Tracker Installations Are Not Created Equal: Understanding Tracker Configuration of Form Data Collection

Julia B. Kieserman
New York University
New York, NY, USA

Athanasios Andreou
New York University
New York, NY, USA

Chris Geeng
Northeastern University
Seattle, WA, USA

Tobias Lauinger
New York University
New York, NY, USA

Damon McCoy
New York University
New York, NY, USA

Abstract

Targeted advertising is fueled by the comprehensive tracking of users' online activity. As a result, advertising companies, such as Google and Meta, encourage website administrators to not only install tracking scripts on their websites but configure them to automatically collect users' Personally Identifying Information (PII). In this study, we aim to characterize how Google and Meta's trackers can be configured to collect PII data from web forms. We first perform a qualitative analysis of how third parties present form data collection to website administrators in the documentation and user interface. We then perform a measurement study of 40,150 websites to quantify the prevalence and configuration of Google and Meta trackers.

Our results reveal that both Meta and Google encourage the use of form data collection and include inaccurate statements about hashing PII as a privacy-preserving method. Additionally, we find that Meta includes configuring form data collection as part of the basic setup flow. Our large-scale measurement study reveals that while Google trackers are more prevalent than Meta trackers (72.6% vs. 28.2% of websites), Meta trackers are configured to collect form data more frequently (11.6% vs. 62.3%). Finally, we identify sensitive finance and health websites that have installed trackers that are likely configured to collect form data PII in violation of Meta and Google policies. Our study highlights how tracker documentation and interfaces can potentially play a role in users' privacy through the configuration choices made by the website administrators who install trackers.

Keywords

Online tracking, PII, Form data collection, Form data configuration, Meta Pixel, Google Tag.

1 Introduction

Targeted advertising is a ubiquitous marketing technique that is fueled by tracking users online to create comprehensive profiles of their activity and (presumed) interests. In order to increase the completeness of a profile, activity from an individual across devices, apps, and websites is commonly linked using "hard identifiers" like email addresses and phone numbers. To collect such activity data, online advertising companies offer third-party analytics and tracking code. They encourage website administrators to include these trackers in their websites and configure them to extract Personally Identifying Information (PII) from forms that users may fill out on the website. The third party then uses this data for tracking and targeted advertising purposes.

Data extraction is of particular concern in verticals that handle sensitive consumer data, including health and finance data, and is thus subject to additional (federal) privacy regulations. In 2023, the U.S. Federal Trade Commission issued fines against two health companies, *BetterHelp* and *GoodRx*, for leaking sensitive data to third party tracking code providers [78]. This issue was not limited to two companies — journalists at *The Markup* have uncovered instances of similar data leakage at addiction service, finance, and college preparation companies [46, 51, 52].

While prior studies have analyzed leakage of PII to third parties [13, 15, 18, 21, 74, 76], as well as the prevalence of trackers and data collection on sensitive websites [20, 43, 72, 85], they have typically treated tracker installations as a binary, either installed or not installed. However, tracker installations are not created equal. They must be *configured* to enable extraction of PII from forms, and to date, it is unclear how often form data extraction is enabled in practice and how the trackers' documentation and configuration user interface may influence website operators' choices.

To the best of our knowledge, no prior work explores tracker configuration across websites and the interplay between tracker documentation and real-world configurations. In this paper, we address this gap with a qualitative and quantitative approach to study how trackers can be configured for web form data collection, how these configurations are presented in the documentation and interface, and how many trackers are configured to do so in practice.

We focus on *Google Tag* and *Meta Pixel*, provided by Google and Meta respectively, as they are the two most popular web trackers with form data collection capabilities [1]. First, we qualitatively code the documentation and configuration user interfaces of each tracker for dark patterns and other potentially confusing language (Section 3). Second, we conduct a measurement study of 40,150 websites, including 3,406 health and 1,633 finance websites, to quantify how often the two trackers are installed and configured to extract PII data from web forms (Section 4).

