Privacy Analysis: Network Requests on Civic Websites

Project submission by: Tin Le
Supervised by Joel Reardon

CPSC502.04 Final Report
Faculty of Science
University of Calgary
April 12, 2022

Abstract – **Internet tracking has become the norm in the current internet landscape. New techniques and the involvement of third parties in user data collection have caused tracking to be more pervasive and widespread than ever, all while often remaining hidden from users. These privacy-infringing practices are now used by campaign websites for political parties and candidates, as well as web services that employees, students, and educators must nowadays work with. This research project and paper accordingly discovers and analyzes the tracking performed by numerous websites, including: campaign websites for the 2021 Canadian Federal Election, campaign websites for the 2021 Calgary General Election, and web services used by those at the University of Calgary community, such as D2L and Microsoft Outlook. In this paper, we review work and literature similar and relevant to our research. We then discuss our methodology and metrics to determine the extent of tracking on our websites of interest, followed by our results and findings. Finally, we propose future work to be done in this field that could further our understanding of websites' privacy practices.**

I. INTRODUCTION

Every website on the internet makes use of Hypertext Transfer Protocol (HTTP) requests in order to transfer data between users and the website's servers [1]. This includes the site's content, images loaded from other places, and scripts to dynamically control its execution. While some requests are critical to properly load websites and provide services to the user, other requests may covertly collect personal information about the user's visit to the website for advertisement, tracking, and other purposes. These data collection requests often occur unbeknownst to the user as they interact with the website. The collected information may then be sent back to the first party web servers, or sent instead to a number of third party companies that track users online.

In order for voters to understand the platforms of various candidates and political parties, it has become increasingly necessary for them to visit the candidates' and parties' websites. It is likely in the interests of political entities to gather information about site visitors. However, users often have little awareness about the data that the websites they visit collect, since most users do not generally check their browser's network requests or read privacy policies.

In addition, employees doing their job in this day and age must visit a variety of online services regardless of the privacy risk they may face. For example, teachers and also students at the University of Calgary must use D2L, among other services, to fulfill their job and course requirements.

This project therefore investigates and measures the privacy compromises that one may encounter when using civic and other necessary websites relevant to the University of Calgary community. Specifically, we investigate 3 distinct groups of websites:

- Group 1: Consists of the campaign websites for the six major Canadian federal parties in the 2021 Canadian Federal Election:
  - Liberal Party
  - Bloc Quebecois
  - Conservative Party
  - Green Party
  - New Democratic Party
  - People's Party of Canada
- Group 2: Consists of the campaign websites of most candidates with websites in the 2021 Calgary General Election, spanning across all four positions:
  - Mayor
  - Councillor
  - Public School Board Trustee
  - Separate School Board Trustee
- Group 3: Consists of several websites relevant to students and employees at the University of Calgary:
  - d2l.ucalgary.ca
  - outlook.office.com
  - portal.my.ucalgary.ca
  - gradescope.ca
  - ucalgary.ca

Our research was accomplished by capturing and analyzing HTTP requests and network traffic across a variety of different types of websites. Various programs were also developed and used in order to facilitate the data collection and analysis processes.

There are many reasons for which a website may choose to implement tracking. Companies may collect data on their websites' customers to tailor and improve their service and products [2]. User data can be sold to 3rd party advertising companies, who use the collected information to target users with specific ads [2]. Modern tracking has evolved to a point where companies can even create profiles for individual users; browser fingerprinting is a technique that combines collected details of a user's device with their online activity to track them across the internet [2]. These practices clearly pose a strong risk to our right to privacy on the internet, and it is therefore important to understand the manner in which our information is collected and handled by the many websites that pertain to our community.

Section II of this paper discusses related works done for this topic. We then discuss the methodology used in our research to obtain our data in section III. Sections IV, V, and VI examine the results of our investigations into Groups 1, 2, and 3 respectively. We discuss possible future work for this field in section VII, and we finally conclude the paper in section VIII.

II. RELATED WORKS

There exist many works and resources that discuss the Hypertext Transfer Protocol (HTTP), the protocol that is central to this project. Regardless, a general understanding of how HTTP works will suffice: clients send HTTP requests to web servers, and web servers send HTTP responses back to clients [3]. There exist various HTTP request commands, called HTTP methods, each of which do different things.

For example, HTTP GET requests ask for resources from the server, while HTTP POST requests are designed to send client data to the server's databases [3]. HTTP messages may also include headers, which can include supporting information such as the client's browser and operating system. It should be noted that the Internet Engineering Task Force has developed HTTP/2 and HTTP/3—newer versions of the protocol [4][5]. Nevertheless, the basic concepts of the original HTTP still remain, and the aforementioned concepts offer adequate knowledge to read the HAR (HTTP archive) file format—a file format commonly used to store HTTP network traffic data.

The primary goal of analyzing network traffic in this project is to identify signs that a website is compromising users' privacy or tracking them. Gourley et al. discuss the various ways in which web services collect and use personal information [6]. Collection methods range widely, from basic well-known techniques such as storing cookies, to complex ones such as operating system instance fingerprinting to gain access to a user's webcam. A considerable portion of methods are related to HTTP, which is relevant to our study. Examples include HTTP cookies, scanning of HTTP headers, and HTTP caching. While Gourley et al.'s paper [6] offers a broad survey of the various privacy threats that users may face and ample statistics and case examples to back up its claims, it lacks specific test results and analyses of specific categories of websites. Our project hones in on the privacy practices of certain sets of websites, including election campaign web pages and sites that are a daily necessity to those in the Calgary community.

The usage of software to analyze a large number of websites' privacy practices is not new. Mayer and Mitchell present results generated by FourthParty, a web measurement technology which they developed [7]. The software comes in the form of a Mozilla Firefox extension that automatically generates a logging database, which stores information regarding dynamic web content. Users can then find specific data in the results using basic SQL queries. Notable findings from previous FourthParty testing include:

- OkCupid—a dating website—sending data to a third party about how often users drink, smoke, and do drugs;
- Epic Marketplace—an advertising network—publicly exposing its interest segment data in 2011, which included "menopause, getting pregnant, repairing bad credit, and debt relief" (Mayer and Mitchell 415);
- anonymizeIp: In a crawl of the Alexa top 10,000 global websites, an opt-in feature to anonymize user IP addresses was only activated on 1.3% of reports to Google Analytics.

While Mayer and Mitchell's focus in their paper [7] seemed to be on covering the privacy practices of the most popular websites on the Internet, this project instead focuses on certain groups of websites.

Another example of a software designed to measure privacy on the Internet is Jensen et al.'s iWatch [8], a Java-developed web crawler "designed to catalogue and analyze online data practices and the use of privacy related indicators and technologies." (Jensen et al. 29). Given a starting set of URLs to initially visit, iWatch will automatically download the pages it lands on while finding other links to visit in the downloaded pages. During this process, the web crawler searches for certain HTTP tokens and other signs that may indicate any unusual data-handling or collection techniques. In their paper [8], they used iWatch to crawl a total of 240,340 web pages in 81 countries. Unlike Mayer and Mitchell's work however, Jensen et al. do not provide any particular examples of sites with poor privacy practices; instead, they include tables of statistics about the percentages of sites in the crawl that use tools such as cookies, pop ups, and banners. While this work is highly impressive and wide-reaching, the facts that their paper was published in 2007 and the crawls were actually performed in 2005–2006 mean that the information provided is outdated and may no longer apply to today's internet landscape. For instance, iWatch checked if websites used the Privacy Preferences Project (P3P), a disused protocol that allowed websites to declare and communicate their privacy policies to users. Some of the concepts mentioned in their paper, such as cookies and popups, still remain in use to this day. Despite that, the number and complexity of tracking and other privacy-compromising techniques have since skyrocketed, and therefore their data is less applicable nowadays.

Federal and provincial political parties often "fall between the cracks" of Canada's privacy regulations [9]. Bennett and Bayley's 2012 report [9] notes that the Personal Information Protection and Electronic Documents Act (PIPEDA), arguably Canada's most important data privacy law, does not cover political parties. Other laws that parties are exempted from include an anti email-spam act[1] and the Telecommunications Act's "Do not Call List". Out of the ten provinces, only British Columbia had privacy legislation comparable to PIPEDA that also covered political parties. Since the report was published in 2012, there has been some improvement in Canada's privacy legislation. In 2018, the Canadian Parliament enacted Bill C-76, or the Elections Modernization Act, which required parties to ensure that their privacy policies protected personal information, as well as send their policies to Elections Canada for review [10]. The Office of the Privacy Commissioner of Canada's "Guidance for federal political parties on protecting personal information", published in 2019, is designed "to assist federal political parties in complying with their new legal obligations relating to privacy policies" [11]. Nevertheless, our project analyzes the actual information collection practices of federal parties, in addition to provincial candidates.

Finally, similar work to this project has been done on

[1] An Act to promote the efficiency and adaptability of the Canadian economy by regulating certain activities that discourage reliance on electronic means of carrying out commercial activities, and to amend the Canadian Radio-television and Telecommunications Commission Act, the Competition Act, the Personal Information Protection and Electronic Documents Act and the Telecommunications Act (S.C. 2010, c. 23), https://laws-lois.justice.gc.ca/eng/AnnualStatutes/2010_23/page-1.html

the 2020 United States elections. The International Digital Accountability Council's (IDAC) "Privacy Issues in 2020 U.S. Campaigns' Apps & Websites: An IDAC Investigation & Recommended Best Practices" report [12] provides findings and data from several election/campaign mobile apps and websites. Mobile testing was done by manual and automated interaction with the apps, after which network traffic would be analyzed. This type of testing revealed that both the official Trump and Biden apps sent sensitive user data to multiple third parties, some of which was done covertly. For example, the Trump app sent geolocation to a third party called Phunware, but did not disclose that it would share location data with other parties in its privacy policy. More analogous to this project was the website testing, which was performed on 132 campaign sites using a website behavior-analysis platform called Netograph. The IDAC study kept track of how many third parties and third party cookies were used by each webpage. Results from testing showed that both Trump and Biden's official campaign websites used upwards of 20 third parties, including some ad networks. Unfortunately, the website analysis section of the report is not very detailed, which is what this project aims to improve upon by providing more specific information regarding the privacy practices of analyzed websites.

