terms of identifying campaign sites without privacy policies, with a false positive rate of 1.29%.
- **Study:** We use Polityzer to perform the *first large-scale analysis of the privacy practices of campaign websites*, analyzing 2060 sites of House, Senate, and Presidential candidates from the 2020 U.S. election.
- **Findings:** Our analysis leads to ***20 key findings*** that demonstrate significant privacy gaps. For instance, we find that 70.78% campaigns do not provide privacy disclosures, of which 64.27% that collect sensitive data. Similarly, even where privacy policies are present, 41.22% campaigns do not properly disclose data collection. Moreover, we find that 144/162 (88.89%) campaigns among those with privacy policies may be *inadvertently* (and without disclosure) sharing data with other campaigns through the use of the common fundraising platforms. We also find security weaknesses and use of trackers in campaigns that collect user data. These findings echo prior concerns regarding the privacy practices of political campaigns [37], [58], and demonstrate how websites are indeed used by campaigns to collect private data at scale, but without transparency and accountability.
- **Dataset:** To enable future research, we curate a dataset of 2060 campaign websites and 507 privacy policies belonging to senate, house, incumbents, and presidential candidates. Our artifact is available in our online appendix [83].

2. Motivation

Political campaigns collect user data from three key sources: publicly available information (*e.g.*, voter rolls), commercial sources (*i.e.*, data brokers or other campaigns), and their campaign websites and apps. Campaigns do not consider one source more important than the other, but instead, aggregate data collected from all sources to form complete profiles on individual voters [8], [88]. Websites, in particular, are critical for three reasons. First, websites enable campaigns to scale their data collection beyond what door-to-door campaigning allows. More importantly, websites provide campaigns with "organic traffic", *i.e.*, people who naturally navigate to the site or find it via search, and are hence much more likely to donate, volunteer, or register and provide email and other information [95]. As we see later in Section 6.2, this importance is evident from the fact that 40.91% of the campaigns registered to the FEC in 2020, including almost all of the winning campaigns, had active websites at the time of the election, many of which collected data outside of what the FEC mandates.

Second, websites also enable campaigns to reliably fill in the gaps in their voter profile databases obtained from other sources. For instance, voters have expressed the desire and ability to opt out from providing their contact information on voter rolls to prevent spam messages [25]; however, campaign websites and apps have been observed collecting the contact information of voters' friends and social connections [97], thereby completing profiles potentially against the voters' wishes. We see similar cases of data collection (*e.g.*, partner's name and email, friends' names and email) in Section 7.2. Finally, *after elections*, it is a standard practice for candidates to rent out or sell their databases to other candidates, PACs, political parties, or private brokers [84], potentially to recuperate campaign expenditure [97]. That is, campaign websites form an integral data collection vantage point for candidates, helping them collect, aggregate and monetize the data of citizens, in a manner that is quite similar to for-profit social networks, or other commercial websites that face far more scrutiny.

Therefore, given the critical position of campaign websites in a candidate's data aggregation apparatus, this paper seeks to empirically understand their privacy posture, further motivated by the increasing interest from governments, evidence of user concern over collection of sensitive political data, and the harm that may befall users if researchers overlook campaign websites, as this section describes.

2.1. Expectations of Governments and Regulators

Governments, at least in Europe, are cognizant of the privacy risk from data collection by campaign websites, and hold them to the same privacy standards as for-profit organizations. To elaborate, the European Union has taken the lead in protecting the privacy of political data collected by (campaign or other) websites, by classifying "political data" as a special, *opt-in*, category of personal data under the GDPR [49], *i.e.*, which cannot be processed without the owner's explicit consent. The GDPR also requires election campaigns to inform the users about collected data and the purpose behind the collection, and also to hold the collected data securely [39]. Individual countries in Europe have also issued specific regulatory guidance for political campaign websites, *e.g.*, the UK's guidance for processing personal data for political campaigning purposes [59] in compliance with both GDPR and the UK's Data Protection Act [99].

In contrast, the United States does not have a regulation that specifically governs private data collection by political campaigns. Instead, the U.S. has so far specified bare-minimum expectations that prevent the government (including members of the U.S. Congress) from misusing private voter data, such as the "Franking Privilege", which prevents members of congress from using such data for election campaigns [29]. However, *there are signs that this status quo is changing*, potentially due to the significant push for consumer data privacy around the globe, and calls by privacy advocates for presidential candidates' websites to be held to a higher standard than for-profits [97].

To elaborate, the U.S. Congress is currently considering the *Voter Privacy Act* [102] which seeks to grant voters access to personal data that campaigns have on them, and rights to erase their data from campaign databases and prohibit targeted ads. This act targets any political candidate, campaign, or entity using an "interactive computer service" for data collection, *i.e.*, a campaign website or mobile app. Among other things, this proposed act mandates campaigns to disclose the categories of personal information collected on an individual, thus significantly strengthening the transparency around campaign data practices. Such emerging regulations and existing precedents from the EU motivate our data-driven analysis of political campaign websites.

2.2. User Expectations and Desire for Privacy

There is strong evidence that users are increasingly concerned about data collected for political purposes, *e.g.*, a survey following the Cambridge Analytica scandal found that 73.9% of users were concerned about websites using their data for political purposes [93]. However, to our knowledge, there is no prior work that systematically studies what users precisely expect from political campaigns in terms of digital data privacy. To better understand user expectations in this context, we build upon prior work in Web privacy.

To elaborate, prior work [80] shows that in the general context of data collected by websites, *uninformed users* have no expectations of online privacy. However, *users desire strong privacy guarantees once they are exposed to privacy policies*, and informed on the ways in which their data is collected and used [80]. This shift in behavior is particularly evident in the for-profit context, wherein privacy studies [85], [87], [107] and pertinent regulation [3], [48] have caused user-awareness and privacy expectations to mature over the last decade. For example, a 2021 survey [26] showed that 86% users cared about data privacy, with 79% willing to act to protect their privacy and 47% already switching companies over data privacy practices.

We anticipate that the case of political data is no different, *i.e.*, *users only care when informed*. That is, even in the early stages of the use of Web data in political campaigning, *users expressed their discomfort when explicitly asked about profiling* and data aggregation by campaigns, with one noting that "I would not know any person who would be okay with some outside group having access to that much personal data" [86]. Thus, even if user expectations are unclear at present, it would be premature to conflate the lack of awareness with a lack of desire for data privacy in the political context, given ample evidence about how strongly the public feels about data privacy in general [23], [26], [35], [51], [78], [93]. Our hope is that our timely, empirical, evaluation of campaign websites, particularly during a major US election cycle, would inform users on what data campaigns collect and share, and motivate users to expect stronger guarantees comparable to the for-profit context.

2.3. Why Should Researchers Care?

Although Section 2.2 provides evidence that users are likely to expect privacy guarantees for political data, we also entertain a counter-possibility: *what if users don't care at all?* One might argue that since users support the campaign, they may not expect the privacy guarantees they expect from for-profits, effectively "donating" their data for a cause. This argument motivates a pivotal question for researchers: *should we exempt campaign websites from analysis simply because users blindly trust them and want to help them?*

The answer to this question is no, both due to the debatable premise that users trust campaigns to such a degree, and the severe harms that will result from overlooking political campaigns just because users trust them. To elaborate, there is no evidence to suggest that users trust political campaigns enough to forego the rights to their private information in perpetuity. In fact, a prior survey demonstrates that users have a very dim view of the public sector, with only 11% considering them as "trusted" to protect their private information [73]. This potential lack of immediate trust could explain why campaigns have to resort to aggressive tactics to collect data, *e.g.*, such as staging a photo op with Santa Claus for kids, and then required voters to sign up with their email addresses to download their children's photos [100].

Alternately, even if a user did trust a campaign and willingly provided data for a cause, such blind faith may be unwarranted and harmful, due to campaign behavior *after elections*. For instance, a candidate in the 2016 presidential election was found to have *sold the email list collected via their website to rivals*, and separately, rented the email list out, charging $10,500 per unsolicited email sent to their 675k subscribers [82]. Similarly, an app from a 2016 presidential campaign shared the contact list and location data of voters with Cambridge Analytica [101], potentially for building "psychological profiles" of the user and their contacts [98]. Regardless of how much the user supports a candidate's cause, it is implausible that they would condone behavior such as selling out their data to rivals.

Finally, campaigns often switch party affiliation after the election [18], or alternately, voters themselves change their minds [79], [106]. This transience of voter-campaign relationships makes it ill-advised to assume that voters providing data to a campaign relinquish their rights in perpetuity. Considering these factors, it is incumbent on researchers to analyze the privacy posture of campaign websites, and usher in increased transparency into their data collection and usage, to protect users even if, and precisely because, they may blindly trust campaigns.

3. Ethical Considerations

Politics is a sensitive subject, and we are cognizant of the several ethical “lines” that this work could cross if performed without significant care. Therefore, to preempt harm, and with the goal of *uncovering privacy gaps agnostic of political implications*, we imposed a set of ethical constraints on this work, based on four guiding principles: (P_1) *Focus on privacy, and not politics*, (P_2) *Limiting harm to candidates*, (P_3) *Limiting harm to campaign resources*, and (P_4) *Transparency*. This section describes these principles and the constraints we impose to adhere to them.

3.1. P$_1$: Focus on Privacy, and not Politics

Our goal is to highlight the gaps in the privacy postures of US political campaigns, and understand their implications *on user privacy*. As the loss of privacy affects users regardless of their political inclination, our position is that all political campaigns, regardless of affiliation, should adhere to privacy best practices. Therefore, we seek to limit our analysis and discussion to only what is relevant with respect to user privacy, and prevent a *partisan* interpretation of our results, as *user privacy should be a bipartisan issue*.

With this rationale, we impose the following constraint on the study: we refrain from analyzing our data in terms of specific political parties, affiliations, or the known political positions of individual candidates. To elaborate, we strip the party designation of candidates from the collected data before performing any analysis on it (which also prevents biasing ourselves), and do not later seek to attach party-specific insights in our findings. Instead, our analysis considers general congressional designations; *e.g.*, House, Senate and Presidential candidates, and committee memberships.

3.2. P$_2$: Limiting Harm to Candidates

Although prior work explicitly discloses the names of organizations with privacy gaps (*e.g.*, PolicyLint [15], PoliCheck [16], TaintDroid [38]), we deliberately refrain from disclosing the identity of candidates/campaigns in our findings to prevent reputational harm. To elaborate, we do not name candidates and anonymize any identifiable information when describing the data, results, findings, or our interaction with candidates. Similarly, we anonymize the composition of sub-samples of campaigns chosen for our sharing analysis in Section 9 to further mitigate harm. Finally, we paraphrase our interactions with campaigns rather than quoting them verbatim and redact any personal information to prevent harm and de-anonymization of the candidate. As our only correspondence was for the responsible disclosure of findings, the interaction does not warrant seeking an IRB approval [74].

3.3. P$_3$: Limiting Harm to Campaign Resources

Over the course of the study, we take several decisions to minimize or prevent harm to campaign resources or personnel. Particularly, we do not interact with any human subject or collect information about any human at any stage of our analysis. To analyze data sharing among campaigns, we use the same *black-box analysis* approach employed in prior work [81] [70]. Specifically, our experiment also involves providing a *valid but fictional email address* to the campaign website and monitoring sharing by analyzing emails automatically delivered to our inbox, without the involvement of any human subject. As expected, we only received automated campaign ads during the entire study, which causes negligible harm to the system. Our only interaction with the campaigns occurred after the study, during the disclosure of findings, as described in Section 12.2. Finally, our automatic crawler respects `robots.txt` files in campaign websites, which we then collect manually.