In summary, we are guided by the following research questions. Qualitatively, we explore (i) *how are form data collection features*

Proceedings on Privacy Enhancing Technologies 2025(4), 679–695

https://doi.org/10.56553/popets-2025-0151

configured, (ii) *how are those configurations explained in the documentation*, and (iii) *what measures do third parties take to ensure that web administrators working with sensitive data are protecting that data as required by US law*? Quantitatively, we investigate (iv) *how prevalent is form data collection for Google Tag and Meta Pixels*, (v) *do tracker installations in health and finance verticals have different incidences of form data collection than non-sensitive verticals*, and (vi) *what types of PII are trackers configured to collect*?

We find that while *both Google and Meta encourage web administrators to enable form data collection by recommending the least private default configuration without addressing the potential risks*, they have significantly different interfaces. *Meta Pixel* has a streamlined set-up workflow that guides the web administrator to decisions that maximize data collection and often omits privacy considerations. *Google Tag* has a complex flow with contradictory statements that might make it challenging for a web administrator to assess the state of data collection configuration.

In line with this finding, our measurement reveals that *Meta Pixels are frequently configured to collect email addresses, names, and phone numbers*, either by enabling the default configuration settings that collect all supported PII fields or by custom configurations. 93.5% of the websites that enable form data collection for Meta collect phone numbers, 93.7% full names, and 99.5% email addresses. Furthermore, *Meta Pixels are more frequently configured to collect form data compared to Google Tags (62.3% vs. 11.6%).* This may be a result of Meta's aforementioned tracker configuration flow, which guides the website administrator to configure form data collection. In contrast, Google's tracker does not include form data collection as part of the setup workflow.

We also find that *both Meta and Google address federal regulatory privacy restrictions on health and finance data* by requiring website administrators to indicate the vertical of their website during the account creation or tracker configuration process. However, neither provides detailed explanations that would help web administrators understand the implications of this designation.

In our measurement, we find that *form data collection is less common on finance and health websites for Meta Pixels but not for Google Tags*. Specifically, 68% of websites with *Meta Pixel* in non-sensitive categories collect form data, compared to only 30.8% for health and 20.3% for finance websites, respectively.

In analyzing the configuration of trackers from multiple perspectives, we make four primary contributions:

- We create methodologies for qualitatively analyzing tracker configuration documentation and a data collection and analysis pipeline that enables large-scale measurement of website tracker configurations.
- We expose how Google and Meta recommend website administrators enable automatic PII collection from form data and are providing privacy advice that has been repeatedly debunked by the US FTC.
- We reveal that popular websites configure *Meta Pixel* to collect form data much more frequently than *Google Tag*.
- We identify specific finance and health websites that have likely configured *Google Tags* and *Meta Pixels* to automatically collect PII form data in violation of Google's and Meta's policies.

We believe that our study offers a unique perspective on how instructions and setup guides can drive configuration decisions, thus providing potential technical defenses and guidance to regulators seeking to improve user privacy.

2 Background and Related Work

Online tracking is pervasive on modern websites [8, 69]. In some cases, tracking is installed on a website through a joint effort between advertising third parties and website administrators, who maintain individual websites. These trackers take the form of a library of functions developed and made available by the third party. Not only do they record visits to websites where they are installed, but they also aim to collect Personally Identifiable Information (PII), such as email addresses or phone numbers, so that website visits can be linked to an (advertising) identity that the third party has established for a website visitor. Prior work has found that trackers do in fact facilitate the collection of website visitors' PII back to the third parties that developed them [13, 53]. This PII is collected from forms filled out by website visitors including account registration forms [15], login forms [18], and contact forms [76]. In some cases, PII can even be collected from a form before it has been submitted [74].