III. METHODOLOGY

As mentioned previously, our research investigates three separate groups of websites of interest:

- Group 1: Consists of the campaign websites for the six major Canadian federal parties in the 2021 Canadian Federal Election:
  - Liberal Party
  - Bloc Quebecois
  - Conservative Party
  - Green Party
  - New Democratic Party
  - People's Party of Canada
- Group 2: Consists of the campaign websites of most candidates with websites in the 2021 Calgary General Election, spanning across all four positions:
  - Mayor
  - Councillor
  - Public School Board Trustee
  - Separate School Board Trustee
- Group 3: Consists of several websites relevant to students and employees at the University of Calgary:
  - d2l.ucalgary.ca
  - outlook.office.com
  - portal.my.ucalgary.ca
  - gradescope.ca
  - ucalgary.ca

[2]https://github.com/duckduckgo/tracker-radar

For each website of interest, our first step in our methodology was to visit the website on a browser and download the network traffic as a HAR file. This is done on both Firefox and Google Chrome using the "Network" section of the browser's web development tools. While this could have been done completely manually, two Python browser automation scripts—one controlling the Firefox browser and one controlling the Google Chrome browser—were created to repeat this process automatically for a large number of websites at a time. This allows us to easily obtain a large data set of website behaviour. Our browser automation scripts also clear the history and cookies after visiting each website, in order to ensure the clean slate environment.

Once the HAR files containing the requests were downloaded, the next step involved analyzing the data. Since the amount of data is excessive to analyze completely manually, this project also included the creation of several helper Python programs that sift through the HAR files and:

- mark noteworthy and relevant HTTP requests, such as to trackers and other analytics companies,
- provide summaries of the data and tracking performed,
- output statistics into CSV files for importing into spreadsheets.

The DuckDuckGo Tracker Radar[2], a database consisting of common third party domains on the web, was used to help detect tracking. The database maintains a JSON file for each common third party domain, which stores additional information such as its prevalence on the internet, tracking methods, and a list of rules that requests to the domain follow. For this project, we consider any domain found in this database as a tracking domain. Additionally, an HTTP request is marked as a potential tracking request if:

1. the destination domain can be found in the tracker database, and,
2. its URL (web address) contains or matches one of the rules found in the file for that domain in the database.


```
"resources": [
    {
        "rule": "moatads\\.com\\/addthismoatframe568911941483\\/moatframe\\.js",
        "cookies": 0,
        "fingerprinting": 0,
        "subdomains": [
            "z"
        ],
```


Figure 1: Example of a rule in the tracker database

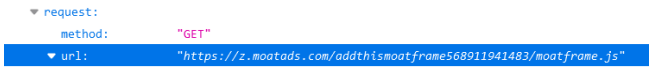


Figure 2: Example of a request that matches the rule above and is therefore marked as a tracking request

The version of the tracker database used for this project was the latest version as of December 23, 2021.

Finally, the results were charted, compared, and contrasted, revealing any trends and patterns. This process provides an overview of the privacy practices used by the organizations studied.

Information, code, and additional details for this project can be found at the following GitHub repository: `https://github.com/tincangit/research-project-502`.

IV. GROUP 1: MAJOR CANADIAN FEDERAL PARTIES

The six parties and their corresponding campaign websites for the 2021 Canadian Federal Elections in this group are:

- Liberal Party of Canada - http://www.liberal.ca/
- Bloc Quebecois - http://www.blocquebecois.org/
- Conservative Party of Canada - http://www.conservative.ca/
- Green Party of Canada - http://www.greenparty.ca/
- New Democratic Party - http://www.ndp.ca/
- People's Party of Canada - https://www.peoplespartyofcanada.ca/

Two sets of HAR files were manually collected for this group on September 18th, 2021; one using a clean slate Firefox browser, and the other using a clean slate Google Chrome browser. The "Disable cache" option was enabled for both browsers while obtaining data. At this point, the browser automation programs had not been fully developed yet, and were therefore not used to collect these HAR files.

A. Number of HTTP Requests

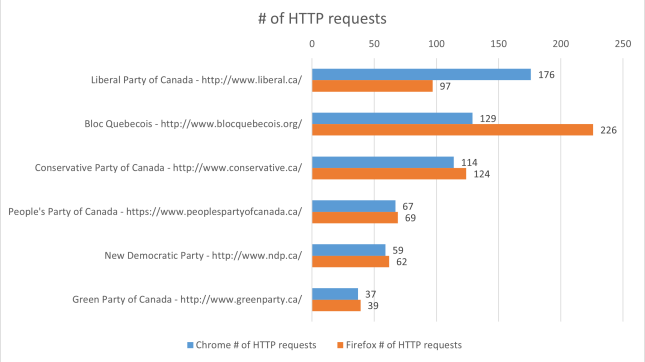
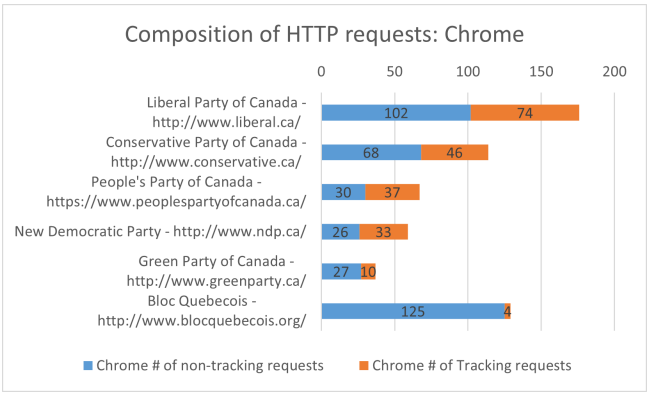


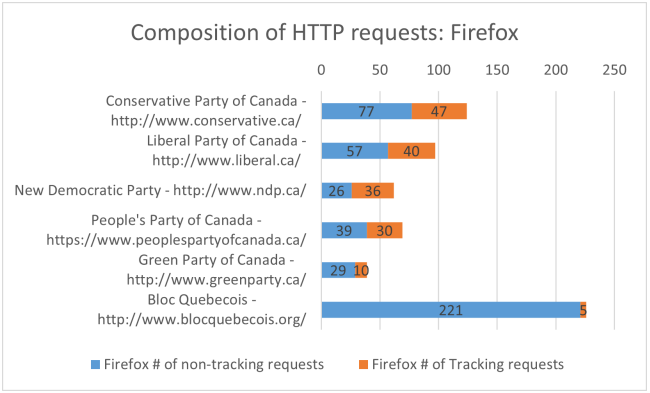
Figure 3: Number of HTTP requests made by party website

The six parties made a total of 582 requests on Google Chrome and 617 requests on Firefox. Figure 3 shows the number of HTTP requests each website made on each of the two browsers. On its own, this statistic does not provide much of an insight into the extent of a website's tracking. A website could make hundreds of requests to obtain images and other benign resources, whereas a single request to a nosy analytics company could reveal significant personal information about a user. It is only when this data is combined with other statistics that privacy practice patterns emerge.

B. Tracking/Non-Tracking HTTP Requests



(a) Chrome



(b) Firefox

Figure 4: Number of tracking vs non-tracking requests made by party website

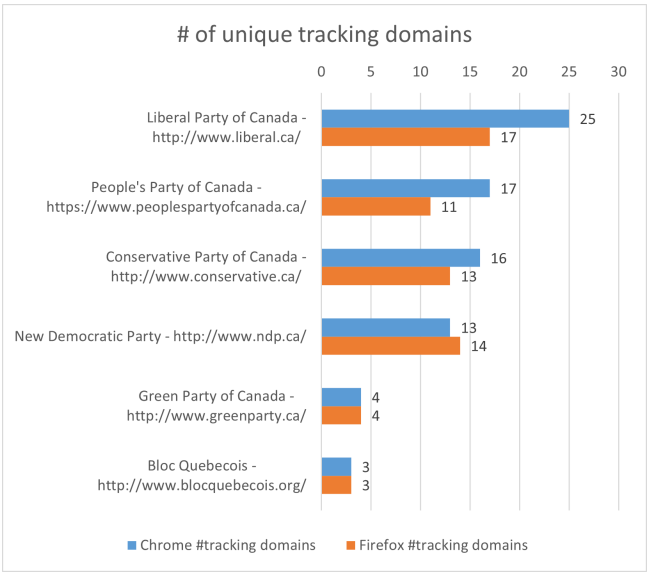
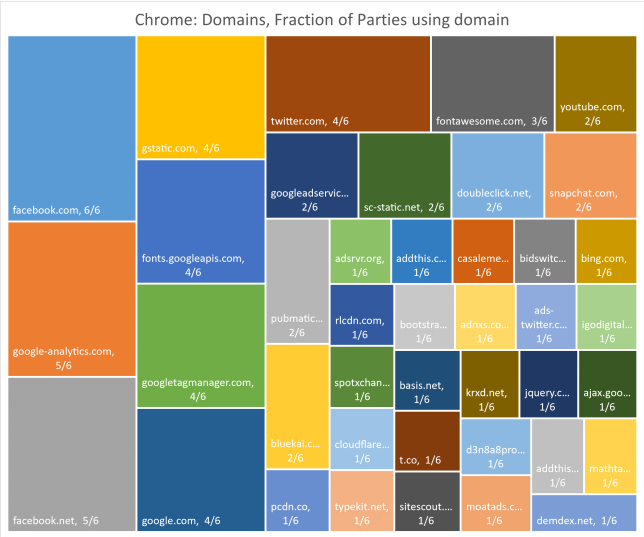
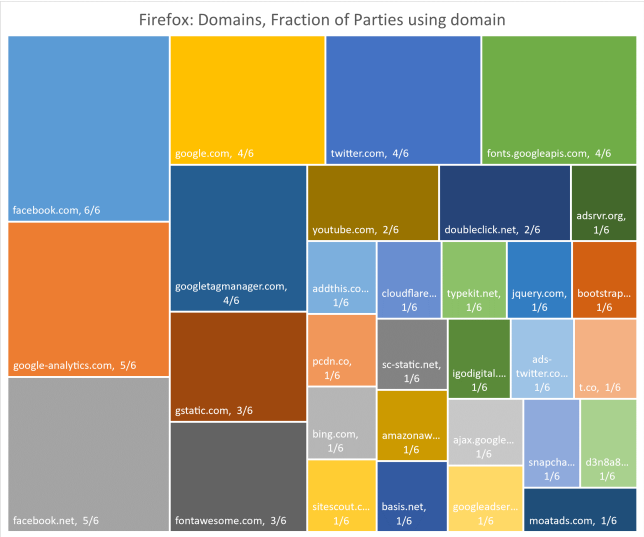