3.4. P$_4$: Transparency

While we take significant care to prevent a partisan interpretation of the paper (**P**$_1$), one might argue that not discussing political affiliations may in fact have a partisan effect, *i.e.*, that of hiding patterns of misbehavior in one political party. That is, the need for *full transparency* in terms of revealing the political affiliations is directly at odds with that of keeping the focus on privacy and off politics.

The position of the authors is that the benefits of a non-partisan interpretation of our results, in the form of enabling future research, informing the public, and motivating policy-makers to regulate campaigns regardless of affiliation, are far superior to the perceived loss of transparency, *i.e.*, we choose to adhere to **P**$_1$, at the cost of **P**$_4$. However, to enable future researchers (or the general public) to make full (transparent) use of our data and analysis, we will release our raw data and code, including the crawled campaign websites. As this data is already publicly accessible, releasing it does not cross any ethical boundaries, while our framework, Polityzer, will allow researchers to perform similar analysis based on political factors such as party affiliation, if they so choose.

4. Background

Federal elections in the U.S. can be divided into two categories (termed ’regular elections’ from hereon): i) the *Presidential election* that elects the President, and ii) the *Congressional elections* that elect the members of the Congress. The presidential election occurs every four years while the congressional elections occur every two years. Further, the US Congress is divided into two branches – the House of Representatives (termed simply ‘House’ from hereon), which consist of 435 voting members and six delegates, and the Senate, which consist of 100 members. The members of the House and Senate serve terms of two and six years respectively. Hence, all 435 members (and 6 delegates) of the House get elected every two years while only a third members of the Senate get elected every two years, as was the case in the 2020 election where 32 Senate seats and 441 House seats (including six delegates) were contested.

Outside of the regular elections that occur every two years, *special elections* are held when there are vacancies for any Congressional seats before the regular elections. There

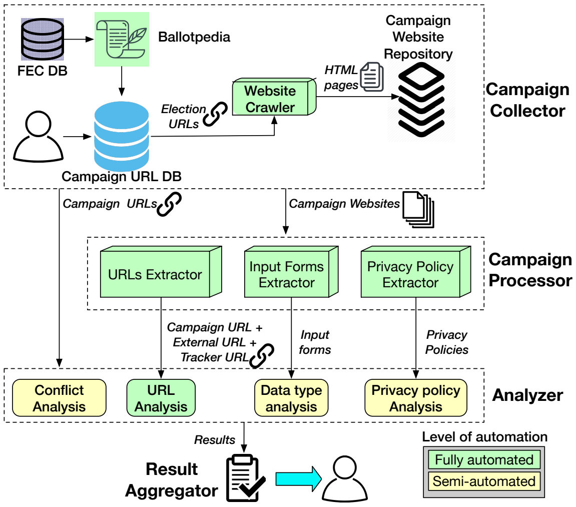

Figure 1. Overview of the Polityzer framework

were 11 special elections in 2020: eight for House seats and three for Senate seats [17]. Furthermore, all the regular elections to the House, the Senate and the Presidency are preceded by *primary* elections, where the political parties select a candidate that advances to the regular elections [69]. The analysis in this paper incorporates the election campaigns of all candidates (including those eventually elected) for both the Presidential and Congressional elections of 2020, including special elections and primaries.

Finally, all candidates who raise/spend over $5000 must register their campaigns and file financial reports with the Federal Election Commission (FEC) [43]. The FEC discloses this information in a searchable database, including the state and district that a candidate is registered in as well as the amount of money raised and spent [41]. In this work, we treat the FEC database as the ground truth regarding the candidates participating in the election.

5. Overview

Motivated by the criticality of campaign websites for user data collection, emerging expectations from governments regarding political data privacy, increasing user concerns, and the significant harms to users if campaign websites escape scrutiny, our analysis takes a *normative* position, *i.e.*, that *the websites of political campaigns be held to similar privacy standards as those of for-profit organizations*. This position reflects in our research questions (**RQ**$_1$→**RQ**$_4$), and guides our design and use of Polityzer. Figure 1 illustrates the design of Polityzer, composed of the following modules:

1. Campaign Collector: The Campaign Collector builds a database of the election campaign websites, and consists of a campaign URL database that houses the input website links of the campaigns, and an automated website crawler which downloads all the websites listed in the campaign URL database and stores them in a campaign websites repository for analysis. While campaign URLs can be directly fed into the Data collector manually, to enable large-scale collection, we develop an automated approach that collects candidate information from the FEC and uses Ballotpedia [19] to extract their campaign website URL (see Section 6.1).

2. Campaign Processor: The Campaign Processor automatically extracts useful information from the website pages using 3 submodules; (i) the *URL Extractor* that extracts the external/outbound links as well as any trackers that may be present, (ii) an *Input Form Extractor* that extracts the input forms in the website through which user data is collected, and (iii) the *Privacy Policy Extractor* which checks whether the website has a privacy policy document and extracts it.

3. Analyzer: The Analyzer performs 4 types of analyses on the processed websites; (i) *Data type analysis (**RQ**$_1$)* to understand the scope of private data collection (Section 7), (ii) *Privacy Policy Analysis (**RQ**$_2$)* to understand disclosure practices (Section 8), (iii) *Conflict analysis (**RQ**$_3$)* to discern conflicts in the privacy disclosures of campaigns (Section 9) and, (iv) *URL analysis (**RQ**$_4$)* to characterize the general security posture of the campaign website (Section 10).

4. Result Aggregator: Once the analysis is complete, the Result Aggregator generates a privacy report containing the aggregated results per campaign. We describe the results of each submodule of Analyzer in Sections 6→10.

6. Campaign Collector

For our analysis to be tractable, we use the campaign collector to download the websites belonging to the following federal races in the U.S.: (1) *House* elections, (2) *Senate* elections, and (3) *Presidential* election, as previously described in Section 4. We now describe our methodology, followed by an overview of the resultant datasets.

6.1. Website Collection Methodology

The collector module takes campaign links as input and downloads and organizes the websites per campaign type i.e. House, Senate or President. The campaign links can be fed to Polityzer automatically or manually, depending on whether an automated process is viable for a given election. For our analysis of the US-based campaigns, we constructed an automated pipeline using two data sources; (i) the FEC, and (ii) Ballotpedia. We now describe this pipeline followed by the methodology to build the campaign repository.

1. Obtaining candidate lists from the FEC: As discussed in Section 4, the FEC database is the ground truth for obtaining the list of all federal election candidates in the US election [43], which we use to obtain candidate names and metadata (*e.g.*, state, district, and party affiliation).

2. Resolving campaign websites of individual candidates: The FEC database does not link to the candidates' websites as candidates are only obligated to provide names, addresses, and party information. Hence, the challenge is *to obtain a candidate's campaign URL with only their name and metadata*, which we address using third-party sources along with web searches.

First, we search for corresponding candidate profiles in a non-profit political encyclopedia, *Ballotpedia* [19]. These profiles include the candidates' campaign URLs, which we seek, the accuracy of which is ensured by dedicated Ballotpedia staff. However, obtaining a Ballotpedia profile for a candidate is non-trivial. To elaborate, a typical URL for a candidate's Ballotpedia profile is simply `https://ballotpedia.org/<candidate-name>`. Just appending a candidate's name to this URL template does not work, as the candidates use their legal name in their FEC registration while Ballotpedia profiles use their informal names that are often different; *e.g.*, "John Doe" (Ballotpedia profile name) could be registered as "Jonathan Doe" (legal name) with the FEC. Further, to distinguish candidates with the same name, Ballotpedia also appends the candidate's home state to the profile URL, *i.e.*, `.../John_Doe_(State1)` and `.../John_Doe_(State2)`.

To address these challenges, we supplement our approach with Google search to identify the correct Ballotpedia profile URL of the candidates. For each candidate, we create a Google search query using both their FEC-provided name and the metadata (*e.g.*, state, party registration). We obtain the correct Ballotpedia profile by performing three additional checks on the top 10 search results, *i.e.*, we check if (1) the URL root is `ballotpedia.org`, (2) the URL format is that of a Ballotpedia profile, *i.e.*, `ballotpedia.org/candidate_name` or `ballotpedia.org/candidate_name_(state)`, and (3) the URL contains the candidate's last name. This approach always finds a candidate's Ballotpedia profile, if present, because the search query involves relevant information (*e.g.*, district/state) that automatically narrows down the search space, never returning >one match.

Once we have the Ballotpedia profile URLs for the candidates, we automatically extract the campaign website link listed in the profile using an HTML parser. Note that not all candidates have websites listed on their Ballotpedia profile. We do not directly perform Web searches for the campaign website for such candidates, as there is no ground truth outside Ballotpedia to validate the mapping.

3. Downloading websites to build the *campaign repository*: The collector module uses an automated crawler built using scrapy-selenium [27] that spawns a headless browser with a user-agent specifying Mozilla as the browser. The crawler takes input in the form of a campaign URL and downloads the HTML pages as well as other website resources (*e.g.*, PDFs and CSVs) into a campaign repository. The crawler uses a breadth-first-search approach, and traverses all links up to a fixed *depth* starting from the homepage. Default depth is set at two to allow the crawler to finish within a reasonable timeframe, regardless of the relative website sizes in terms of number of links present, but can be re-configured as per user needs. Note that the crawler does not traverse outbound links (*i.e.*, links that point to external web pages on other domains), and instead saves the link and terminates that specific search.

TABLE 1. DATASET OVERVIEW

Dataset	No. in our dataset	Total in FEC	Cand./seat
house_active	952 (90.49%)	1052	7.49
house_inactive	710 (31.84%)	2252	
senate_active	112 (94.92%)	118	16.375
senate_inactive	151 (37.44%)	406	
senate_incumbents	68	-	-
president_active	4 (100%)	4	1208
president_inactive	63 (5.32%)	1204	
total	**2060** (40.91%)	5036	

6.2. Results of Campaign Website Collection

We collected all the House, Senate and Presidential candidates that were registered for the U.S. general election of 2020 in the FEC database [41] as of September 10, 2020, which yielded a total of 5036 candidates. This list contained duplicates, as the same candidate may register for multiple offices simultaneously. Removing such duplicates brought the total number of unique candidates to 4885.

As described in Section 6.1, our automated pipeline to resolve candidate websites helped us map 3259 (66.7%) of the candidates to their respective Ballotpedia profiles. The missing 33.3% can generally be attributed to missing Ballotpedia profiles. Of these 3259 Ballotpedia profiles, 2630 provided a URL to the campaign website, *i.e.*, we obtained the campaign website links for 53.83% (2630/4885) of the overall sample of unique candidates identified from the FEC database. We further manually added 68 incumbent senators who were not participating in the election and hence, were absent from the FEC database, using the same methodology to obtain their campaign websites. We did not consider four retiring senators and one vice-presidential candidate whose website redirected to the presidential candidate's website.