The specific type of PII collected, as well as the method of collection, varies across third-party trackers. This paper focuses on *Meta Pixel* and *Google Tag* as they are the two most popular web trackers with form data collection capabilities [1]. *Meta Pixel* can automatically search for the following categories of user data by regular expression: email, gender, address, name, phone number, date of birth, and advertising user id (external-id) [66]. By contrast, *Google Tag* identifies only email addresses (also using a regular expression) but offers the option to define website-specific CSS selectors or JavaScript variables for email, address, name, and phone number [33]. Both Meta and Google trackers have an automatic and manual form data collection method; automatic collection is configured through the user interface and manual collection is configured by modifying website source code.

As a concrete illustration, we discuss how `techcrunch.com`, a top-ranked news website, is configured to collect data using *Meta Pixel*. The website has a form for subscribing to an email listserv on the landing page. This flow is illustrated in Figure 1. To install the *Meta Pixel* tracking code, a website administrator had to create a Facebook account and walk through the setup steps. The website administrator may or may not be internal to the TechCrunch organization. During the process of configuring the tracking code, they would be prompted to turn on automatic data collection for several properties, including email. In this instance, `techcrunch.com`'s web admin chose to enable automatic data collection for the following PII: email, first and last name, phone number, gender, zip code, city, and state. They then installed the tracking code on their website by copying and pasting a few lines of JavaScript code from Meta's user interface. This code will query a Meta server to load the configuration options selected by the website administrator. Now, when a website visitor decides they would like to subscribe to a newsletter, enters their email address, and clicks submit, their data will silently be passed to Meta. Meta's tracker may also collect data in response to other click events, including other buttons on the page or anchor tags not associated with the subscription form.[1]

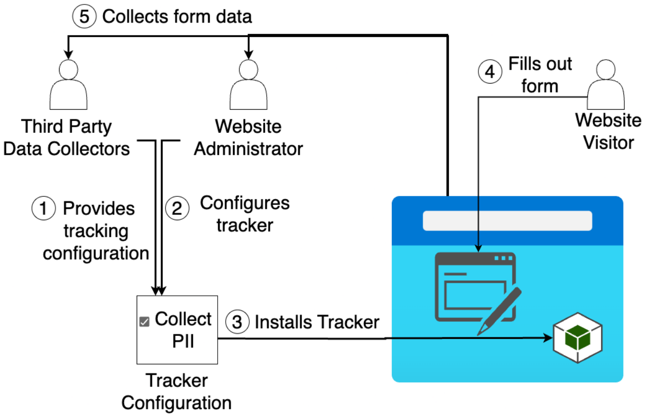


Figure 1: Tracking code setup flow.

Although always privacy-invasive, this data collection becomes particularly problematic when it occurs on websites that handle legally protected data. In the United States, health and finance data are protected by the Health Insurance Accountability Act of 1996 (HIPPA) [81] and the Gramm-Leach-Bliley Act [28], respectively. Prior research has surfaced the role trackers have played in leaking website data in these sensitive verticals. Robinson et al. found *Google Tag* was installed on the majority of Illinois' hospital websites [72], Huo et al. identified *Meta Pixel* and *Google Tag* data breaches across electronic health record portals [43], and Bekos et al. found that *Meta Pixel* tracks users across time in website verticals considered sensitive by GDPR [8]. Additionally, reporting by news publication *The Markup* anecdotally found that *Meta Pixel* installed on hospital and tax preparation websites sent website visitors' names, doctors' names (hospital websites) and income amounts (tax preparation websites) back to Meta [25, 52].

Ultimately, a tracker will only collect form data if it is configured to do so, which is determined by a website administrator at the time of installation. Throughout this paper, we use the term "website administrator" to refer to anyone who decides to configure and install tracking code, regardless of their technical expertise or relationship to the organization that owns the website. In practice, the decision to install a tracker and configure it for data collection may come from a combination of internal teams, like Marketing, Legal, and IT, or be outsourced to a third-party contractor. In some cases, this may leave the original website owners unaware of the configuration setting [50]. Typically, website administrators install trackers for marketing purposes, namely to monitor website traffic, identify online customers that have followed an ad on another platform (i.e. conversions), and perform "dynamic re-marketing" [40], which targets ad campaigns to former website visitors on different platforms.