Figure 5: Number of unique tracking domains by party website



(a) Chrome



(b) Firefox

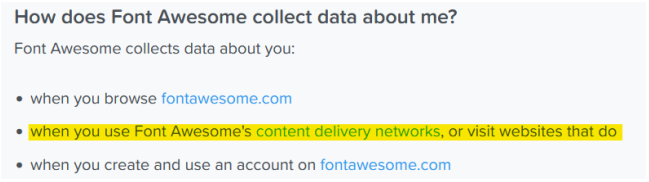
Figure 6: 3rd party domains and fraction of parties using the domain

The six campaign websites from Group 1 made a total of 204 tracking requests on Google Chrome versus 168 tracking requests on Firefox. Those requests interacted with 43 unique tracking domains across the two browser data sets.

Figure 4 divides requests into 2 categories: non-tracking and tracking. A request is considered a tracking request if its URL matches one of the rules found in the DuckDuckGo Tracker Radar. A list of matched rules can be found in the project's GitHub repository. Figure 5 shows the number of unique tracking domains found for each party; that is, the number of unique domains each party makes a request to.

Figure 6 illustrates all of the tracking domains found and the fraction of the six parties that use each domain. The more popular a tracker is, the larger its corresponding rectangle in the treemap.

Some of these domains may seem benign, such as the font services (fonts.googleapis.com, fontawesome.com, etc.) and the social media domains, which likely are necessary for certain functions on the websites. However, these domains' companies' privacy policies and the HTTP data suggest that requests to these domains still collect some user information, even if that is not the primary purpose. Therefore we consider these domains to be trackers; even if organizations don't intend for tracking to happen, it may still occur.



How does Font Awesome collect data about me?

Font Awesome collects data about you:

- when you browse fontawesome.com
- when you use Font Awesome's content delivery networks, or visit websites that do
- when you create and use an account on fontawesome.com

Figure 7: Excerpt from fontawesome.com's privacy policy. Websites (such as the PPC's website) use content delivery networks to implement their Font Awesome's fonts on their websites.

The following list displays each unique tracking domain used by each party across both browsers:

- Liberal Party of Canada - http://www.liberal.ca/
  - ads-twitter.com
  - basis.net
  - bing.com
  - bluekai.com
  - doubleclick.net
  - facebook.com
  - facebook.net
  - fonts.googleapis.com
  - google-analytics.com
  - google.com
  - googleadservices.com
  - googletagmanager.com
  - gstatic.com
  - igodigital.com
  - krxd.net
  - pcdn.co
  - pubmatic.com
  - rlcdn.com
  - sc-static.net
  - sitescout.com
  - snapchat.com
  - spotxchange.com
  - t.co
  - twitter.com
  - youtube.com
- Bloc Quebecois - http://www.blocquebecois.org/
  - cloudflare.com
  - facebook.com
  - fontawesome.com
- Conservative Party of Canada - http://www.conservative.ca/
  - adsrvr.org
  - ajax.googleapis.com
  - bidswitch.net
  - casalemedia.com
  - doubleclick.net
  - facebook.com
  - facebook.net
  - fonts.googleapis.com
  - google-analytics.com
  - google.com
  - googletagmanager.com
  - gstatic.com
  - jquery.com
  - pubmatic.com
  - typekit.net
  - youtube.com
- Green Party of Canada - http://www.greenparty.ca/
  - facebook.com
  - facebook.net
  - google-analytics.com
  - twitter.com
- New Democratic Party - http://www.ndp.ca/
  - amazonaws.com
  - bootstrapcdn.com
  - facebook.com
  - facebook.net
  - fontawesome.com
  - fonts.googleapis.com
  - google-analytics.com
  - google.com
  - googleadservices.com
  - googletagmanager.com
  - gstatic.com
  - sc-static.net
  - snapchat.com
  - twitter.com
- People's Party of Canada - https://www.peoplespartyofcanada.ca/

- addthis.com
- addthisedge.com
- adnxs.com
- bluekai.com
- d3n8a8pro7vhmx.cloudfront.net
- demdex.net
- facebook.com
- facebook.net
- fontawesome.com
- fonts.googleapis.com
- google-analytics.com
- google.com
- googletagmanager.com
- gstatic.com
- mathtag.com
- moatads.com
- twitter.com

Table 1 (on the next page) lists the top 10 most popular tracking domains in the federal election data, as well as the the parties that use each tracker. These domains, which are also highly prevalent on the web, mainly belong to well-known companies such as Facebook and Google.

The data indicates a trend that the Liberal and Conservative webpages generally perform the most tracking requests, while the Bloc Quebecois and Green Party webpages perform the least.

The discovered tracking requests perform their tasks using a variety of methods, including using cookies, inserting tracking pixels, and uploading information to databases. Some of these methods will be explored in further detail in later sections.

C. POST Requests

POST requests are a type of HTTP request that send client data to a server, after which the server usually stores the data in a database [3]. While these requests are normal and often invoked when a user submits information to the website, our methodology consisted of simply visiting websites without any further interaction. Therefore, in this case, POST requests are a potential indicator of user data being gathered and collected by third parties.

POST requests often contain data passed through the URL's query string—a portion of a URL that assigns values to parameters. Data can also be attached in the postData section of the request. The information that is sent to the server can occasionally be deduced by analyzing the query string and the postData section of the request.

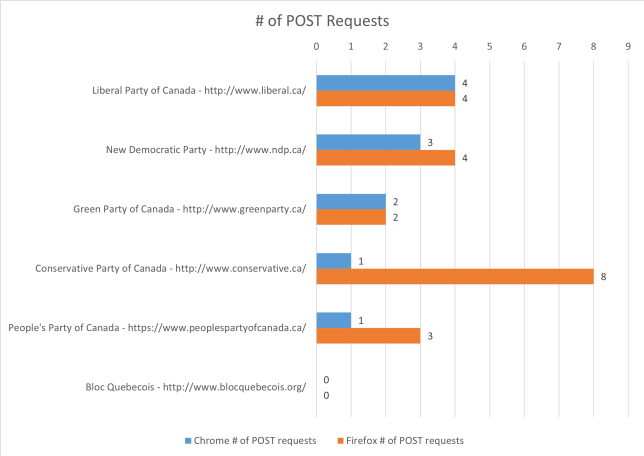


Figure 9: Number of POST requests made by party website

Figure 9 shows how many POST requests were made by each party campaign website on each browser. Table 2 shows the domains that the POST requests made by each party's site contact. A complete list of all the POST request URLs can be found in the GitHub.

Candidate - Website	POST request domains
Liberal Party of Canada - http://www.liberal.ca/	tr.snapchat.com www.facebook.com www.youtube.com bat.bing.com
Bloc Quebecois - http://www.blocquebecois.org/	N/A
Conservative Party of Canada - http://www.conservative.ca/	www.youtube.com www.facebook.com
Green Party of Canada - http://www.greenparty.ca/	www.google-analytics.com
New Democratic Party - http://www.ndp.ca/	www.google-analytics.com tr.snapchat.com
People's Party of Canada - https://www.peoplespartyofcanada.ca/	v1.addthisedge.com m.addthis.com www.google-analytics.com

Table 2: Unique POST request domains across both browsers

D. Tracking Pixels

A tracking pixel is a minuscule image that is stealthily embedded in websites in order to collect data about visitors [13]. When a user visits a webpage with a tracking pixel, their browser downloads the pixel, and in the process provides information to the web servers [13]. This may include details such as their operating system, interactions with the website, and IP address [13].

(a) The pixel (white dot)

```
name:    "User-Agent"
value:   "Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:92.0) Gecko/20100101 Firefox/92.0"
```

(b) Some of the data collected by the tracking pixel

```
▼ content:
    mimeType:    "image/gif"
    size:        43
    encoding:    "base64"
```

(c) HTTP response for the pixel

Figure 10: Example of a tracking pixel

Tracking pixels can be found by searching through HTTP responses for very small images. For this project, any HTTP response with

1. the string "image" in the mimeType (media type) value, and,
2. a content size of between 1 and 99 bytes inclusive

is marked as a tracking pixel.

The main privacy concerns with tracking pixels lie with the fact that they are designed to stay hidden and often collect user data without consent. Additionally, by obtaining the user's IP address, third parties can determine the user's location, and furthermore combine additional data about the user in order to create a detailed profile of the user.