Finally, we used the crawler to download the campaign websites. Not all campaigns were active at this point, since the primaries (see Section 4) had completed and many candidates had dropped out. As a result, some of the website links were dead or led to 404 errors. The collector successfully downloaded ***2060 (40.91%) websites***, including 13 that belonged in multiple groups *e.g.*, dropping their presidential campaign to run for house or senate. The downloads were completed on November 2, 2020, a day before the election.

To better understand the privacy practices in terms of candidates that appeared on the final ballot, versus those that had already dropped out, or those that are in office (and may or may not be re-contesting), we define a campaign "status" terminology. Campaigns whose candidates appeared on the final ballot in November were categorized as *active*, while those that did not (potentially due to a primary election loss) were classified as *inactive*. This classification was applied to the Senate, House, and Presidential campaigns, resulting in **6 datasets** (*i.e.*, *house_active, house_inactive, senate_active, senate_inactive, president_active, president_inactive*). Finally, the 68 incumbent senators not up for re-election were classified as *senate_incumbents*, the **seventh dataset**.

Table 1 shows the number of websites in the 7 datasets. The collector was able to download a higher percentage of websites for *active* candidates in all categories, which is expected, as active candidates were likely to have operational websites until the election. The lowest download rate occurred for *president_inactive* dataset, which is also explainable, as the number of candidates was quite large for what would eventually be one seat. Finally, *all campaigns in senate_incumbents had operational campaign websites even though they were not running*, potentially accruing user data even when the candidates may not be actively campaigning.

7. Analysis of Data Collection (RQ_1)

To obtain an estimate of what is at stake, Polityzer first seeks to understand the collection of private user data by campaigns. Certain data types must be collected by US campaigns to fulfill donor reporting obligations imposed by the FEC. Hence, Polityzer covers all data types, irrespective of the FEC requirements, but distinguishes FEC vs non-FEC data types when discussing privacy implications.

7.1. Methodology

Polityzer approximates the types of private data collected by campaign websites through an analysis of HTML forms, using the Input Forms Extractor in the Campaign Processor module shown in Figure 1. The Input Forms Extractor automatically extracts all forms from each campaign website and extracts the "labels" adjacent to each input field, producing a label-set for each campaign website to enable analysis.

We use a manual approach to identify *private* data types from the label-set for each campaign. Our goal is to identify private data in a political context, such as voter registration information or party affiliation, as these political types may not be present in existing ontologies, *e.g.*, that of PoliCheck [16]. For coverage, we ensure that the final set for each campaign contains the union of types identified from our labeling and those in PoliCheck's ontology.

7.2. Results and Findings

Table 2 shows the data types collected by campaigns and their distribution across our datasets. We split the collected data types into three categories: (1) *FEC-required* data types such as fine-grained location, employer information and occupation, (2) *non-FEC* data objects, *i.e.*, private data such as gender and party affiliation collected by campaigns at their own discretion, and (3) *FEC* data types*, *e.g.*, information such as credit card numbers, or banking information that are not explicitly required by the FEC, but need to be maintained as a part of the donation records/receipts. We have manually identified and associated the data types to the corresponding campaign websites in each relevant finding in this section.

***Finding 1* – 1462 (70.97%) of campaigns collected personal information through their websites ($\mathcal{F}_1$).** Table 2 shows the 28 unique data types collected, with at least one candidate collecting data of each type. We found an additional 139 (6.75%) campaigns collecting political opinions on issues such as education, guns, and abortion rights. Although these opinions are not PII, they are often collected along with at least an email or a phone number on the same form, and hence, are sufficient to allow campaigns to build user profiles without explicit consent.

TABLE 2. DATA COLLECTED BY CAMPAIGN WEBSITES. **FEC** INDICATES DATA REQUIRED BY THE FEC, !−**FEC** INDICATES DATA NOT REQUIRED BY THE FEC, **FEC*** INDICATES THE DATA NOT EXPLICITLY REQUIRED BY THE FEC, BUT OFTEN A PART OF DONATION RECEIPTS SHARED WITH THE FEC.

Class	**Data type**	**House**		**Senate**			**President**		**total**
		act.	*inact.*	*act.*	*inact.*	*incumb.*	*act.*	*inact.*	
FEC	*name*	661	397	82	72	47	4	34	1297
	location_coarse	532	228	66	40	51	4	19	940
	location_fine	313	134	36	23	24	3	8	541
	employer_info	100	52	18	13	13	4	5	205
	occupation	95	45	16	11	14	2	4	187
!-FEC	*email_address*	709	443	88	80	58	4	42	1424
	phone_number	478	252	70	39	39	4	19	901
	website	78	52	11	5	3	2	5	156
	education_info	16	5	2	1	0	5	0	29
	social_media	15	5	1	2	1	2	1	27
	language	6	9	0	0	0	0	1	16
	friend_email	2	1	0	1	4	0	1	9
	date_of_birth	2	4	1	0	1	1	0	9
	friend_name	2	1	0	1	4	0	0	8
	party_affiliation	4	2	0	0	0	2	0	8
	resume	4	1	0	1	0	1	0	7
	union_status	2	0	0	0	0	1	0	3
	photo_self	1	1	0	0	0	0	0	2
	fax_number	4	0	0	0	0	0	0	4
	age	0	3	0	0	0	0	0	3
	gender	4	3	0	0	0	0	0	7
	partner_name	1	0	1	0	0	0	0	2
	partner_employer	1	0	1	0	0	0	0	2
	parent_phone	1	0	0	0	0	0	0	1
	parent_email	1	0	0	0	0	0	0	1
	parent_name	1	0	0	0	0	0	0	1
	citizenship_status	0	0	0	1	0	0	0	1
	race	0	1	0	0	0	0	0	1
FEC*	*creditcard_info*	34	15	4	7	6	1	0	67
	banking_info	6	3	2	1	3	0	0	15
	pay_method	24	20	9	6	4	1	0	64
	retired_info	34	12	7	3	9	1	0	66

***Finding 2* – Campaigns collect highly private data types that are not required by the FEC ($\mathcal{F}_2$).** As shown in Table 2, all categories of campaigns, *i.e.*, house, senate or presidential campaigns as well as active or inactive categories, collect multiple data types that do not need to be reported to the FEC, indicating that the purpose of the website extends beyond just collecting donation. Email and phone numbers are the most common, which campaigns may associate with other sources to form complete profiles on users, as discussed in Section 2. Other sensitive non-FEC data types are *education information, date of birth, resume, gender, party affiliation* and in rare cases, even *race*, union status and *photo* of the user. Such data is highly private, and may enable invasive campaigns to profile voters, and sell/share the profiles in perpetuity (see Section 2.3).

***Finding 3* – Communication-related PIIs are the most collected data types ($\mathcal{F}_3$).** Among campaigns that collect at least one data type, email is collected by at least 98.96% campaigns (Table 7 in Appendix). Additionally, phone numbers are collected by 62.45%, meaning that communication related data types are among the most collected, and is indicative of the fact that communication with likely voters is one of the primary objectives of a campaign website. Note that some websites also collect username and passwords, most likely due to the presence of template login pages, as we verified with five randomly chosen websites. We have removed these two fields from the list of data types.

***Finding 4* – Campaigns collect data that impacts the privacy of people *other than the user* ($\mathcal{F}_4$).** We found that 12 campaigns collect contact information of the user's friends and parents, *i.e.*, their names and email addresses, echoing prior evidence of campaign apps collecting contact lists [10] (see Section 2). These requests are often presented as a means of *sharing the campaign*, *i.e.*, wherein the user "shares" the campaign with their friends by submitting the friend's name and email. Similarly, a campaign website collected information on the user's parents as part of a fellowship application. At least two campaigns also collected partner's name and employer information as part of the donation form. *Such collection is by nature without consent*, since the entity whose data is being shared cannot consent, and harms the privacy of people that are not directly interacting with the campaign. As we discuss later in $\mathcal{F}_6$, "friend's email and name" were among the data types not disclosed in the privacy policies of several campaigns.

***Finding 5* – Campaigns collect social and economic opinions of users along with their PII, thereby gaining the ability to directly associate individual users to their political leanings. ($\mathcal{F}_5$).** 69 campaigns collect opinions on various issues such as abortion, immigration, guns, and taxes through a survey page or their volunteer sign up forms. A user's beliefs on these issues is often a reliable indicator of their political leanings [46]. As it is a common practice to share voter data [81], this may lead to the exposure of the user to unsolicited micro-targeted ads from campaigns they did not directly interact with, even if the user trusts the specific campaign collecting the data (see Section 2.3). Of the 141 webpages of the 69 campaigns that collect such opinions, 127 pages (90.0%) from 61 campaigns (88.41%) do so while also collecting either an email or a phone number from the same page, allowing the campaigns to map the political stances to an individual. More importantly, using the analysis from Section 8, we find that *33/61 (54.10%) of these campaigns lack a privacy policy*.

In summary, the collection of such a wide-range of private data underscores the critical position of websites in the data aggregation apparatus of campaigns.

8. Analysis of Privacy Disclosure (RQ$_2$)

Privacy policies play a critical role in conveying how campaigns handle the significant sensitive data they collect via their websites. Moreover, privacy policies may also reveal if the data is being shared with third-parties such other political campaigns. This analysis focuses on understanding whether the privacy policies in campaign websites properly convey such relevant information accurately to the user.

8.1. Methodology

As existing privacy regulations (*e.g.*, CCPA [3]) do not apply to political campaign websites, we do not assume that a campaign website will have a "Privacy Policy" link on the main page, as is the best practice, also mandated by CCPA. Instead, we seek to find *any* form of privacy disclosure in the website, which allows us to deduce *if* campaigns disclose their privacy practices to users, and how accurate the disclosure are with respect to the private data they collect. Our analysis is organized in two steps:

Step 1 – Checking the campaign website for *a* privacy disclosure: For each campaign website, we first attempt to check if the website provides a privacy disclosure, *i.e.*, not a privacy policy per se, but any document that describes the collection and sharing of private data. For this, Polityzer searches campaign websites using a bag of words approach which searches for any hyperlinks (*e.g.*, https://⟨campaign-url⟩.com/privacy-policy/) or link-text (*e.g.*, "Privacy Statement") that may contain terms indicative of a privacy disclosure. We obtain this set of disclosure-related terms from prior work [67], and increment it with additional words that may indicate privacy or any legal disclosure, specifically, "terms", "conditions" and "disclosure", leading to the following set of privacy disclosure-related terms: [privacy, terms, conditions, notice, statement, disclosure].

Finally, after Polityzer automatically extracts all the hyperlinks containing the disclosure-related terms in a campaign website, we manually check each shortlisted link to confirm whether it truly leads to a privacy policy page, which we extract for further analysis. In the spirit of performing a conservative analysis, we consider cases where the link pages led to empty/error pages as "having a privacy policy", given the presence of the hyperlink.

Step 2 – Analyzing the privacy policies for collection and sharing accuracy: We use Polisis [53] to extract collection and sharing statements from the privacy policies, followed by manual annotation of data mentioned in the statements, and a comparison with the data actually collected by the campaigns, as found in Section 7. To elaborate, for each privacy policy, we automatically use the Polisis API [54] to obtain a category prediction for each sentence (*e.g.*, "collection", "sharing"). To elaborate, we extract sentences receiving the highest confidence scores for "collection" or "sharing", which are most relevant to our goal.