In order to take full advantage of the marketing features of trackers, website administrators must install them such that they can collect website visitor PII. This requires that they are installed on a web page with input form fields, since trackers are scoped to a single web page, and that they are configured to use form data collection features. Configuring form data collection may not always be straight-forward to implement or understand, and as such website administrators may end up misconfiguring trackers. The consequences of misconfiguration can include legal non-compliance and data leaks. Maass et al. identified websites that had misconfigured *Google Tags*, thus violating German regulation that required anonymizing visitors' IP addresses [54]. Prior work limited to the mobile space has further demonstrated challenges that website administrators may face in attempting to configure other kinds of privacy-impacting configurations, including working with challenging or misleading privacy APIs [84] or having an incomplete understanding of third-party SDKs [5].

Third parties assist web administrators with these decisions through a combination of documentation and user interface prompts in a configuration portal. Many third parties design tracking code to be primarily configured through a user interface, which allows less technically-versed people to manage the configuration. This expectation is made clear in the documentation with prompts such as "ask your web developer" when outlining technical steps [37].

Prior research has identified the importance of quality documentation [4, 80], with a specific focus on the quality and thoroughness of privacy documentation. In the mobile space, research has shown that data collection guidance from third-party SDKs is lacking in compliance information [48] and contains discrepancies between documented and actual behavior [44]. Even when configurations are left untouched, relying on the default configuration behavior may lead to less private behavior [73, 79], a pitfall known as "bad defaults" that affects both developers [12] and consumers [42, 55].

Prior work comprehensively demonstrates the prevalence of trackers that collect form data and the challenges of documenting and configuring technical tools. However, to our knowledge, we are the first to examine and measure the specific configurations of *Meta Pixel* and *Google Tag* that lead to form data collection, particularly in the context of how third parties explain and guide website administrators to configure them.

3 Documentation and User Interface Analysis

In order to effectively evaluate how form data configurations are explained in third-party documentation, we first needed to build an adequate understanding of how form data configurations worked. To do so, we played the role of a website administrator by creating a test website that included an input form and installing *Meta Pixel* and *Google Tag* as directed by the user interfaces and documentation for each tracker. Two of the paper authors then followed a reverse engineering process, modifying the tracker and observing the subsequent changes in both the form data collected and the JavaScript code loaded by the browser. We relied on this expert knowledge to evaluate how the documentation and user interface explained form data collection configurations.

3.1 Methodology

3.1.1 Setup. We reviewed 17 pages of Meta documentation and 60 pages of Google documentation produced specifically by each third

[1] This information was provided to the authors in personal communications with another researcher and verified by the authors.

party, including screenshots from the user interface. To identify documents to review, we started from the user interface, specifically focused on setup flows that did not rely on a CMS or website builder (such as Shopify or Wordpress). We then collected any documentation that was explicitly referenced or directly linked in the user interface. Finally, we included documentation we found through a Google search for how to setup the *Meta Pixel* or *Google Tag* or related to specific polices or best practices regarding form data collection configuration. All documentation was accessed between May and June 2024.

3.1.2 Process. When reviewing the selected documentation, we were guided by the following research questions:

RQ1. How are form data collection features configured and how are those configurations explained in the documentation?
RQ2. What measures do third parties take to ensure that web administrators working with sensitive data are protecting that data as required by US law?

The two authors who participated in the reverse engineering process analyzed the user journey [79] by reviewing the setup flow and documentation with a mix of deductive and inductive coding techniques. We took inspiration from existing dark patterns [56, 79], software documentation [80], and behavioral economics theory [45] literature. All documents were independently coded by each author before discussing. Documents were coded in an iterative process until a consensus was reached. Disagreements were resolved with the help of two additional researchers.