Despite these privacy concerns, tracking pixels are still common on the web. Out of the six federal parties, all parties except for Bloc Quebecois use tracking pixels in their campaign websites. Figure 11 presents the number of tracking pixels in each website's homepage.

domain	parties using this domain
facebook.com	Liberal Bloc Quebecois Conservative Green Party NDP PPC
google-analytics.com	Liberal Conservative Green Party NDP PPC
facebook.net	Liberal Conservative Green Party NDP PPC
gstatic.com	Liberal Conservative NDP PPC
fonts.googleapis.com	Liberal Conservative NDP PPC
googletagmanager.com	Liberal Conservative NDP PPC
google.com	Liberal Conservative NDP PPC
twitter.com	Liberal Green Party NDP PPC
fontawesome.com	Bloc Quebecois NDP PPC
youtube.com	Liberal Conservative

(a) Chrome

domain	candidates using this domain
facebook.com	Liberal Bloc Quebecois Conservative Green Party NDP PPC
google-analytics.com	Liberal Conservative Green Party NDP PPC
facebook.net	Liberal Conservative Green Party NDP PPC
google.com	Liberal Conservative NDP PPC
twitter.com	Liberal Green Party NDP PPC
fonts.googleapis.com	Liberal Conservative NDP PPC
googletagmanager.com	Liberal Conservative NDP PPC
gstatic.com	Liberal Conservative NDP
fontawesome.com	Bloc Quebecois NDP PPC
youtube.com	Liberal Conservative

(b) Firefox

Table 1: Top 10 most used 3rd party domains within Group 1

Figure 8: Excerpt of a sample POST request to youtube.com. The query string contains information about the user's browser and operating system.

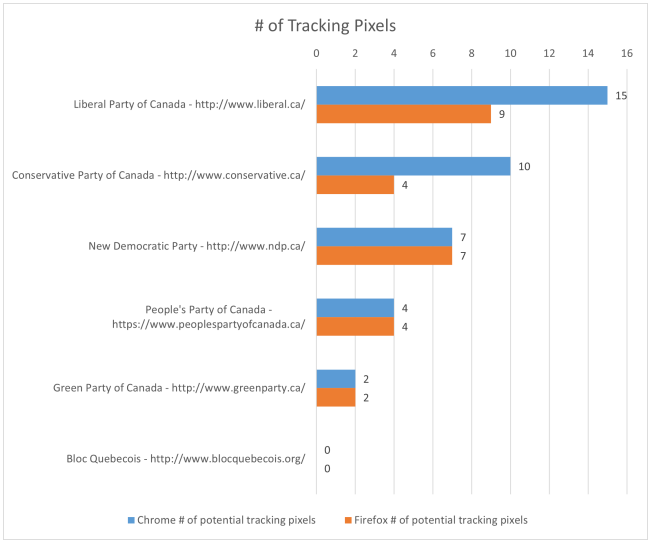


Figure 11: Number of tracking pixel used by party

E. Information Collected

Determining the exact information collected by the trackers can be challenging. Most of the data in the tracking HTTP requests and responses consist of strings and numbers whose meanings and purposes are hard to establish by themselves. Nonetheless, we can determine to varying levels of certainty what actions some of the trackers carry out.

All of the Group 1 websites—except for Bloc Quebecois—make at least one Facebook request that contains a so-called "fr" cookie. According to Facebook, the "fr" cookie is "used to deliver, measure and improve the relevancy of ads, with a lifespan of 90 days" [14]. However, it has been noted that if a user visits their Facebook profile on the same browser, the exact same cookie will be sent to Facebook [15]. This means that Facebook is now able to ascertain that the user (with all their account details) has visited that specific campaign website [15][16]. Facebook is able to accomplish this feat on any website that implements the "fr" cookie re-

quest, and is thus able to follow a Facebook user's surfing behavior across the internet. Snapchat also appears to implement a similar cookie system with the "sc_at" cookie, which according to their privacy policy, is used to identify a visitor across multiple domains" [17].

Figure 12: Facebook's "fr" cookie as seen in a HAR file

Google Analytics is another popular tracker which is used by all Group 1 campaign sites except Bloc Quebecois. Based on evidence from the HAR files and the tracker database, they track user settings such as browser language, timezone, and screen/viewport (the browser window) dimensions. While the latter may seem insignificant, the size of a user's screen and/or viewport can be a very effective way to identify a user. A user's screen size can provide clues as to whether they are using a small laptop or a large desktop PC, for instance. Moreover, if the user's viewport is set to an arbitrary non-default dimension, it becomes even easier to identify the user, because their viewport size will likely be unique when compared to other users who use default viewport sizes [18]. It is for this reason that the Tor Browser, a famous browser designed around privacy, warns users against adjusting or maximizing the browser's window size [19].

Major companies like Google and Facebook often have public resources on their analytics services, making it easier to deduce the information they collect. Smaller companies may have less information available online, which makes the same task harder. However, considering that for this group, the websites were simply visited on a clean slate browser with no additional personal information, the data collected by trackers was likely limited. Further studies that involve more interaction with the websites and pre-existing user data on the browser could explore the true extent to what information these tracking companies collect.

F. Summary

Overall, the data shows a trend of the Liberal and Conservative parties employing the heaviest tracking. In contrast, Bloc Quebecois appears to collect the least data, followed by the Green Party. Widely-known companies such as Google, Facebook, and Twitter show up frequently in tracking requests and domains.

Bloc Quebecois stands out for its exceptionally clean privacy practices—abstaining from the use of tracking pixels and POST requests, while only making 4 or 5 tracking requests. Moreover, none of the tracking requests made by the Bloc Quebecois site seem to have tracking as its main focus; these requests are made to obtain fonts/additional content and connect to social media. On the other hand, all of the five other federal parties' websites make several requests to URLs clearly belonging to analytics, tracking, or advertisement companies. These tracking companies apply an assortment of techniques to quietly gain user data, including cookies, POST requests, and tracking pixels.

V. GROUP 2: CALGARY MUNICIPAL ELECTION CANDIDATES

Group 2 contains the campaign websites of almost every candidate in the 2021 Calgary General Election with an official webpage. The list of websites was compiled from the official Calgary Elections site[3], and encompasses candidates for all four offices: Councillor, Mayor, Public school board trustee, and Separate school board trustee. All candidates with a campaign website listed on their profile in the Calgary Elections list of candidates are included in this group, with a few exceptions.

The numbers of campaign websites/candidates analyzed for this group are:

- 88 Councillor candidates,
- 24 Mayor candidates,
- 29 Public school board trustee candidates, and
- 15 Separate school board trustee candidates

for a total of 156 candidates across the four positions. The HAR files were collected on October 17th and 18th, 2021, using both the Google Chrome and Firefox browser automation programs. The full list of the analyzed candidates and their websites can be found on this project's GitHub.

Due to the large pool of websites in this group, this section will mainly focus on the candidates that perform the most tracking.

A. Councillor Candidates

Out of the four offices, the councillor position had by far the most candidates. We were unable to collect valid HAR files for the 3 following candidates:

- Carla Evers - http://carla-evers.com/
- Joe Magliocca - http://www.ward2.ca/
- Michael Streilein - http://michaelstreilein.ca/

Carla Evers' website was unable to be reached, while the failure to collect HARs for the latter two was caused by URL typos in our project implementation. This occurred when using the Google Chrome and the Firefox browser automation programs. The issue was not discovered until after the election had ended and the campaign websites had been closed. For this reason, these websites are excluded from the 88 sites included in this analysis.

Each tracking domain and rule in the DuckDuckGo database is given a fingerprinting score which "represents the likelihood that they're using browser APIs for fingerprinting" [20]. Four councillor candidates—Aryan Sadat (http://electsadat.com/), Mike LaValley (http://mikelavalley.ca/), Rob Ward (http://robward11.ca/), and Teresa Hargreaves (http://teresaforward12.ca/)—have a specific Cloudflare script request in their websites that the database assigns a fingerprinting score of 3 to—the highest fingerprinting score possible.

[3]https://www.calgary.ca/election/information-for-voters/candidates.html

Figure 13: Example of the Cloudflare request made

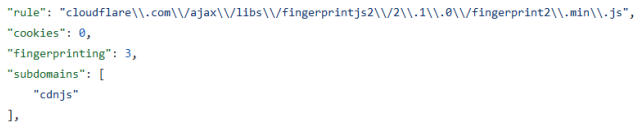


Figure 14: The corresponding rule in the DuckDuckGo database

Cloudflare is an American web performance and security company that also offers web analytics services [21]. The tracker database indicates that the specific Cloudflare request found collects vast data about the user's device, including their operating system and browser, system language, timezone, screen color depth, number of logical processors, and an approximate value for the device's memory size.

Contrary to DuckDuckGo's privacy rating, Cloudflare emphasizes that their web analytics service puts privacy first and doesn't fingerprint individuals [22]. This request may also be related to Cloudflare's distributed denial-of-service (DDoS) services. Further research would be required to determine which is the case.

B. Mayor Candidates

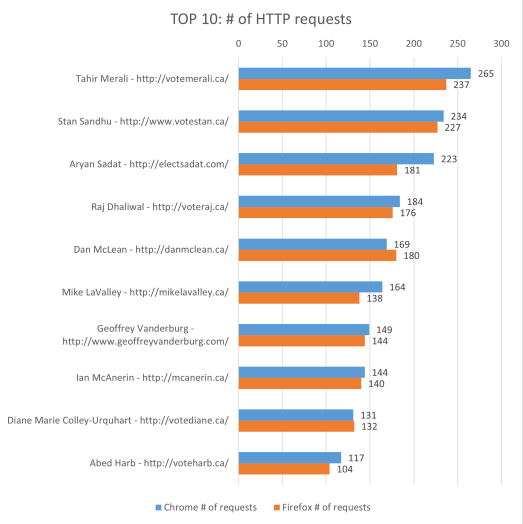
24 mayor candidates and their websites were analyzed. Details of the analysis can be found on page 11.