Once the collection/sharing sentences are identified, we manually annotate them to identify data objects in the sentences (*e.g.*, name, address), looking for *precise* data objects and ignoring generic terms. For instance, in the sentence "we may collect personal information including your name and email", we annotate only the name and the email address. This annotation was performed by two authors in 10 days, wherein the first author identified the data objects in the sentences, which were then confirmed by the second author.

Finally, for each campaign, we compare the data objects extracted from the sentences with the corresponding set of data objects collected by the campaign's website (*i.e.*, obtained from our analysis in Section 7). This analysis allows us to identify several types of anomalies, including data objects that the campaign collects but does not disclose in the privacy policy. Further, we also analyze the sharing

TABLE 3. MISSING PRIVACY POLICIES IN CAMPAIGN WEBSITES

	w/o priv.policy	collecting priv. data
house_active	640/952 (67.22%)	441/640 (68.98%)
house_inactive	589/710 (83.38%)	359/589 (60.95%)
senate_active	52/112 (45.54%)	34/52 (65.39%)
senate_incumbents	15/68 (23.53%)	11/15 (73.33%)
senate_inactive	122/151 (82.12%)	65/122 (53.28%)
president_active	0 (0%)	0 (0%)
president_inactive	40/63 (66.67%)	27/40 (67.5%)
total	1458/2060 (70.78%)	937/1458 (64.27%)

sentences to identify explicit mentions of sharing data with other candidates, campaigns, or committees, and perform a general search for mentions of FEC disclosure requirements.

8.2. Results and Findings

Out of the 2060 websites we analyzed, Polityzer's automated approach for privacy link extraction led to 975 campaigns with potential privacy disclosures (*i.e.*, 975 links), and conversely, 1085 websites without disclosures. Recall that Polityzer's automated approach is conservative, *i.e.*, it gives significant benefit of the doubt to campaigns and over-approximates to find *all potential* privacy disclosures. Hence, its effectiveness is in terms of a low false positive rate, with a positive being the lack of a privacy disclosure. Upon manual validation of this result, we find only 14/1085 false positives, *i.e.*, a false positive rate (FPR) of 1.29%. Of the 14/1085, four used different terms to describe their disclosure (*i.e.*, disclaimer, transparency, fine print, and ToS), five were hosted with third party sites (*e.g.*, *pastebin*), and five resulted from link extraction errors in Polityzer.

We further manually refined the 975 potential privacy disclosure links from Polityzer, and found 560 actual privacy policies, and 415 were not. Of these 560 we were able to extract a ***final 507 disclosures for analysis***, with the rest leading to 404 errors. Table 8 in the Appendix shows the distribution of analyzed privacy policies per dataset.

We also validated the 415 websites without disclosures from our manual refinement of the 975 potential disclosure links, and found that 28 were marked in error. To summarize, we identified 1458 websites as lacking privacy disclosures, considering 42/1500 false positives (i.e., overall 2.8% FPR), with Polityzer's automated keyword-based approach suffering from only 1.29% (14/1085) FPR.

Finally, recall that we used Polisis to identify collection and sharing statements from privacy policies, and filter out the rest. Thus, "effectiveness" in this context would be the ability of Polisis to identify most of the collection/sharing sentences, and conversely, filter out or miss as few as possible, *i.e.*, have few false "negatives" (with a positive being a relevant collection/sharing sentence). To understand if using Polisis filters out relevant sentences, we manually validated each of the 14,454 sentences filtered out by Polisis. We found that only 465/14454 sentences were incorrectly labeled as not being "collection" (339 sentences) or "sharing" (126 sentences), *i.e.*, a false negative rate of 3.22%. We manually integrated these sentences into our analysis.

Our analysis of privacy disclosures led to the following findings, which have all been manually validated.

TABLE 4. DATA COLLECTION WITHOUT PRIVACY DISCLOSURE.

Dataset	Undisclosed
house_active	118/267 (44.19%)
house_inactive	49/90 (54.44%)
senate_active	10/56 (17.86%)
senate_incumbents	16/49 (32.65%)
senate_inactive	7/19 (36.84%)
president_active	3/4 (75%)
president_inactive	6/22 (27.27%)
total	209/507 (41.22%)

***Finding 6* – 1458/2060 (70.78%) of the campaign websites did not have a privacy disclosure, of which, 937/1458 (64.27%) collect private data ($\mathcal{F}_6$).** As shown in Table 3, of the 70.78% campaigns lacking privacy disclosures, it is concerning that 937 (64.27%) collect private user data, including sensitive information such as credit card, employer information, phone number, and location. We also observe that active campaigns were more likely to offer a privacy disclosure than inactive campaigns, likely due to the longer period of time the campaign websites were actively engaging with users and potentially subject to scrutiny.

***Finding 7* – 209/507 (41.22%) of campaigns do not fully disclose all private data in their privacy policy ($\mathcal{F}_7$).** Similar to $\mathcal{F}_5$, Table 4 shows how inactive campaigns are more likely not to disclose all collected private data, relative to active campaigns. Table 15 in the Appendix lists the top 10 undisclosed types. These data types that were not disclosed include (1) data types that were most collected such as phone number (111/507 or 21.89%), email (88/507 or 17.36%) and location (47 or 9.27%), (2) data types known to be shared with third-parties (including the FEC) such as occupation (24/507 or 4.73%) and employer information (29/507 or 5.72%), and (3) sensitive demographic data such as retirement status (10/507 or 1.97%), party affiliation (4/507 or 0.79%) and race (2/507 or 0.39%). Even data types that affect the privacy of the user's friends (*e.g.*, friend's email, 2/507 or 0.39%) were not disclosed. Conversely, 316/507 (62.33%) campaigns disclose data types that they do not collect in practice, potentially due to the use of templates (see Appendix B).

***Finding 8* – 389/507 (76.73%) campaigns disclose sharing with third-parties in their privacy policy ($\mathcal{F}_8$).** However, this does not mean they explicitly mention who the third-party is; *i.e.*, campaigns may not mention sharing with other political campaigns, or even with the FEC. As campaigns are known to sell voter data after elections (see Section 2.3), this lack of transparency is particularly concerning.

***Finding 9* – 179/507 (35.31%) campaigns mention sharing data with other political campaigns ($\mathcal{F}_9$).** Sharing with other campaigns, especially to the campaign's central party, may be a standard practice [81]. However, only 35.31% of mention sharing their data with other political campaigns, despite its privacy implications on the user. This trait is more prevalent among *active* campaigns (see Table 10 in the Appendix). Similarly as in the case of $\mathcal{F}_8$, this finding reflects the norm for campaigns to share data with other campaigns, sometimes even rivals [82].

***Finding 10* – None of the campaigns explicitly discuss their retention practices in relation to the completion of campaign. ($\mathcal{F}_{10}$).** We used keyword-matching using ‘retain‘ or ‘retention‘ to find 165/507 campaigns that discuss retention practices. Of these 165 campaigns, none of them explicitly discuss what happens to user data upon the *completion of the campaign*. When campaigns do discuss retention of data for a period of time in 54 instances, they do so by providing vague and non-descriptive explanation. For instance, campaigns may say they retain data as long as ”necessary for business purpose“, or ”necessary for fulfillment of purpose for which data was given“. Further, with the campaign likely ending, users have no recourse to prevent data misuse once it is collected. This finding shows why researchers must investigate campaign websites regardless of how much users trust them (Section 2.3), as without committing to a retention policy, campaigns gain perpetual access to user data, which is undesired given the transience of user-campaign relationships, and general campaign (mis)behavior after elections.

9. Inter-campaign Sharing Analysis (RQ$_3$)

Users may expect their shared data to stay with the campaign they engaged with. This section explores data sharing among campaigns through an experiment, and by analyzing the conflicts between the privacy disclosures of political campaigns and major fundraising platforms.

9.1. Methodology

We performed two studies to understand the privacy implications of data sharing among campaigns. First, we conducted an *email experiment* to evaluate if candidates share private email addresses with others. Second, we identified campaigns connected to the two most prominent fundraising platforms and analyzed the privacy policies of the platforms as well as the connected campaigns for conflicts.

1. Experimentally studying email data proliferation: We signed up for the newsletters of 26 campaigns evenly distributed among major political parties, consisting of ten Senate, ten House and two Presidential candidates, as well as four House and Senate campaigns from our state, to increase the likelihood of a response owing to the local reference.

To observe the effects of sharing our emails with each of the 26 entities in isolation from the others, we used 26 dedicated email accounts. We examined the domains of the senders to identify emails from external parties, *i.e.*, if the domain differed from that of the campaign website we signed up with. Additionally, we visited the external domain to identify its affiliation (*i.e.*, another campaign, or a PAC). We refer the readers to the study by Podob et al. [81] for a more exhaustive analysis of the sharing practices involving all major campaigns in the 2016 election.

2. Understanding conflicts among the privacy policies of platforms and campaigns: We observe that political campaigns often create parallel instances on fundraising platforms such as ActBlue [1], WinRed [5], Anedot [2] or DonorBox [7], which act as payment providers and also central avenues for attracting voters. As campaigns point to these platforms from their websites, data exchange between the campaign website and the fundraising platform is highly likely. Therefore, we explore conflicts between the privacy policies of campaigns and their fundraising platforms.

TABLE 5. USE OF Platform$_1$ OR Platform$_2$ FOR FUNDRAISING.

Dataset	Use platform	No privacy policy
house_active	567/952 (59.56%)	341/567 (60.14%)
house_inactive	269/710 (37.89%)	210/269 (78.07%)
senate_active	63/112 (56.25%)	11/63 (17.46%)
senate_incumbents	55/68 (80.88%)	13/55 (23.64%)
senate_inactive	60/151 (39.74%)	46/60 (76.67%)
president_active	3/4 (75.0%)	0/3 (0%)
president_inactive	22/63 (34.92%)	10/22 (45.45%)
total	1039/2060 (50.44%)	631/1039 (60.73%)

We design a simple approach to identify campaign websites connected to two major fundraising platforms, Platform$_1$ or Platform$_2$. For each of the 2060 candidate websites we automatically extract all the outbound links, and identify a connection if the root of any of the outbound link is Platform$_1$ or Platform$_2$. We then compare the collection and sharing statements from the privacy policies of the two platforms with those of the connected campaigns (extracted using the methodology described in Section 8).

9.2. Results and Findings

We categorize the emails received between November 5, 2020 to February 20, 2021 as *during-election* emails, as the period includes the November election and the Senate runoff election in Georgia. The emails from February 21, 2021 to September 14, 2022 are categorized as *after-election* emails. We received 1708 during-election emails and 933 after-election emails, with an average of 65 during-election and 35 after-election emails per campaign. All the findings from the analysis of these emails have been manually validated.

***Finding 11* – 3/26 campaigns shared our email with another entity without disclosing in the privacy policy ($\mathcal{F}_{11}$).** In total, 8/26 (30.77%) campaigns we studied shared our emails with other political entities (such as PACs), five during-election and three after-election. Of these eight, *three make no mention of sharing user data with other political entities* in their respective privacy policies.