3.1.3 Codebook. We created five codes, which are additionally enumerated in Table 1 with specific examples.

Hidden risk. A feature presented in such a way that a web administrator can reasonably believe they are presented with all the relevant information to make an informed choice when, in reality, additional privacy-related risks are hidden. This code takes inspiration from dark pattern “hidden costs” [11].

Least private defaults. Default configurations set by the third party that maximize data collection. Since web administrators might not modify defaults, this may lead to more data collection than would otherwise occur. This code takes inspiration from dark pattern “bad defaults” [10].

Least private recommendations. Instances where third-party documentation has explicit recommendations or best practices that maximize data collection. This code was created inductively.

Loss aversion. Language that creates the perception that a web administrator would be missing out on opportunities or not getting the full offering of a certain feature by not making data maximization decisions. This code comes from the field of behavioral economics [45].

Contradictory language. Instances where third-party guidance appears to be in contradiction to the underlying mechanisms of the data collection feature, similar to clarity issues previously explored in the context of documentation [80].

3.2 Results

3.2.1 How are form data collection features configured, and how are those configurations explained in the documentation? We found that Meta and Google recommend web administrators configure *Meta Pixel* and *Google Tag* for data collection across the documentation and user interface and hide the possible risks of sharing web visitors’ PII data with a third party.

Meta and Google often recommend the least private configuration option, which might create hidden risks for the website administrator. For example, across the documentation, Meta recommends that web administrators use both manual and automatic data collection (as defined in Section 2) to increase the amount of data collected for “maximum performance” [60]. Similarly, Google recommends that website administrators install a *Google Tag* “on every page of your website” rather than consider what pages are most appropriate for its intended purpose [30]. Since these recommendations come directly from the third parties, it is reasonable for a web administrator to take them as a form of best practice, even though they present a least private approach.

Google and Meta also insufficiently address the risks of configuring form data collection. Both Google and Meta’s documentation contains language implying that hashing is sufficient for privacy, a claim that has been debunked both by researchers and by the US Federal Trade Commission (FTC) twice in the past twelve years [19, 26, 27]. Meta’s documentation states that they “hash the customer information on the website...to help protect user privacy” [57]. However, by using Meta’s preferred hashing function (SHA-256) [59], Meta can link customer data collected from the website to an existing Meta profile. Therefore, by omitting any other details, this language hides the risk of configuring *Meta Pixel* to share visitors’ PII with Meta. Similarly, Google refers to “sending hashed first party conversion data from your website to Google” as sending data “in a privacy safe way” [29].

We found that the documentation for both providers contains the same language multiple times across different setup guides. For example, Meta states their preferred matching setup, ensuring that a page has “form fields” that collect PII information, three different times [57, 60, 61]. Google states a preference for receiving user email addresses to help identify customer leads five times across two documents [38, 39]. This duplication could reinforce the effect on the reader and, since the documentation may not be accessed or read in a specific order, increase the opportunity for someone installing the tracking code to encounter these recommendations.

Although both providers generally encourage configuring data collection, we did find some discrepancies between how Meta and Google describe and present data collection configuration.

Meta Pixel. Meta provides website administrators with a setup workflow that heavily recommends configuring data collection. When installing a new *Meta Pixel* through the basic setup flow, a wizard includes a screen dedicated to automatic data collection, described as using “information that your customers have already provided to your business” (Figure 3). Including a prompt to turn data collection on in the setup flow can be read as a least private recommendation by Meta to turn it on, insinuating that data collection is part of normal *Meta Pixel* use. They further omit any descriptive text about privacy considerations related to PII data