C. Public School Board Trustee Candidates

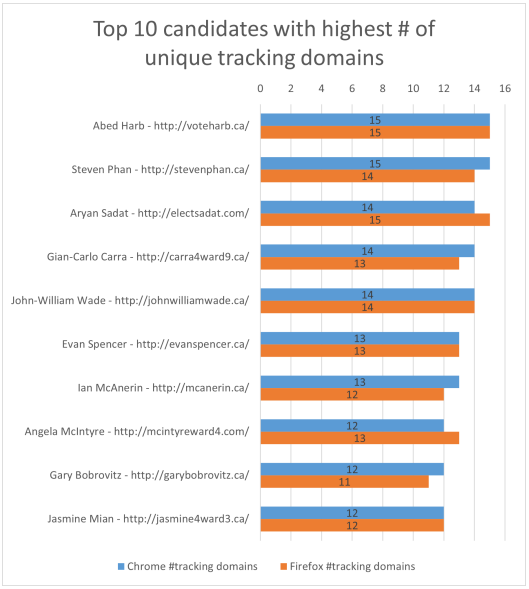
29 public school board trustee candidates and their websites were analyzed. Details of the analysis can be found on page 12. One of the candidates, Susan Vukadinovic (http://susanvukadinovic.com/), also uses the fingerprinting Cloudflare request aforemetioned.

D. Separate School Board Trustee Candidates

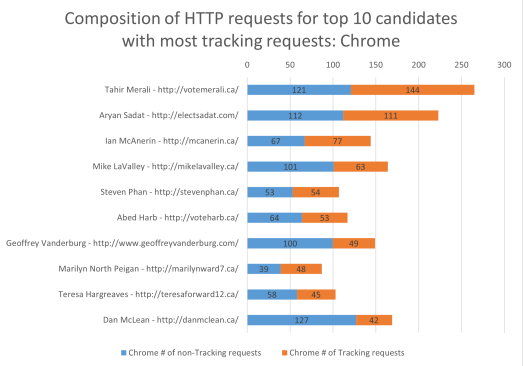
15 public school board trustee candidates and their websites were analyzed. Details of the analysis can be found on page 13.



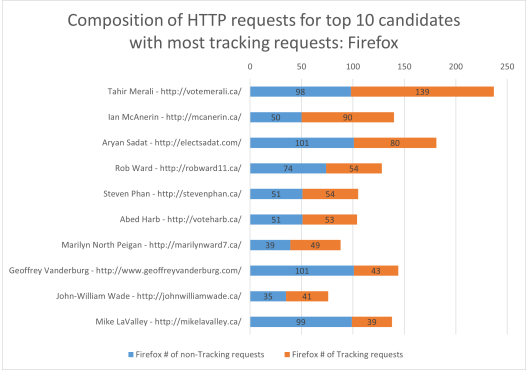
(a) Top 10 councillor candidates: Most HTTP requests



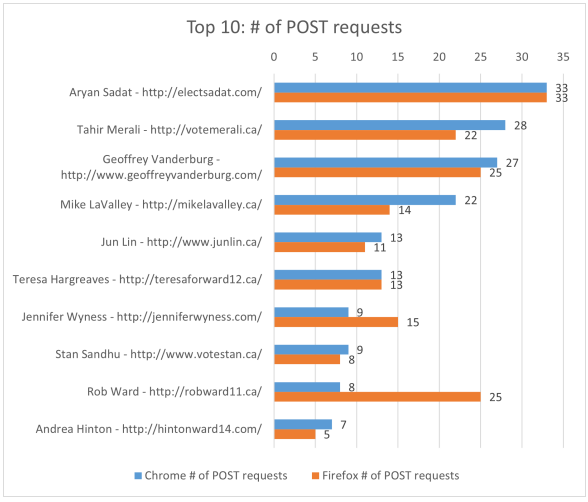
(b) Top 10 councillor candidate websites with the most unique tracking domains



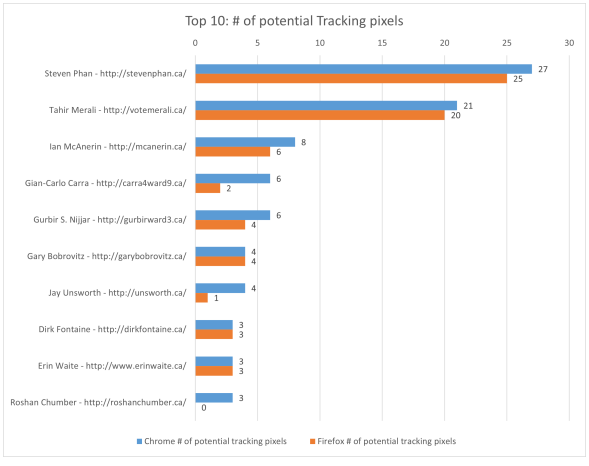
(c) HTTP requests compositions - Top 10 councillor candidates: Most tracking requests (Google Chrome)



(d) HTTP requests compositions - Top 10 councillor candidates: Most tracking requests (Firefox)

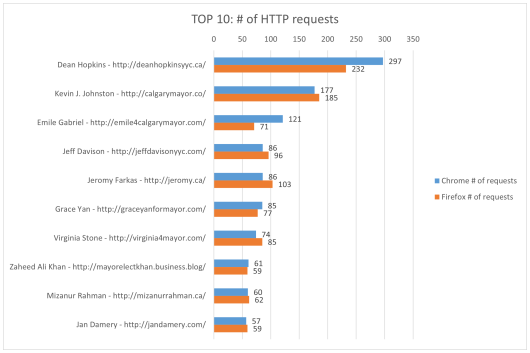


(e) Top 10 councillor candidates with the most POST requests

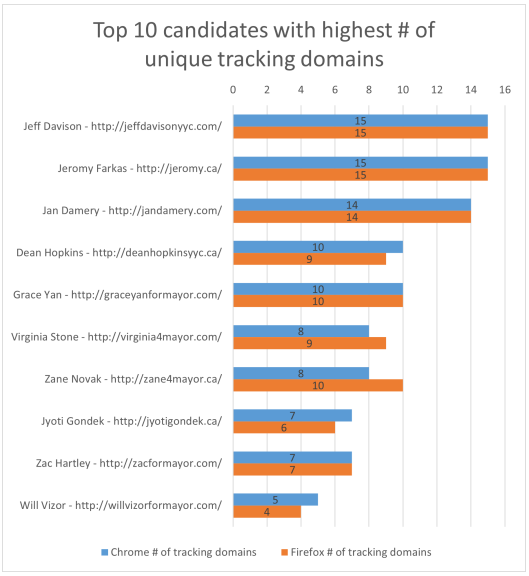


(f) Top 10 councillor candidates with the most potential tracking pixels

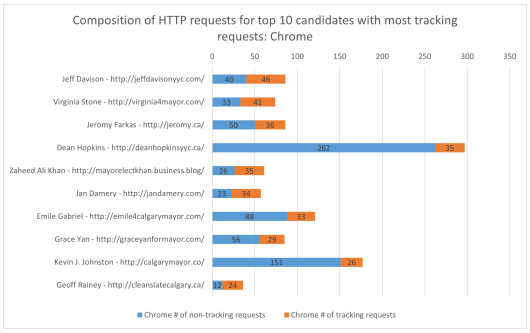
Figure 15: Graphs for councillor candidates and websites



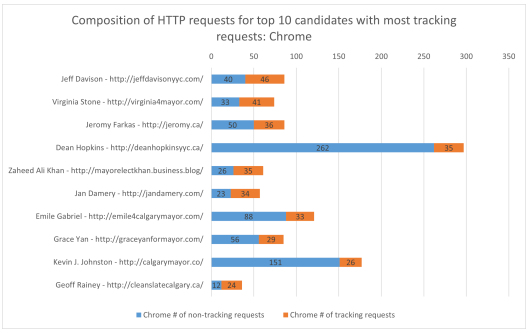
(a) Top 10 mayor candidates: Most HTTP requests



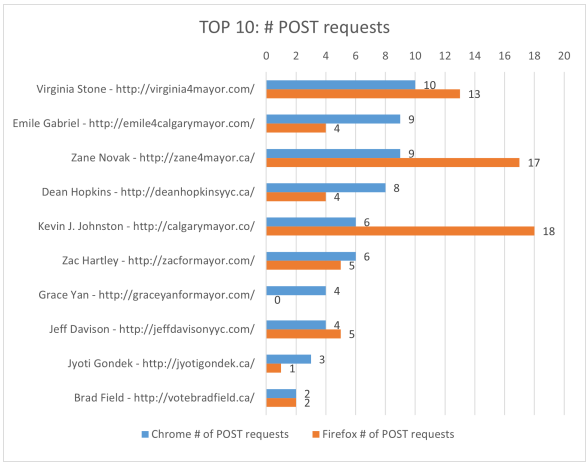
(b) Top 10 mayor candidate websites with the most unique tracking domains



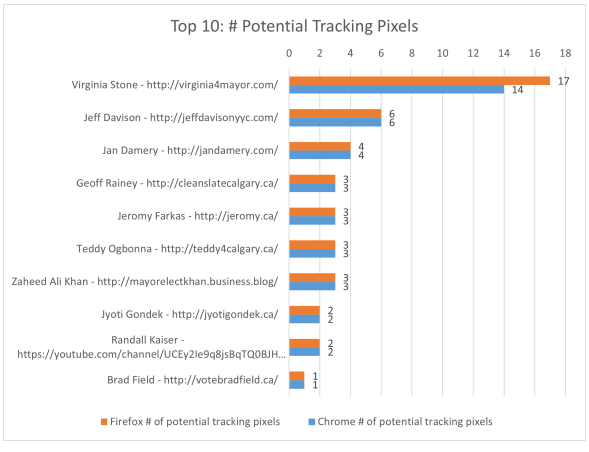
(c) HTTP requests compositions - Top 10 mayor candidates: Most tracking requests (Google Chrome)