***Finding 12* – Of the 1039 campaigns that use fundraising platforms, 631/1039 (60.73%) do not have a privacy policy ($\mathcal{F}_{12}$).** According to their privacy policies, both Platform$_1$ and Platform$_2$ *share user data with the campaigns*. However, since 631 such campaigns do not have privacy disclosures, the privacy policies of both platforms are rendered ineffectual in practice, *i.e.*, the privacy guarantees promised by the platforms to donors do not hold, due to data sharing with campaigns that provide no disclosure or guarantees at all.

***Finding 13* – Campaigns using Platform$_1$ for fundraising may be indirectly sharing with other campaigns ($\mathcal{F}_{13}$).** Platform$_1$ in its privacy policy states that it may share user data with third parties for marketing purposes, including *other political committees or campaigns that may be of interest* to the user. This means that users donating to one campaign may have their data shared with other campaigns.

Hence, we argue that in the spirit of good disclosure, campaigns should explicitly specify such sharing in their privacy policies. However, of the 162 campaigns that use $\mathsf{Platform}_1$ and have a privacy policy, 144 (88.89%) campaigns do not disclose sharing with other campaigns or $\mathsf{Platform}_1$, despite using $\mathsf{Platform}_1$ for fundraising.

10. Security Risk Analysis (RQ$_4$)

Campaign websites collect extensive amounts of private and sensitive user data (Section 7), often without adequate disclosure (Section 8, Section 9). Therefore, it is important to evaluate the general security hygiene of these websites to develop an understanding of the risks associated with malicious or unintended data disclosure. We performed three analyses, described below, followed by the findings.

1. Identifying malicious/phishing URLs: Building upon prior work [24], [56], [66] that use Virustotal [13] as ground truth for identifying malicious websites, we first analyzed each campaign website by passing each URL (including outbound links) included in the website through VirusTotal's API. We then aggregated the results from VirusTotal by checking how many of the scanning engines in VirusTotal marked a URL as either "malicious" or "phishing".

2. Identifying and characterizing trackers: To check for the presence of trackers, we used an existing tracker list [36] that was originally designed for AdBlock [6]. For each campaign website, we check if any of the associated URLs match the regular expression-based rules from the tracker list. We also extract the root domain of the tracker upon identification. As a final step to identify *malicious* (*i.e.*, and not just undesirable) trackers, we run the list of URLs of discovered trackers through VirusTotal.

3. Checking whether websites use TLS: We send a GET request to the website URL using https and label a website as "using TLS" if the response is successful.

Finally, we use APIVoid [12] to obtain general hosting information for deducing the jurisdiction that would apply to the campaign website in case security problems were found. For each candidate URL, we query APIVoid using REST APIs and obtain the following information: the IP address of the server hosting the URL, server's hosting company, and the country where the website is hosted. We further analyze this data to identify campaigns hosted outside of the US, since this study focuses on US political campaigns.

Finding 14 **– Campaign websites are generally secure (**$\mathcal{F}_{14}$**).** Although 17/2060 (0.82%) campaign websites were flagged as unsafe by at least one of the engines in VirusTotal (Table 11, as we show in the Appendix, only four among them were flagged by at least two engines and only one by more than two engines. This shows that at least 2052 (99.18%) campaign websites were marked as secure by VirusTotal. Our results are conservative as we cannot analyze false negatives, *i.e.*, confirm the absence of malicious code in candidate websites, since Polityzer's dataset only consist of html pages and not the associated scripts.

Finding 15 **– Campaign websites are hosted on servers outside of the US (**$\mathcal{F}_{15}$**).** In all, 53/2060 websites were hosted by servers located outside the US, 15 of which were well-known service providers such as CloudFlare and Google. After their removal, we finally got 38 campaign websites hosted outside the US, in countries including Czechia, Denmark, and Japan (full list in Table 13 in the Appendix). We find that 33 of these sites belonged to *inactive* campaigns while 5 belong to *active* campaigns.

As this analysis was performed after the election, it is unclear if the websites were always hosted offshore, or bought by an offshore entity after the election. It is also possible that once the campaign ends, URLs are bought by offshore entities for potentially malicious future use, *e.g.*, one URL in an *inactive* House campaign server is hosted by a company called the *Iranian Research Organization for Science & Technology* located in Hong Kong.

That five *active* campaigns are hosted offshore is concerning, as they were still active and collecting data at the time of the analysis. Due to the diverse laws govering data in different countries, such offshore storage can have serious privacy implications for users; *e.g.*, recent changes in Hong Kong's data security laws that allow the government to access data stored in Hong Kong's data centers [9], [11].

Finding 16 **– 168 (8.16%) campaign websites do not use HTTPS for communication (**$\mathcal{F}_{16}$**).** We observe that HTTPS adoption rate among the campaign websites may be better than HTTPS adoption in general, which is around 80% for Alexa top 100,000 [105] [104]. As shown in Appendix Table 12, 86 (51.19%) of these non-HTTPS campaign websites collect PII including phone numbers, fine-grained location, and credit card data.

Finding 17 **– Campaign websites have malicious outbound links (**$\mathcal{F}_{17}$**).** 71 campaigns had at least one outbound link that was classified as malicious by at least two engines in VirusTotal. While the campaign website is unlikely to be malicious, this result indicates that the links that campaigns include within their website may not be adequately vetted.

Finding 18 **– 1504 (73.01%) campaign websites use trackers (**$\mathcal{F}_{18}$**).** Trackers are used extensively among the campaign websites, but are more likely in *active* campaigns (see Table 14 in the Appendix). We found 280 unique trackers in the websites with `www.google-analytics.com` and `connect.facebook.net` being the two most common. Among the 280 trackers, two were identified as malicious: `ad.yieldmanager.com` (by one VirusTotal engine) and `www.freeresultsguide.com` (by three engines).

Finding 19 **– 974/1504 (64.76%) of campaign websites with trackers do not have a privacy policy (**$\mathcal{F}_{19}$**).** Similar to Findings $\mathcal{F}_5$ and $\mathcal{F}_{10}$, *inactive* campaigns across each data set are more likely to not have privacy policies despite having trackers in their websites, which may lead to their users not even being aware of possible data collection.

Finding 20 **– 112/446 (25.11%) campaign websites do not mention trackers in their privacy policies (**$\mathcal{F}_{20}$**).** This is in keeping with the privacy policies of campaign websites missing key data types, as we detailed in Finding $\mathcal{F}_6$. In both Findings $\mathcal{F}_{17}$ and $\mathcal{F}_{18}$, it is important to note the loss of user data privacy resulting from trackers [40] *e.g.*, trackers from

Facebook or google analytics can collect privacy-sensitive user data from websites [71].

11. Related Work

This work is the first to analyze the privacy practices of political campaign websites at scale, and is closely related to work in the following areas.

Election data and security analysis: The Internet Society has performed data protection, privacy, and security analysis of presidential campaigns since 2016 [62]. The 2020 audit was limited to 20 presidential campaigns, whereas our work covers a 2060 senate, house, and presidential campaigns, and provides specific analysis of collected data-types relative to privacy policies. Further, our email study builds upon Podob et al. [81]'s work, which found evidence of sharing among campaigns and PACs. In addition to what Podob et al. study, we also explore contradictions between a campaign's sharing practice and their privacy policy, and show the *types* of data being shared and the issues in disclosure practices.

Consolvo et al. [31] conducted interviews with campaign personnels to understand their security practices and perceptions about the use of digital assets, finding vulnerabilities and risk due to this use. To safeguard against such vulnerabilities, various organizations have outlined recommendations for campaigns [44], [45], [60], [75]. Our work instead focuses on the campaign websites and complements prior work by outlining the gaps in privacy and security posture of campaign websites. Finally, prior work [70] has also studied the deceptive and clickbait-oriented tactics in campaign email contents to incentivize interaction. We instead seek to understand how the campaign websites collect the contact information (including email) of the users, and how its use and sharing is disclosed to the users.

Targeted ads, profiling using social media: Prior work has focused on the privacy impact of social-media based voter profiling and targeted advertising, especially following the Cambridge Analytica scandal [63] [91] [55] [89]. Additionally, prior research has analyzed the impact of voter profiling through big data on election outcomes [50] and tried to predict election results based on user activities collected from twitter [94] [47]. In contrast, our work focuses on campaign websites and user data that is collected directly through that medium by the campaigns, rather than targeted advertising through social media. Further, prior work has also raised concerns about voter privacy in the age of online political campaigning [57], [65]. We build upon such concerns and study the privacy impact of political campaigns through the overall privacy posture of their websites.

Privacy policy analysis: Prior work has analyzed various aspects of privacy policies such as their availability in mobile apps [21], [34], [52], [109], readability and comprehension [22], [64], [72], as well as analyzing their vagueness [16], [20], [53], consent and opt-out choices [76], [90], contradictions [15], [33], [53], [108], and regulatory compliance [21], [22]. Our work instead focuses solely on the privacy policy availability of campaign websites and performs a semi-automated analysis on the policy text to glean disclosed data objects, for which we leverage past works (*e.g.*, Polisis [53]) where appropriate.

12. Discussion and Conclusion

The severe privacy violations demonstrated in our analysis are all legal in the U.S., given the lack of a dedicated privacy regulation applying to political campaigns. However, "legal" does not mean "appropriate" here, *i.e.*, similar violations by commercial websites would have attracted significant scrutiny and criticism from both regulators and researchers, as they have in the past [28], [32]. Our position is succinctly captured in this quote from the Online Trust Alliance [61], who audited apps belonging to presidential candidates in the 2016 election: *In light of worldwide privacy concerns and the court of public opinion, are the candidates' practices considered responsible or ethical? Should the next president of the United States be held accountable to the same standards as a business?* [97]. That is, we believe that the findings from this study expose inconsistencies that go beyond what is expected in keeping with the spirit of good disclosure, as well as laws in prominent (non-US) jurisdictions, and general consumer expectations of privacy. Several campaigns that follow privacy best-practices may adhere to this view as well, such as the 29.22% that provide privacy disclosures, 58.78% of which disclose all data collection.

In summary, the last decade has seen significant advancements in consumer data privacy due to the concerted efforts of researchers to change the perceptions of both governments and consumers, and this work seeks to initiate a similar transformation in this highly relevant domain. To this end, we organize the discussion along three key areas: We summarize the privacy implications of the findings (limitations discussed in the Appendix A). We then explore *why* the campaigns' privacy postures are this way, leveraging the responses from campaigns contacted for responsible disclosure. Finally, we conclude with actionable outcomes from this study that would help researchers as well as regulators bring about tangible change and accountability in data privacy in this domain.

12.1. The Privacy Posture of Campaign Websites

Our findings show that campaign websites collect extensive amounts of highly sensitive data ($\mathcal{F}_1 \rightarrow \mathcal{F}_5$), and confirm the important position websites occupy in the campaigns' data aggregation apparatus, as outlined in Section 2. While the significant collection of private data *not required by the FEC* ($\mathcal{F}_2$) is indeed concerning, we find that the privacy risks from this collection are made severe due to the sharing practices, and a general lack of transparent disclosure.

To elaborate, aside from common privacy violations, such as the lack of a privacy policy ($\mathcal{F}_6$) or incomplete policies ($\mathcal{F}_7$), we find that many campaigns use boilerplate language regarding sharing ($\mathcal{F}_8$), and some even share data with other campaigns without disclosing such sharing at all ($\mathcal{F}_{11}$). Similarly, most campaigns with access to data from fundraising platforms lack privacy disclosures entirely ($\mathcal{F}_{12}$), rendering ineffective the guarantees promised in the platform's disclosure, as well as the disclosures of other campaigns that provide data to the platform ($\mathcal{F}_{13}$).