Code	Definition	Example	Third Party
Hidden risk	A feature presented in such a way that a web administrator can reasonably believe they are presented with all the relevant information to make an informed choice when, in reality, additional privacy-related risks are hidden	*We hash the customer information on the website before they're sent to Meta technologies to help protect user privacy*	Meta
Least private defaults	Default configurations set by the third party that maximize data collection	*You have the option to enable/disable the collection of granular location-and-device data on a per-region basis. Analytics collects this data by default.*	Google
Least private recommendations	Instances where third-party documentation has explicit recommendations or best practices that maximize data collection	*Toggle ON the parameters you want to share from your website or app. We suggest selecting at least Email and Phone number for the best results.*	Meta
Loss aversion	Language that creates the perception that a web administrator would be missing out on opportunities or not getting the full offering of a certain feature by not making data maximization decisions	*The more items you complete on the checklist, the more complete your GA4 [Google Analytics] data will be. Many configurations determine what data is collected in your property, so it will only be available from when you complete them. That's why it's valuable to do them as soon as possible.*	Google
Contradictory language	Instances where third party guidance appears to be in contradiction to the underlying mechanisms of the data collection feature	*Google policies mandate that no data be passed to Google that Google could use or recognize as personally identifiable information (PII)*	Google

Table 1: Code book with examples observed in Google's and Meta's web tracker documentation and configuration user interfaces.

collection, creating a hidden cost for web administrators who may not consider the risks without further information. Further, if the web administrator takes the prompt, all of the attributes Meta can collect are automatically selected. By assuming the web administrator would like to collect all possible attributes, Meta has set up a least private default that relies on the web administrator to actively deselect attributes to increase customer privacy (Figure 4).

The documentation also mentions inconsistent behavior that may occur if a website has multiple *Meta Pixels* installed [58]. If one pixel collects form data and another pixel does not, both pixels may collect form data collection when one is triggered to do so. Although it is explicitly mentioned in the documentation, it is not included as part of the setup flow nor does the documentation adequately explain the implications for potential data leakage.

Data submission events also behave in unexpected ways. When automatic form data collection is enabled, Meta's data collection logic activates on any button click event, whether or not the button is connected to a form with data [65]. This could create instances of form leakage behavior, not just of an email and password, as documented by Senol et al. [74], but also of other supported PII types such as name, address, or gender. We further found instances where a form data collection event was triggered by a completely different type of click event, including clicking on a hyperlink, which was not specified in the documentation.

Google Tag. Google's tracking ecosystem is complex. Form data collection is configured across several different parts of the user interface and *Google Tag* can send data to different Google products, including Google Ads and Google Analytics. While this complexity does not necessarily increase the likelihood that a website administrator will configure form data collection, it can make it hard for website administrators to determine if a *Google Tag* is configured to collect form data. Neither the documentation nor the user interface make it clear how to ensure that tracking code will behave as intended. Unlike *Meta Pixel,* the data collection feature is not explicitly mentioned in the tracker set-up flow and thus requires self-direction to turn on.

We also found that the documentation uses contradictory statements. For example, Google asserts that "Google policies mandate that no data be passed to Google that Google could use or recognize as personally identifiable information (PII)" [35]. However, other parts of the documentation make it explicit that tracking code "uses first-party user-provided data from your website" [38]. Although this discrepancy is likely drawing a distinction between hashed and un-hashed PII, this distinction is not made explicit (and, as stated, hashing is insufficient as a privacy method), therefore this can be read as a direct contradiction in the documentation.

3.2.2 What measures do third parties take to ensure that web administrators working with sensitive data are protecting that data as required by US law? In the United States, sensitive data in health and finance verticals are provided with specific legal protections. Both Meta and Google address regulatory restrictions by applying constraints on automatic form data collection for websites belonging to either of these verticals. However, a website's vertical classification is entirely self-reported and neither Google nor Meta offer reasonable explanations that would help a web administrator understand the importance of the designation.

Meta states that "businesses...may not have certain features available to them if they're categorized as being in a restricted vertical" [64], omitting that the reason for this is because they presumably deal with especially sensitive data. They further state that businesses "learn how to set up [data collection] manually" [60], suggesting these restrictions can be circumvented by using manual data collection techniques. This may result in a web administrator modifying the website source code to send the same data to the third party that was intentionally restricted in the user interface. Similarly, *Google Tag* states that data sharing "is not available to Analytics accounts with properties in the "Health" and "Finance" property industry categories" [31] without further explanation.