(d) HTTP requests compositions - Top 10 mayor candidates: Most tracking requests (Firefox)

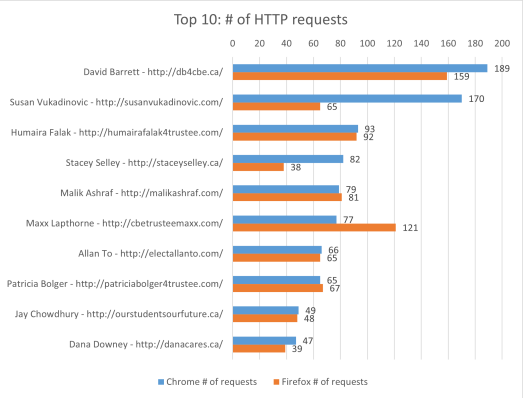


(e) Top 10 mayor candidates with the most POST requests

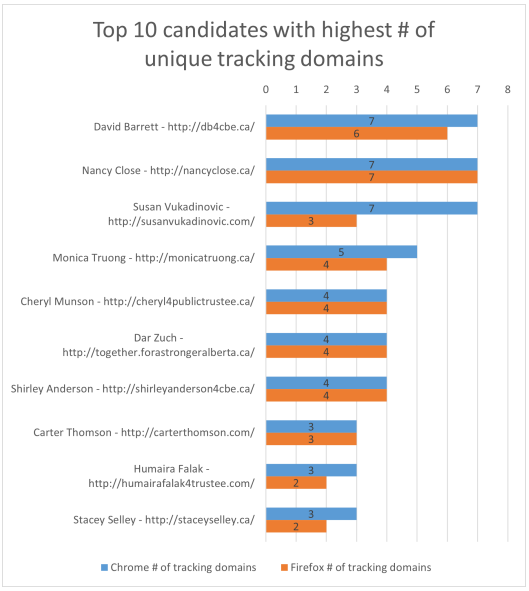


(f) Top 10 mayor candidates with the most potential tracking pixels

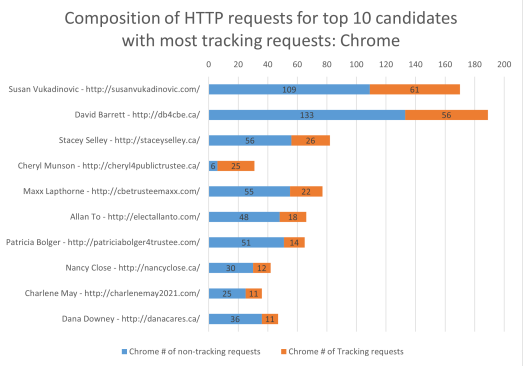
Figure 16: Graphs for mayor candidates and websites



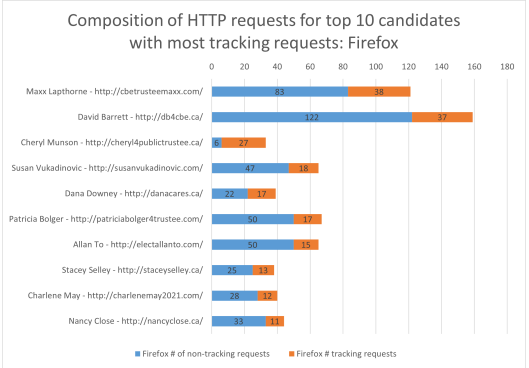
(a) Top 10 public school board trustee candidates: Most HTTP requests



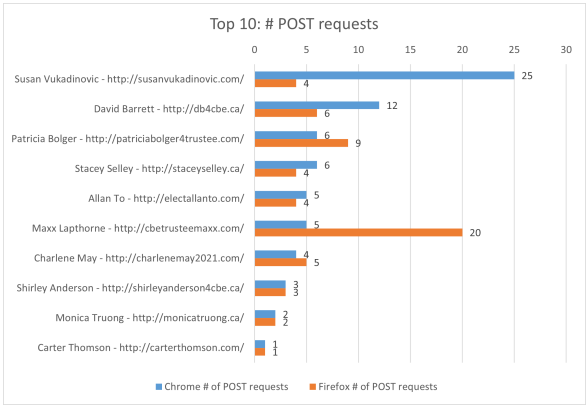
(b) Top 10 public school board trustee candidate websites with the most unique tracking domains



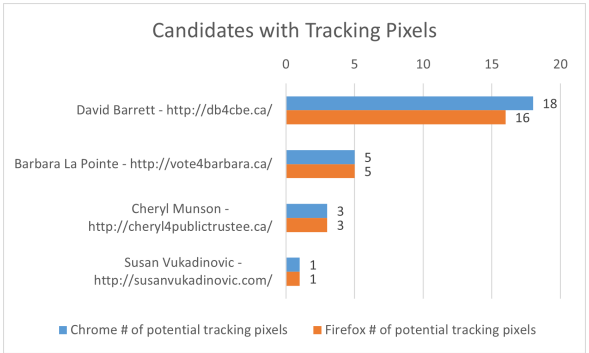
(c) HTTP requests compositions - Top 10 public school board trustee candidates: Most tracking requests (Google Chrome)



(d) HTTP requests compositions - Top 10 public school board trustee candidates: Most tracking requests (Firefox)

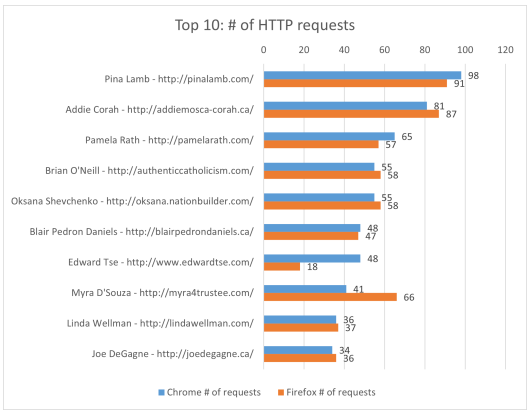


(e) Top 10 public school board trustee candidates with the most POST requests

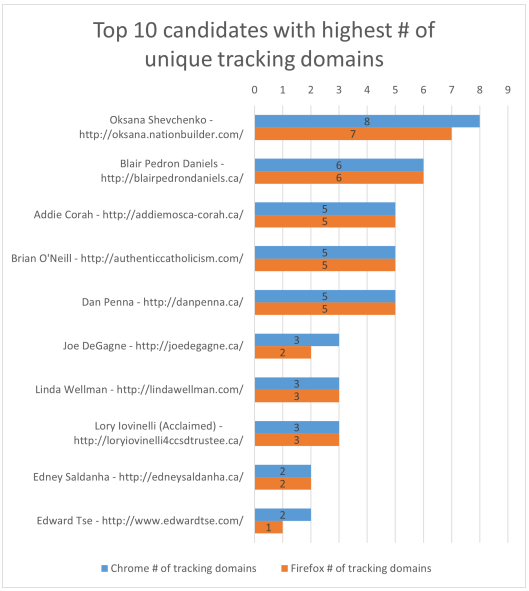


(f) Top 10 public school board trustee candidates with the most potential tracking pixels

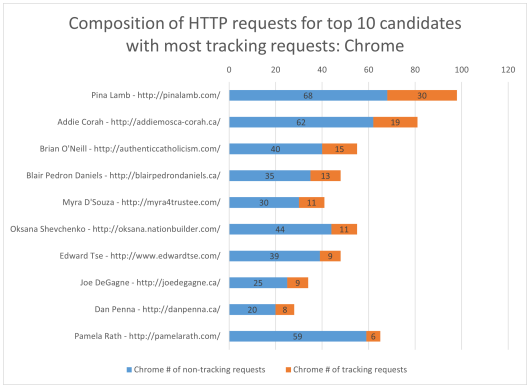
Figure 17: Graphs for public school board trustee candidates and websites



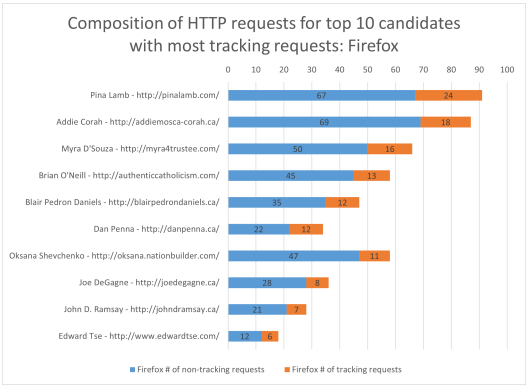
(a) Top 10 separate school board trustee candidates: Most HTTP requests



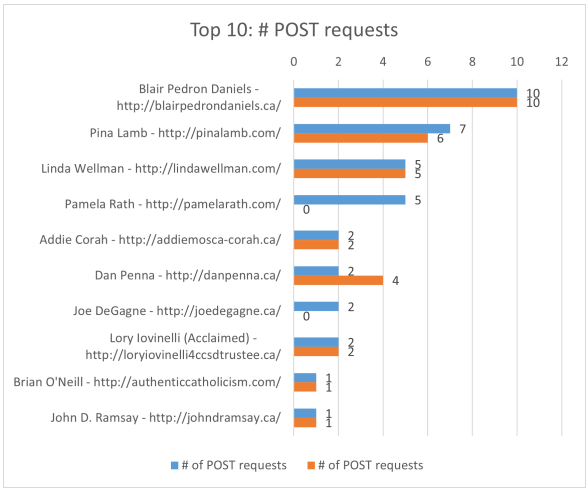
(b) Top 10 separate school board trustee candidate websites with the most unique tracking domains