What is worse is that no campaign precisely discloses what happens to the data after completion of the campaign ($\mathcal{F}_{10}$). The implication here is that once users provide data, campaigns may use, share, and sell the data in perpetuity, without the data owner's consent. This perpetual ownership campaigns acquire not only exposes users to privacy harms (*e.g.*, data being sold to rivals [82]), but also to security risks such as identity fraud and surveillance given the potential for data leaks [14], [96], especially in cases where the websites are hosted in non-US jurisdictions ($\mathcal{F}_{15}$), use vulnerable communication ($\mathcal{F}_{16}$), or are connected to malicious, non-vetted, entities ($\mathcal{F}_{17}$). The undisclosed presence of aggressive trackers in many campaign websites ($\mathcal{F}_{19}$, $\mathcal{F}_{20}$) compounds these harms, by exposing the user's general browsing habits to the campaigns as well.

To summarize, the collection of a significant range of private data, coupled with insufficient disclosure, bad security practices, and undisclosed sharing, exposes users to significant privacy risks and loss of agency. To put these findings into context, we make two final observations:

Observation 1 ($\mathcal{O}_1$) – The campaign websites of 253 current *lawmakers* did not have a privacy policy, of which, 200/253 collect personal information.

Observation 2 ($\mathcal{O}_2$) – 99 of these 253 lawmakers serve on privacy-relevant congressional committees on cyber, technology, or consumer protection. Such committees often scrutinize the security or privacy practices of businesses, *e.g.*, 4 of these lawmakers participated in a Senate hearing titled *"Does Section 230's Sweeping Immunity Enable Big Tech Bad Behavior?"*. We hope that the findings from this study help responsible members of congress in holding their own campaigns to the same standards they govern.

12.2. Rationale for the Present Privacy Postures

During the responsible disclosure of our findings to campaigns without privacy policies, we received 20 responses that provide insight into the campaigns' rationale regarding data privacy (see the online appendix [83] for details).

Particularly, 6/20 campaigns were open to adding a privacy disclosure to their websites, but were ill-equipped due to the lack of technical support or privacy know-how, some even asking us for a template. These responses are encouraging as they show a *willingness to follow privacy best-practices*. In contrast, 5/20 campaigns misunderstood the rationale behind privacy disclosures, (incorrectly) arguing that privacy policies are non-binding, and hence ineffectual. Another argued that since they did not collect data (which we verified to be correct), they did not need one, which goes against commonly understood best-practices.

Further, some (3/20) did not consider their campaigns active, and hence saw no need to retroactively add a privacy policy. However, we note that the websites were still active at the time of this exchange, and could have been collecting data. Some others confused our inquiry with their stance on privacy in general, or mistook us as service providers proposing to create a policy for them (which could also explain the lack of responses from campaigns).

Finally, 2/20 campaigns admitted that the absence of the privacy policy was directly *because of the lack of federal privacy regulation* for campaigns. One candidate expressed frustration at their party's privacy posture, and suggested us to convince their party to require their candidates to have a privacy policy. The same candidate stated that they were asked by the central party to share the campaign's donor list, corroborating our findings. Finally, the candidate also expressed the need for dedicated resources for campaigns, such as a website that explains the best practices, provides templates, and how-tos, thereby aiding the largely volunteer-run campaigns to develop a good privacy posture.

12.3. Towards Privacy, Transparency, and Accountability in Political Campaign Websites

This paper develops artifacts and insights that will benefit researchers, users, and policymakers alike. Particularly, our data, results, and the Polityzer framework will help researchers further explore privacy in the context of political campaigns, and extend our methodology and analysis pipeline to analyze other relevant artifacts, such as campaign-related mobile apps. Moreover, researchers will be able to use Polityzer to periodically evaluate campaigns, enabling longitudinal understanding of the privacy postures of political campaign websites.

Similarly, we are encouraged to see legislative efforts [102] towards regulating the privacy practices of political campaigns in the U.S., particularly their digital components, such as websites and apps. We envision that the measurement results and findings from this study, as well as future research that builds upon it, will provide empirical grounding for such legislative efforts. For instance, our findings motivate the dire need to require campaign websites to provide privacy disclosures, particularly including details on how long they retain user information. Such a criteria will not only force campaigns to be transparent about their routine sharing or sale of data after the election, but also enable users to make informed choices when committing their data to a particular campaign.

Finally, we see significant privacy benefits to users down the line. Particularly, we find the general obscurity on the users' privacy expectations from campaigns unsurprising, given the lack of privacy studies to that end. The data and findings from our large scale study provide a unique opportunity for researchers to address this gap, repeating prior work on gauging user privacy expectations [80] in this critical context. More importantly, we hope that just as prior work [80] found, presenting the public with the key measurements and findings regarding the privacy practices of campaign websites may also educate them on their privacy implications, motivating *informed* voters, who will then demand increased privacy, accountability, and transparency, from the campaigns. This, in turn, may be the final push needed for strong privacy legislations governing political data in the U.S., just as user privacy concerns motivated regulations such as the GDPR and CPRA.

References

[1] Actblue. https://secure.actblue.com/. Accessed August 2022.

[2] Anedot — powerful giving tools made easy. https://www.anedot.com/. Accessed August 2022.

[3] California consumer privacy act (ccpa). https://www.oag.ca.gov/privacy/ccpa. Accessed August 2022.

[4] California privacy rights act (cpra). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.\&part=4.\&lawCode=CIV\&title=1.81.5. Accessed August 2022.

[5] Directory — winred. https://winred.com/. Accessed August 2022.

[6] Easylist - overview. https://easylist.to/. Accessed August 2022.

[7] Free donate button - donorbox nonprofit fundraising software. https://donorbox.org/. Accessed August 2022.

[8] How social media is shaping political campaigns. https://knowledge.wharton.upenn.edu/article/how-social-media-is-shaping-political-campaigns/. Accessed August 2022.

[9] In hong kong, a proxy battle over internet freedom begins. https://www.nytimes.com/2020/07/07/business/hong-kong-security-law-tech.html. Accessed August 2022.

[10] Privacy of no concern for ted cruz mobile app in campaign's massive data mining operation. http://www.allgov.com/news/top-stories/privacy-of-no-concern-for-ted-cruz-mobile-app-in-campaigns-massive-data-mining-operation-160212?news=858277. Accessed August 2022.

[11] Tech companies grapple with hong kong's new security law. https://www.datacenterdynamics.com/en/news/tech-companies-grapple-hong-kongs-new-security-law/. Accessed August 2022.

[12] Threat analysis apis. https://www.apivoid.com/. Accessed August 2022.

[13] Virustotal. https://www.virustotal.com/gui/home/upload. Accessed August 2022.

[14] Alexandra Parker. Personal data leaked from fulton county, according to election officials. https://www.atlantanewsfirst.com/2022/09/23/personal-data-leaked-fulton-county-according-election-officials/. Accessed Nov 2022.

[15] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *Proceedings of the USENIX Security Symposium*, 2019.

[16] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 985–1002. USENIX Association, August 2020.

[17] Ballotpedia. Special elections to the 116th united states congress (2019-2020). https://ballotpedia.org/Special_elections_to_the_116th_United_States_Congress_(2019-2020). Accessed August 2022.

[18] Ballotpedia. State legislators who have switched political party affiliation. https://ballotpedia.org/State_legislators_who_have_switched_political_party_affiliation. Accessed May 2023.

[19] Ballotpedia. What should you have on your campaign website? https://ballotpedia.org/Ballotpedia:About. Accessed August 2022.

[20] Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. A Theory of Vagueness and Privacy Risk Perception. In *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, 2016.

[21] Jasmine Bowers, Bradley Reaves, Imani N. Sherman, Patrick Traynor, and Kevin Butler. Regulators, Mount Up! Analysis of Privacy Policies for Mobile Money Services. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[22] Jasmine Bowers, Imani N Sherman, Kevin Butler, and Patrick Traynor. Characterizing Security and Privacy Practices in Emerging Digital Credit Applications. In *Proceedings of the ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 2019.

[23] Allison J Brown. "should i stay or should i leave?": Exploring (dis) continued facebook use after the cambridge analytica scandal. *Social media+ society*, 6(1):2056305120913884, 2020.

[24] Onur Catakoglu, Marco Balduzzi, and Davide Balzarotti. Automatic extraction of indicators of compromise for web applications. In *Proceedings of the 25th international conference on world wide web*, pages 333–343, 2016.

[25] CBS4 News. How political campaigns are able to text you with personal information. https://cbs4indy.com/news/how-political-campaigns-are-able-to-text-you-with-personal-information/. Accessed May 2023.

[26] Cisco. Building consumer confidence through transparency and control. https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-cybersecurity-series-2021-cps.pdf?CCID=cc000742&DTID=odicdc000016&OID=rptsc027438. Accessed May 2023.

[27] clemfromspace. scrapy-selenium. https://github.com/clemfromspace/scrapy-selenium. Accessed August 2022.

[28] CNET. Google fined $57 million under new european data privacy law. https://www.cnet.com/tech/tech-industry/google-fined-57-million-under-european-privacy-law/. Accessed May 2023.

[29] Congressional Research Service. Franking privilege: Historical development and options for change. https://crsreports.congress.gov/product/pdf/RL/RL34274/20. Accessed May 2023.

[30] Joan L Conners. Social media use in us senate campaigns: Initial tactics with twitter. *Social Media and Politics: A New Way to Participate in the Political Process [2 volumes]*, page 129, 2016.

[31] Sunny Consolvo, Patrick Kelley, Tara Matthews, Kurt Thomas, Lee Dunn, and Elie Bursztein. "why wouldn't someone think of democracy as a target?": Security practices & challenges of people involved with u.s. political campaigns. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, August 2021.

[32] CPO Magazine. Lessons learned from gdpr fines in 2023. https://www.cpomagazine.com/data-protection/lessons-learned-from-gdpr-fines-in-2023/. Accessed May 2023.

[33] Lorrie Faith Cranor, Pedro Giovanni Leon, and Blase Ur. A Large-Scale Evaluation of US Financial Institutions' Standardized Privacy Notices. *ACM Transactions on the Web*, 2016.

[34] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the ISOC Network and Distributed Systems Symposium (NDSS)*, 2018.

[35] Katharine Dommett. Regulating digital campaigning: the need for precision in calls for transparency. *Policy & Internet*, 12(4):432–449, 2020.

[36] EasyList. Easylist filter. https://easylist.to/. Accessed January 2022.

[37] EFF. Voter privacy: What you need to know about your digital trail during the 2016 election. https://www.eff.org/deeplinks/2016/02/voter-privacy-what-you-need-know-about-your-digital-trail-during-2016-election. Accessed May 2023.

[38] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)*, 32(2):1–29, 2014.

[39] Encompass. The effect of gdpr on the political industry. https://encompass-europe.com/comment/the-effect-of-gdpr-on-the-political-industry. Accessed May 2023.

[40] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1388–1401, 2016.

[41] FEC. Browse data. https://www.fec.gov/data/browse-data/?tab=candidates. Accessed August 2022.

[42] FEC. Recording receipts. https://www.fec.gov/help-candidates-and-committees/keeping-records/recording-receipts/. Accessed August 2022.

[43] FEC. Registering as a candidate. https://www.fec.gov/help-candidates-and-committees/registering-candidate/. Accessed August 2022.

[44] Center for Democracy and Technology. Election cybersecurity 101 field guides. https://cdt.org/area-of-focus/cybersecurity-standards/election-security/. Accessed August 2022.

[45] Belfer Center for Science and International Affairs. Cybersecurity campaign playbook. https://www.belfercenter.org/publication/cybersecurity-campaign-playbook. Accessed August 2022.

[46] Gallup. Partisan differences growing on a number of issues. https://news.gallup.com/opinion/polling-matters/215210/partisan-differences-growing-number-issues.aspx. Accessed Nov 2022.

[47] Manish Gaurav, Amit Srivastava, Anoop Kumar, and Scott Miller. Leveraging candidate popularity on twitter to predict election outcome. SNAKDD '13, New York, NY, USA, 2013. Association for Computing Machinery.

[48] GDPR EU. What are the 7 main principles of gdpr? https://www.gdpreu.org/7-main-data-protection-principles-under-gdpr/. Accessed May 2023.

[49] GDPR-info. Art. 9 gdpr - processing of special categories of personal data. https://gdpr-info.eu/art-9-gdpr/. Accessed May 2023.

[50] Roberto J. González. Hacking the citizenry?: Personality profiling, 'big data' and the election of donald trump. *Anthropology Today*, 33(3):9–12, 2017.

[51] Government Technology. Voter data modeling: Does it threaten our privacy? https://www.govtech.com/data/voter-data-modeling-does-it-threaten-our-privacy.html. Accessed May 2023.

[52] Catherine Han, Irwin Reyes, Álvaro Feal, Joel Reardon, Primal Wijesekera, Narseo Vallina-Rodriguez, Amit Elazari, Kenneth A. Bamberger, and Serge Egelman. The Price is (Not) Right: Comparing Privacy in Free and Paid Apps. In *Proceedings on Privacy Enhancing Technologies (PETS)*, 2020.

[53] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, Baltimore, MD, August 2018. USENIX Association.

[54] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *Proceedings of the USENIX Security Symposium*, 2018.

[55] Joanne Hinds, Emma J Williams, and Adam N Joinson. "it wouldn't happen to me": Privacy concerns and perspectives following the cambridge analytica scandal. *International Journal of Human-Computer Studies*, 143:102498, 2020.

[56] Geng Hong, Zhemin Yang, Sen Yang, Lei Zhang, Yuhong Nan, Zhibo Zhang, Min Yang, Yuan Zhang, Zhiyun Qian, and Haixin Duan. How you get shot in the back: A systematical study about cryptojacking in the real world. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1701–1713, 2018.

[57] Christopher Hunter. Political privacy and online politics: How e-campaigning threatens voter privacy. *First Monday*, 7(2), Feb. 2002.

[58] iapp. Closing in on the us election with voter privacy and election security. https://iapp.org/news/a/closing-in-on-the-u-s-election-with-voter-privacy-and-election-security/. Accessed May 2023.

[59] Information Commissioner's Office. Guidance for the use of poersonal data in political campaigning. https://ico.org.uk/for-organisations/direct-marketing-and-privacy-and-electronic-communications/guidance-for-the-use-of-personal-data-in-political-campaigning-1/. Accessed May 2023.

[60] Election Cybersecurity Initiative. Usc protect elections - our candidate is democracy. https://www.electionsecurity.usc.edu. Accessed August 2022.

[61] Internet Society. Online trust alliance (ota). https://www.internetsociety.org/ota/. Accessed May 2023.

[62] Internet Society. Online trust audit – 2020 u.s. presidential campaigns. https://www.internetsociety.org/resources/ota/2019/online-trust-audit-2020-u-s-presidential-campaigns/. Accessed Nov 2022.

[63] Jim Isaak and Mina J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[64] Carlos Jensen and Colin Potts. Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2004.

[65] Daniel Kreiss and Philip N Howard. New challenges to political privacy: Lessons from the first us presidential race in the web 2.0 era. *International Journal of Communication*, 4:19, 2010.

[66] Li Li, Daoyuan Li, Tegawendé F Bissyandé, Jacques Klein, Yves Le Traon, David Lo, and Lorenzo Cavallaro. Understanding android app piggybacking: A systematic study of malicious code grafting. *IEEE Transactions on Information Forensics and Security*, 12(6):1269–1284, 2017.

[67] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–64, 2020.

[68] Tariq Mahmood, Tasmiyah Iqbal, Farnaz Amin, Wajeeta Lohanna, and Atika Mustafa. Mining twitter big data to predict 2013 pakistan election winner. In *INMIC*, pages 49–54. IEEE, 2013.

[69] Masterclass. A guide to the primary election cycle in the united states. https://www.masterclass.com/articles/a-guide-to-the-primary-election-cycle-in-the-united-states\#how-do-primary-elections-work. Accessed August 2022.

[70] Arunesh Mathur, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M Stewart, and Arvind Narayanan. Manipulative tactics are the norm in political emails. 2022.

[71] Matomo. Google analytics privacy issues: Is it really that bad? https://matomo.org/blog/2022/06/google-analytics-privacy-issues/. Accessed Nov 2022.

[72] Aleecia M. McDonald and Lorrie Faith Cranor. The Cost of Reading Privacy Policies. *I/S Journal of Law and Policy for the Information Society (ISJLP)*, 4, 2008.

[73] McKinsey & Company. The consumer-data opportunity and the privacy imperative. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative. Accessed May 2023.

[74] National Archives. Code of federal regulations. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102. Accessed Nov 2022.

[75] Federal Bureau of Investigation. Protected voices. https://www.fbi.gov/investigate/counterintelligence/foreign-influence/protected-voices. Accessed August 2022.

[76] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On the Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. In *Workshop on Technology and Consumer Protection (ConPro)*, 2019.

[77] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. {TESSERACT}: Eliminating experimental bias in malware classification across space and time. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 729–746, 2019.

[78] Andrew Perrin. Americans are changing their relationship with facebook. *Pew Research Center*, 5, 2018.

[79] Pew Research. Voters rarely switch parties, but recent shifts further educational, racial divergence. https://www.pewresearch.org/politics/2020/08/04/voters-rarely-switch-parties-but-recent-shifts-further-educational-racial-divergence/. Accessed May 2023.

[80] Callum Pilton, Shamal Faily, and Jane Henriksen-Bulmer. Evaluating privacy-determining user privacy expectations on the web. *computers & security*, 105:102241, 2021.

[81] Andrew Podob, Benjamin W Campbell, and Janet M Box-Steffensmeier. Collaboration among congressional campaigns: The sharing of donor and supporter information. In *Political Networks Workshops & Conference*, 2018.

[82] Politico. Inside the 2016 black market for donor emails. https://www.politico.com/story/2015/12/inside-the-2016-black-market-for-donor-emails-216761. Accessed May 2023.

[83] Polityzer. Polityzer data and code. https://github.com/polityzer/polityzer. Accessed January 2022.

[84] Proton Blog. Political campaigns and your personal data. https://proton.me/blog/political-campaigns-and-your-personal-data. Accessed May 2023.

[85] Yu Pu and Jens Grossklags. Valuating {Friends'} privacy: Does anonymity of sharing personal data matter? In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 339–355, 2017.

[86] Pulitzer Center. Consumer data privacy in politics. https://pulitzercenter.org/stories/consumer-data-privacy-politics. Accessed May 2023.

[87] Emilee Rader. Awareness of behavioral tracking and information privacy concern in facebook and google. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 51–67, 2014.

[88] Reuters. How political campaigns use your data. https://graphics.reuters.com/USA-ELECTION/DATA-VISUAL/yxmvjjgojvr/. Accessed August 2022.

[89] Ira Rubinstein. Voter privacy in the age of big data. *SSRN Electronic Journal*, 2014, 01 2014.

[90] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. Identifying the Provision of Choices in Privacy Policy Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[91] Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. The cambridge analytica affair and internet-mediated research. *EMBO reports*, 19(8):e46579, 2018.

[92] Prabhsimran Singh, Yogesh K Dwivedi, Karanjeet Singh Kahlon, Annie Pathania, and Ravinder Singh Sawhney. Can twitter analytics predict election outcome? an insight from 2017 punjab assembly elections. *Government Information Quarterly*, 37(2):101444, 2020.

[93] SlickText. One year after cambridge analytica, survey reveals strong consumer privacy fears remain. https://www.slicktext.com/blog/2019/02/survey-consumer-privacy-fears-after-cambridge-analytica/. Accessed April 2022.

[94] David Sounthiraraj, Justin Sahs, Garret Greenwood, Zhiqiang Lin, and Latifur Khan. Smv-hunter: Large scale, automated detection of ssl/tls man-in-the-middle vulnerabilities in android apps. In *In Proceedings of the 21st Annual Network and Distributed System Security Symposium (NDSS'14*, 2014.

[95] Speakeasy Political. Political digital fact of the day: The importance of your campaign's website. https://www.speakeasypolitical.com/12-facts-political-campaign-website/. Accessed May 2023.

[96] Tactical Tech. Personal data: Political persuasion. https://cdn.ttc.io/s/tacticaltech.org/Personal-Data-Political-Persuasion-How-it-works_print-friendly.pdf. Accessed Nov 2022.

[97] TechCrunch. Donor data is the new currency of presidential candidates. https://techcrunch.com/2016/01/18/donor-data-is-the-new-currency-of-presidential-candidates/. Accessed May 2023.

[98] The Guardian. 'i made steve bannon's psychological warfare tool': meet the data war whistleblower. https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump. Accessed May 2023.

[99] The National Archives. Data protection act 1998. https://www.legislation.gov.uk/ukpga/1998/29. Accessed May 2023.

[100] The New York Times. Ted cruz's new data strategy: Here comes santa claus. https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/12/17/ted-cruzs-new-data-strategy-here-comes-santa-claus/?_r=0. Accessed May 2023.

[101] UPROXX. How ted cruz's app is taking privacy invasion to all new levels. https://uproxx.com/technology/ted-cruz-app/. Accessed May 2023.

[102] U.S. Congress. S. 2398 - a bill to amend the federal election campaign act of 1971 to ensure privacy with respect to voter information. https://www.congress.gov/bill/116th-congress/senate-bill/2398/text. Accessed May 2023.

[103] Haoyu Wang, Zhe Liu, Jingyue Liang, Narseo Vallina-Rodriguez, Yao Guo, Li Li, Juan Tapiador, Jingcun Cao, and Guoai Xu. Beyond google play: A large-scale comparative study of chinese android app markets. In *Proceedings of the Internet Measurement Conference 2018*, pages 293–307, 2018.

[104] WatchGuard. Https content inspection. https://www.watchguard.com/wgrd-solutions/security-topics/https-inspection. Accessed April 2022.

[105] Web Tribunal. 21 ssl statistics that show why security matters so much. https://webtribunal.net/blog/ssl-stats/. Accessed April 2022.

[106] Paul Webb and Tim Bale. Shopping for a better deal? party switching among grassroots members in britain. *Journal of Elections, Public Opinion and Parties*, 33(2):247–257, 2023.