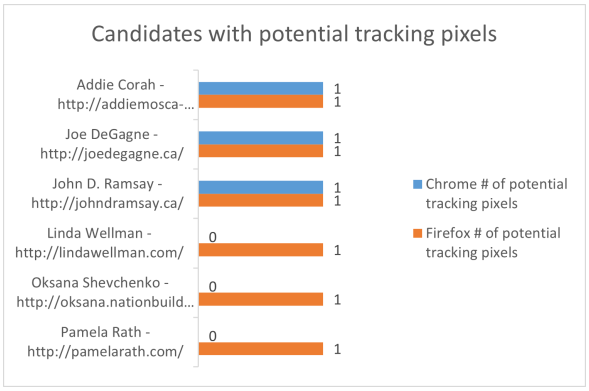
(c) HTTP requests compositions - Top 10 separate school board trustee candidates: Most tracking requests (Google Chrome)



(d) HTTP requests compositions - Top 10 separate school board trustee candidates: Most tracking requests (Firefox)



(e) Top 10 separate school board trustee candidates with the most POST requests



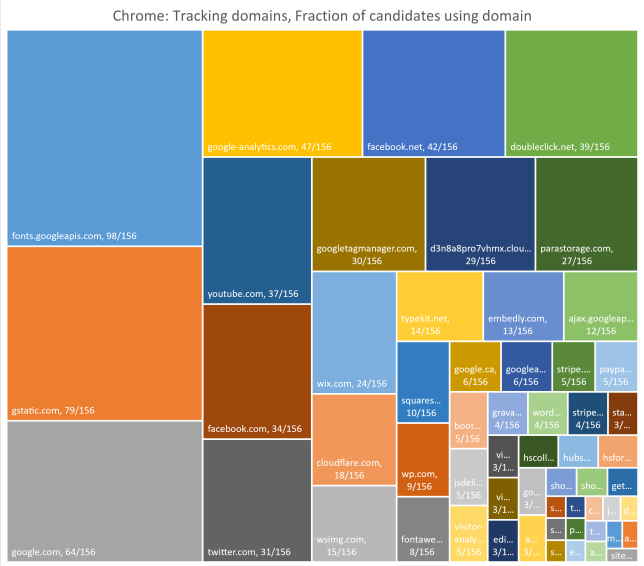
(f) Top 10 separate school board trustee candidates with the most potential tracking pixels

Figure 18: Graphs for separate school board trustee candidates and websites

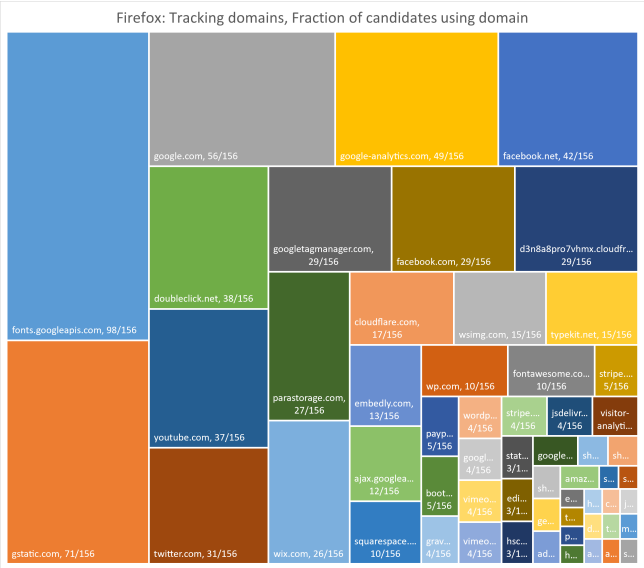
E. Overview

Office Position	Average # of Tracking Requests	Average # of domains
Councillor	19.78	5.81
Mayor	18.81	5.48
Public School Board Trustee	10.90	2.45
Separate School Board Trustee	9.90	3.27
Group 1 (Federal Parties)	31.00	11.67

Table 3: Average number of tracking requests and tracking domains by candidates in each office position (for both Google Chrome and Firefox)



(a) Google Chrome



(b) Firefox

Figure 19: 3rd party domains and fraction of candidates for all 4 positions using the domain

Table 3 shows the average number of tracking requests and domains for each position in each office position, as well as for the 6 party websites in Group 1. Websites for the two school board trustee positions seem to perform less tracking overall compared to the councillor and the mayor candidate websites. The data indicates that campaign sites for the Calgary election track users significantly less than the 6 federal parties in Group 1. However, this statistic may be affected by the much smaller sample size of federal parties. Additionally, the heavier tracking from certain parties (such as Liberal) outshadow the parties that use less tracking.

The treemaps in Figure 19 illustrate all tracking domains used by candidates from all for office positions. The names of the lesser used domains are not fully visible in the graphs; those can be found in the GitHub repository. Much like Group 1, domains from Google and Facebook are prevalent. However, unlike the federal parties, website building services—such as Wix and Squarespace—are also used in Group 2. These companies may track website visitors in addition to providing their primary service. Considering that these candidates likely have substantially less funding for their campaign pages compared to the federal parties, it is possible that a sizable amount of tracking in this section is unintended by the candidates.

VI. GROUP 3: UNIVERSITY OF CALGARY WEBSITES

The goal of this part of the project was to understand the privacy compromises and tracking that a student or staff member at the University of Calgary (UofC) must encounter in their day-to-day work. The services and their corresponding websites analyzed for this group are:

- D2L - d2l.ucalgary.ca
- Outlook - outlook.office.com
- UofC Portal - portal.my.ucalgary.ca
- Gradescope - gradescope.ca
- University of Calgary - ucalgary.ca

Two sets of HAR files were manually collected for this group on April 11, 2022; one using a clean slate Firefox browser, and the other using a clean slate Google Chrome browser. Unlike the other groups, the manual collection involved a short interaction session with the websites, such as clicking buttons and navigating between pages. The exact same sequence of interactions were applied for both Google Chrome and Firefox. The ”Disable cache” and ”Persist logs” options were enabled for both browsers while obtaining data. Since the interactions required a user log-in, the HAR files for Group 3 are not available on the GitHub repository for privacy reasons.

Due to the nature of these websites, their analysis must be treated differently that that of the two previous groups. For example, the sending of data to web servers, such as POST requests, is to be expected for some of the services’ functionalities. Additionally, since each of the websites have a different purpose and will thus require different resources, it would not make sense to compare the HAR files directly against each other. Therefore, each website in Group 3 will be analyzed separately in this section.

A. D2L - d2l.ucalgary.ca

D2L is a learning and teaching platform used by students and teachers at UofC. For this project, data was collected

from the student version of the service. The sequence of interactions with this website consisted of:

1. logging into D2L
2. checking the notifications bar
3. visiting a class page
4. viewing a PDF file

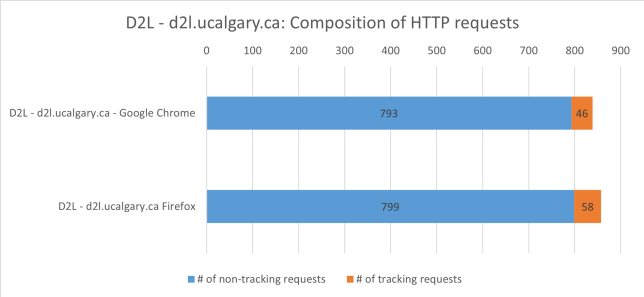


Figure 20: Composition of HTTP requests

D2L makes requests to 2 external domains: brightspace.com and es.io. Not much information can be found about the technical details of these domains, however the percentage of requests shows that D2L is heavily reliant on the former, while the latter's subdomain (d2l-elasticrum.apm.us-east-1.aws.cloud) suggests that it may somehow be related to Amazon's cloud computing services. No tracking pixels were found.

Based on the HARs, there is likely no third party tracking on D2L. However, Brightspace is known to allow instructors to see where students have clicked within a course—a practice known as click tracking [23]. Since requests must be made with the servers every time a student interacts with the website, it is probable that instructors and administrators have a record of students' interactions on D2L.

B. Outlook - outlook.office.com

Microsoft's Outlook is the default email application used by UofC members. The sequence of interactions with this website for the project include:

1. logging into Outlook/UofC account
2. completing 2-step verification
3. opening a previously received email
4. sending a test email to the user's own account

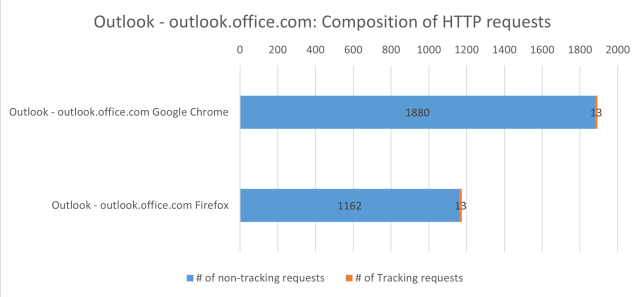


Figure 21: Composition of HTTP requests

The Outlook HAR files are the largest in Group 3. Some of the requests to the microsoft.com are marked as tracking by the DuckDuckGo database; those with URLs starting with

- `browser.pipe.aria.microsoft.com`
- `browser.events.data.microsoft.com`

The marked tracking requests are all POST requests, and seem to report a lot of the user mail settings, as well as generic info such as browser, timezone, and device details. All of the data seems to go through Microsoft's control, so although the information likely doesn't go through third parties, it is hard to determine what Microsoft does with the info. Additionally 15 and 12 tracking pixels were found in the Google Chrome and Firefox HARs respectively.

C. UofC Portal - portal.my.ucalgary.ca

The UofC Portal is used by members of UofC to handle personal administrative tasks. The sequence of interactions with this website for the project include:

1. logging into UofC account
2. viewing the "exams/grades" section
3. viewing the "my financials" section

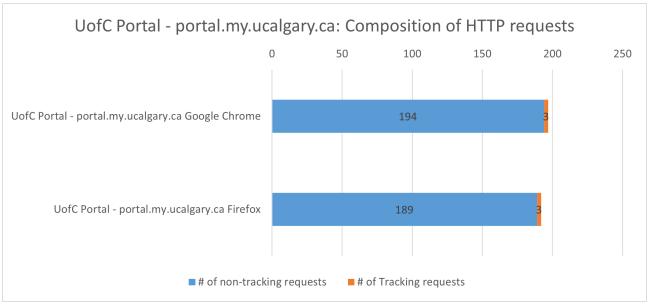


Figure 22: Composition of HTTP requests

The UofC Portal made requests to fonts.googleapis.com and google-analytics.com, both of which are marked as tracking domains. The Google Analytics request is similar to that discussed in Group 1 subsection E. 1 Google Analytics tracking pixel was found on each browser.

D. Gradescope - gradescope.ca

Gradescope is a service used by students and instructors to submit and grade assignments. The sequence of interactions with this website for the project include:

1. logging into Gradescope account
2. viewing a class's page
3. viewing a graded assignment

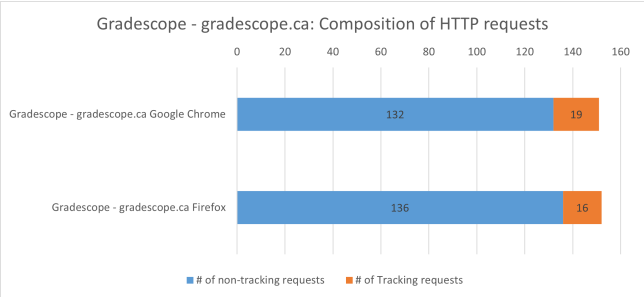


Figure 23: Composition of HTTP requests

The following tracking domains were found on both browsers:

- `google-analytics.com`
- `hsforms.com`
- `bugsnag.com`
- `googletagmanager.com`
- `amazonaws.com`

Google Analytics and Google Tag Manager are both web analytics services. Gradescope likely used Amazon Web Services's to deliver website assets to the user; for example, one of the requests to amazonaws.com loaded a video demonstration of grading on Gradescope. hsforms.com belongs to a company called Hubspot, which offers a online form builder service, and Bugsnag is a "Error Monitoring & App Stability Management" [24][25]. 2 Google Analytics tracking pixels were found on each browser.

E. University of Calgary - ucalgary.ca

ucalgary.ca is the main website of the University of Calgary. The sequence of interactions with this website for the project is as follows:

1. visiting ucalgary.ca
2. visiting "Undergraduate Overview" in the "Future Students" section
3. visiting "Explore Programs" in the "Future Students" section

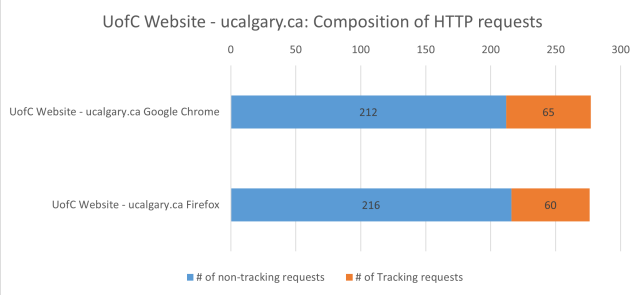


Figure 24: Composition of HTTP requests

The UCalgary website is the only site in Group 3 without any user log-in. The following tracking domains were found:

- `ads-twitter.com`
- `ajax.googleapis.com`
- `doubleclick.net`
- `facebook.com`
- `facebook.net`
- `fonts.googleapis.com`
- `google-analytics.com`
- `google.com`
- `googletagmanager.com`
- `gstatic.com`
- `licdn.com`
- `linkedin.com`
- `t.co`
- `twitter.com`
- `typekit.net`
- `youtube.com`

Out of all the websites in Group 3, ucalgary.ca has the most unique tracking domains and the highest ratio of tracking requests versus non-tracking requests. The website made 16 and 24 POST requests in Google Chrome and Firefox respectively. Unlike the other websites in Group 3, there was no reason to have to send data to the web servers, such as logging in or sending an email. As such, it is highly probable that the POST requests were purely for data collection and tracking purposes. 21 and 16 tracking pixels were also found in the Google Chrome and Firefox HARs respectively, all of which were made to Google Analytics.

VII. FUTURE WORK

Interacting with the websites visited, like the work done for Group 3, yields more tracker data than simply visiting them. As such, future work could involve manual and automated interactions with the webpages of interest, much like the IDAC's work for the 2020 United States presidential election[12]. Additional browsers and multiple redundant sets of HAR files should also be used, in order to reduce the chance of error or outlier HARs. Lastly, more research should be done into the technical details of the data collection performed by smaller tracking companies. Unlike large companies such as Google, their data collection methods and documentation are harder to find in the public domain. It is therefore in the interest of user privacy to understand how their services operate.

VIII. CONCLUSION

It is clear that in this day and age, our private information is regularly exposed to third parties for a variety of purposes. This paper offers an overview of the many covert techniques websites utilize to collect visitor information, from stealthy tracking pixels to long-lasting cookies. By combining experiment results, previous research, and official company policies and documentation, this paper explores the many bits of user information that tracking companies collect, as well as why even seemingly insignificant data can be used to track us. We hope that our research helps contributes to the field of privacy and encourages discussion about this important topic in our community.

IX. REFERENCES

References

[1] H. Nielsen, J. Mogul, L. M. Masinter, R. T. Fielding, J. Gettys, P. J. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol – HTTP/1.1," RFC 2616, Tech. Rep. 2616, Jun. 1999. [Online]. Available: https://www.rfc-editor.org/info/rfc2616

[2] C. Stouffer, "Internet tracking: How and why we're followed online," Jun 2021. [Online]. Available: https://us.norton.com/internetsecurity-privacy-internet-tracking.html

[3] D. Gourley, B. Totty, M. Sayer, A. Aggarwal, and S. Reddy, *HTTP: The Definitive Guide*. O'Reilly Media, Inc., September 2002.

[4] M. Belshe, R. Peon, and M. Thomson, "Hypertext Transfer Protocol Version 2 (HTTP/2)," RFC 7540, Tech. Rep. 7540, May 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7540

[5] M. Bishop, "Hypertext Transfer Protocol Version 3 (HTTP/3)," Internet Engineering Task Force, Internet-Draft draft-ietf-quic-http-34, Feb. 2021, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-ietf-quic-http-34

[6] T. Bujlow, V. Carela-Español, J. Solé-Pareta, and P. Barlet-Ros, "A survey on web tracking: Mechanisms, implications, and defenses," *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476–1510, 2017.

[7] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 413–427.

[8] C. Jensen, C. Sarkar, C. Jensen, and C. Potts, "Tracking website data-collection and privacy practices with the iwatch web crawler," in *Proceedings of the 3rd Symposium on Usable Privacy and Security*, ser. SOUPS '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 29–40. [Online]. Available: https://doi.org/10.1145/1280680.1280686

[9] C. J. Bennett and R. M. Bayley, "Canadian federal political parties and personal privacy protection: A comparative analysis," March 2012.

[10] House of Commons of Canada. Forty-second Parliament, First Session, "Elections modernization act," December 2018. [Online]. Available: https://laws-lois.justice.gc.ca/eng/annualstatutes/2018_31/page-1.html

[11] O. of the Privacy Commissioner of Canada, "Guidance for federal political parties on protecting personal information," April 2019. [Online]. Available: https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/gd_pp_201904/

[12] Q. Palfrey, N. Good, W. Monge, G. Kozemczak, and L. Ghamrawi, "Privacy issues in 2020 u.s. campaigns' apps & websites: An idac investigation & recommended best practices," November 2020. [Online]. Available: https://digitalwatchdog.org/wp-content/uploads/2020/11/IDAC-2020-Elections-Investigation-11232020.pdf

[13] R. Wiki, "What are tracking pixels and how do they work?" [Online]. Available: https://en.ryte.com/wiki/Tracking_Pixel

[14] Facebook, "Cookies policy." [Online]. Available: https://www.facebook.com/policy/cookies

[15] J. Evans, "How tracking pixels work." [Online]. Available: https://jvns.ca/blog/how-tracking-pixels-work/

[16] Longines, "Cookie notice." [Online]. Available: https://www.longines.com/cookie-notice

[17] S. Inc., "Cookie information," Jan 2022. [Online]. Available: https://snap.com/en-US/privacy/cookie-information

[18] A. Roselli, "Google analytics viewport tracking," Nov 2017. [Online]. Available: https://adrianroselli.com/2017/02/google-analytics-viewport-tracking.html

[19] P. C. Stuff, "Tracking browser window size using css," Jan 2017. [Online]. Available: https://www.parckwart.de/computer_stuff/tracking_browser_window_size_using_css

[20] Duckduckgo, "Faq.md," GitHub, November 2020. [Online]. Available: https://github.com/duckduckgo/tracker-radar/blob/main/docs/FAQ.md

[21] Cloudflare, "Cloudflare - the web performance & security company." [Online]. Available: https://www.cloudflare.com/

[22] Cloudflare, "Cloudflare web analytics." [Online]. Available: https://www.cloudflare.com/web-analytics

[23] S. Johnson, "What is click tracking? how should instructors use it to improve their courses?" Sep 2020. [Online]. Available: https://www.vanderbilt.edu/brightspace/2020/09/03/what-is-click-tracking-how-should-instructors-use-it-to-improve-their-courses/

[24] HubSpot, "Free online form builder." [Online]. Available: https://www.hubspot.com/products/marketing/forms

[25] Bugsnag, "Error monitoring & app stability management — bugsnag." [Online]. Available: https://www.bugsnag.com/