[107] Craig E Wills and Can Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 13–18, 2012.

[108] Le Yu, Xiapu Luo, Xule Liu, and Tao Zhang. Can We Trust the Privacy Policies of Android Apps? In *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2016.

[109] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. Automated Analysis of Privacy Requirements for Mobile Apps. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)*, 2017.

Appendix

1. Threats to Validity

We list the threats to validity of this study, as follows:

Completeness of website collection and email study: Our dataset does not include candidates whose Ballotpedia profile was not found or who did not have a campaign website link in their Ballotpedia profile. Similarly, our dataset does not include Political Action Committees (PACs) and Super-PACs, as our primary focus was on the privacy practices of campaigns. Further, we started the crawling from September 15, 2020, which was after the primaries, so we may have "lost" some data that could otherwise have been collected.

Completeness of email study: The email study encompasses emails received between November 5, 2020 to September 14, 2022, and only considers the top donation-earners. For a more exhaustive analysis of the sharing practices of campaigns, we refer the readers to the study by Podob et al. [81] which includes all major campaigns in the 2016 election over a full election cycle.

Completeness of data type collection: For compiling the data types collected by each campaign website, we automatically extract all the forms in all the webpages of the website and extract the "labels" in those forms. In doing so, we may miss cases where input fields are not bounded by proper `<form>` tags in the webpages or cases where input fields that are present within forms are not properly labeled. Additionally, one author manually resolved each of the extracted labels based on the limited context provided by the label texts. The author discarded any labels that could not be reasonably resolved (*e.g.*, label texts such as input_1), including the labels meant for automated crawlers (*e.g.*, 'leave this box empty'). Hence, our methodology of data type collection (and the resultant findings) offers a *lower bound* on the data collected by the campaign websites.

Using VirusTotal: VirusTotal has been widely used as the ground truth for malware classification by prior work [66] [24] [56] [77] [103]. That said, VirusTotal's engines are not without flaws, particularly false positives due to over-approximation, and our findings obtained through them ($\mathcal{F}_{12}$ and $\mathcal{F}_{15}$) should be interpreted as *potential* problems in this context. Further, to counteract the infeasibility of validating the findings of VirusTotal's scanners, prior work generally uses a threshold for trusting the VirusTotal labels [66] [24]. Therefore, like prior work [66], we consider a threshold of 1 engine, but report the number of engines that label a website as malicious or phishing in our findings related to VirusTotal (Section 10), to allow an informed interpretation of our findings, and convey the potential for false positives.

2. Use of Privacy Policy Templates

We found that 316/507 (62.33%) campaigns include data types in their privacy policy that they do not collect in practice, as seen in Table 6. Some unique data types that fall under this category are tax ID number, religion, phone contact list, and mobile device ID number. The presence of such *surplus* datatypes can be explained in a number of ways. First, the privacy policy of both the website and the mobile app of the campaign (if present) could be the same, which means the data may still be collected, just not via the website. Second, the campaign may intend to collect such data in the future or is using a template privacy policy without removing such extra datatypes. Use of policy templates among campaign websites is likely, based on our comparison of privacy policy texts, that we discuss next. Finally, this could also be due to a gap in our analysis, as we only analyze the form labels in webpages, and may miss form inputs if the form is improperly labeled in the html.

TABLE 6. EXTRA DATA MENTIONED IN THE PRIVACY POLICY BUT NOT COLLECTED IN THE WEBSITE.

Dataset	Extra in priv.policy	
house_active	155 (58.05%)	(total=267)
house_inactive	58 (64.44%)	(total=90)
senate_active	36 (64.29%)	(total=56)
senate_incumbents	36 (73.47%)	(total=49)
senate_inactive	12 (63.16%)	(total=19)
president_active	3 (75%)	(total=4)
president_inactive	16 (72.73%)	(total=22)
total	316 (62.33%)	(total=507)

To gauge the use of privacy policy templates across different campaigns, we compared the text of each privacy policy with the rest of the privacy policies in our dataset. We do this by first converting the privacy policy text into TF-IDF vectors and calculating their cosine similarity with each other. For high likelihood of similarity, we only consider policies with over 98% of cosine similarity score as similar. Finally, we randomly choose 5 privacy policies that have at least 1 similar privacy policy to assess their overall privacy implication.

***Finding A1* – 239 privacy policies have at least one highly similar corresponding policy. ($\mathcal{F}_{A1}$).** The cluster with the highest number of similar privacy policies is 23, while the cluster with the least number is 2. In the random sampling of 5 clusters we manually analyze, the main part of the policy text were identical, with the only textual difference being in the introductory paragraph. Due to this, the data types described within the privacy policies were also identical.

***Finding A2* – Campaigns with similar privacy policy did not have similar data collection practices. ($\mathcal{F}_{A2}$).** In our analysis of 5 randomly chosen similar privacy policy clusters of sizes 20, 8, 7, 4 and 2 respectively, we found no similarity in the types of data they collect, despite the fact that the disclosure about the data types being the same. This likely indicates that the privacy policies were simply written from a template (*e.g.*, by replacing the name of the campaign and the candidate) rather than as a way to rigorously inform the user about the privacy practices *corresponding* to that campaign website.

Findings $\mathcal{F}_{A1} \rightarrow \mathcal{F}_{A2}$ further show the need of implementing a rigorous policy standard in this domain, as we discussed in Section 12.3.

TABLE 7. TOP 10 MOST COLLECTED DATA TYPES AMONG CAMPAIGN WEBSITES

Datatype	**count (total=1446)**
email_address	1431 (98.96%)
name	1304 (90.18%)
location_coarse	918 (63.49%)
phone_number	903 (62.45%)
location_fine	535 (37.0%)
password	245 (16.94%)
employer_info	174 (12.03%)
occupation	167 (11.55%)
username	150 (10.37%)
website	134 (9.27%)

TABLE 8. TOTAL PRIVACY POLICY EXTRACTED FOR ANALYSIS PER DATASET

Dataset	**No. of priv.policy extracted**
house_active	267
house_inactive	90
senate_active	56
senate_incumbents	49
senate_inactive	19
president_active	4
president_inactive	22
total	507

TABLE 9. CAMPAIGNS' COLLECTION OF POLITICAL OPINIONS WITH PII IN THE SAME PAGE

Dataset	**Collect political opinion with PII**	**No privacy policy**
house_active	34 (3.6%)	21 (61.76%)
house_inactive	12 (1.69%)	9 (75%)
senate_active	4 (3.57%)	0 (0%)
senate_incumbents	2 (2.94%)	1 (50%)
senate_inactive	5 (3.31%)	2 (40%)
president_active	2 (50%)	0 (0%)
president_inactive	2 (3.17%)	0 (0%)
total	61 (2.96%)	33 (54.10%)

TABLE 10. CAMPAIGN WEBSITES THAT DISCLOSE SHARING DATA WITH OTHER CAMPAIGNS

Dataset	**Share with other campaigns**
house_active	91/267 (34.08%)
house_inactive	23/90 (25.56%)
senate_active	23/56 (41.07%)
senate_incumbents	20/49 (40.82%)
senate_inactive	5/19 (26.32%)
president_active	2/4 (50.0%)
president_inactive	15/22 (68.18%)
total	179/507 (35.31%)

TABLE 11. NO. OF UNSAFE CAMPAIGN WEBSITES ACROSS DATASETS

Dataset	**No. of unsafe sites**
house_active	11/952 (1.16%)
house_inactive	4/710 (0.56%)
senate_active	0/112 (0%)
senate_incumbents	0/68 (0%)
senate_inactive	0/151 (0%)
president_active	0/4 (0%)
president_inactive	2/63 (3.17%)
total	17/2060 (0.83%)

TABLE 12. NO. OF NON-TLS WEBSITES

Dataset	**non-TLS sites**	**Collect PII**
house_active	66/952 (6.93%)	44/66 (66.67%)
house_inactive	65/710 (9.15%)	28/65 (43.08%)
senate_active	12/112 (10.71%)	6/12 (50%)
senate_incumbents	2/68 (2.94%)	2/2 (100%)
senate_inactive	16/151 (10.60%)	2/16 (12.5%)
president_active	0/4 (0%)	0
president_inactive	7/63 (11.11%)	4/7 (57.14%)
total	168/2060 (8.16%)	86/168 (51.19%)

TABLE 13. NON-US COUNTRIES WHERE WEBSITES ARE HOSTED

Country	**Num of campaigns**
Canada	11
Germany	11
Australia	5
Japan	3
France	2
VietNam	1
UK	1
Lithuania	1
HongKong	1
Czechia	1
Denmark	1

TABLE 14. NO. OF CAMPAIGN WEBSITES WITH TRACKERS

Dataset	**w/ trackers**	**w/o priv.policy**
house_active	741/952 (77.84%)	467/741 (63.02%)
house_inactive	457/710 (64.37%)	361/457 (78.99%)
senate_active	90/112 (80.36%)	34/90 (37.78%)
senate_inactive	100/151 (66.23%)	74/100 (74%)
senate_incumbents	67/68 (98.53%)	15/67 (22.39%)
president_active	4/4 (100%)	0 (0%)
president_inactive	45/63 (71.43%)	23/45 (51.11%)
total	1504/2060 (73.01%)	974/1504 (64.76%)

TABLE 15. TOP 10 MOST COMMONLY UNDISCLOSED DATATYPES

Datatype	**count** (total=507)
phone_number	111 (21.89%)
email_address	88 (17.36%)
location_coarse	80 (15.78%)
name	71 (14.01%)
location_fine	47 (9.27%)
password	34 (6.71%)
employer_information	29 (5.72%)
occupation	24 (4.73%)
website	14 (2.76%)
credit_card_info	13 (2.56%)

3. Meta-Review

Summary of Paper:

This paper studies the privacy practices of the websites of the candidates for elected office in the 2020 U.S. elections. The authors developed a measurement infrastructure that analyzes the privacy policies, data collection practices, and third party trackers present on candidates' websites.

Scientific Contributions:

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a Valuable Step Forward in an Established Field
- Establishes a New Research Direction

Reasons for Acceptance:

1. The paper confirms important results in the area of data use by political campaigns, from a computer science and measurement-based approach. The paper finds evidence for sharing of mailing list data between campaigns, overcollection of data for targeting purposes, and incomplete privacy disclosures.
2. The paper provides a step forward in the field of auditing website privacy practices. The authors created new tools and built on existing tools like Polisis for extracting privacy policies from websites, and performed extensive validation to ensure that they transferred to a new domain (political campaign websites).
3. The paper establishes a new research direction in the privacy practices of political campaigns. The authors identify that political campaigns are generally exempt from privacy regulations like GDPR, but many engage in practices that users would likely find problematic. It provides motivation for further studies of how political campaigns handle user data, and for policies to protect the privacy of voters, donors, and users of their sites.

Noteworthy concerns:

Reviewers were concerned about whether there is evidence that users and voters are concerned about political campaign websites' privacy practices. While the authors acknowledge that no empirical data exists as to whether voters/donors have privacy concerns, they do present data from related fields and relate recent election-related events (e.g., Cambridge Analytica scandal) that support the idea that if voters/donors were to learn what data is being collected, and how it is being used, such concerns would very likely arise. The introduction for the paper provides a strong starting point for future work on this topic.