3.2.3 Takeaway. In summary, we observed the following:

- Both Meta and Google recommend configuring data collection without sufficiently addressing the privacy risks of doing so.
- Meta's basic setup flow prompts website administrators to configure data collection and, when configured, collects 11 types of PII by default.
- Both Meta and Google technically restrict websites in sensitive verticals, but vertical identification is at the discretion of the website and not adequately explained in the documentation.

4 Website Configurations

Thus far, we have analyzed how the *Meta Pixel* and *Google Tag* present form data collection to website administrators in the documentation and user interface. We proceed by measuring website configurations to understand what impact this may have had on actual websites.

4.1 Methodology

In order to compare the configurations of websites in health, finance, and other verticals, we needed to generate a list of top websites by vertical. To do so, we joined the top one million websites from Tranco [49] with SimilarWeb, a web traffic estimator that categorizes websites by vertical (e.g., finance, games, health, shopping, travel) [75] and has been used by prior studies [77, 82, 83, 85]. This generated a list of 42,481 websites with quartile ranks of 34,429, 110,756, 355,814, and 999,979 for quartiles 1, 2, 3, and 4, respectively.

To measure data collection configurations on each website, we simulated form data collection, scraped website tracker code and collected network traffic, and then parsed the data to detect the presence of trackers (i.e. *tracker installation*) and whether they were configured to collect form data (i.e. *form data collection (FDC)*). The pipeline is outlined in Figure 2.

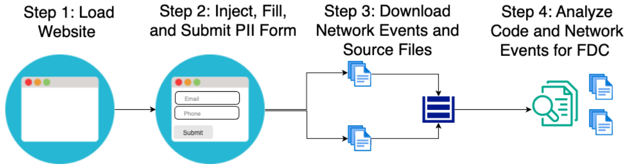

Step 1: Load Website
Step 2: Inject, Fill, and Submit PII Form
Step 3: Download Network Events and Source Files
Step 4: Analyze Code and Network Events for FDC
Email
Phone
Submit

Figure 2: Data collection and analytics pipeline.

4.1.1 Data Collection. On each website we performed the following data collection actions:

(1) *Form Injection Action:* Inject a form with PII fields into the page and submit it. If a tracker is configured to collect form data, it will generate a network event that sends these data to Google or Meta.
(2) *Download Action:* Capture all the JavaScript files and network events loaded by the webpage. This collects any potential *Google Tag* and *Meta Pixel* configuration files present and any form data collection events that might have occurred as a result of the form injection action.

We note that we visited only the website landing page, which may have a smaller incidence of tracker installations than other pages on certain types of websites [43]. We chose not to interact with cookie consent banners present on the page to avoid bias towards any one user behavior (accept or decline). We address the potential impact of this decision in Section 4.2.

Instrumentation. All website visits took place on Linux virtual machines in Google Cloud's US Central region between September and November 2024. We ran a total of 50 Virtual Machines (VMs) in parallel with identical configurations and a different subset of our website list.

Each virtual machine had a Python script that would start on machine boot and iterate through a subset of our input website list. We used Google Cloud infrastructure (buckets and functions) to maintain the visit status of each website and communicate with individual VMs. The Python script opened each website in Chrome for 180 seconds to provide sufficient time for the page to load and our website visit actions to complete. In order to reduce the possibility of bot detection, we chose not to rely on commonly used automation tools like Selenium or OpenWPM [22] and restarted each machine every three hours.

Each website was opened in a Chrome instance with two installed custom Chrome extensions. The extensions performed two actions: a form injection action and a download action. The tasks were triggered after the page finished loading, determined using Chrome's manifest v2 webNavigation feature [36].

Form Injection Action. On each website, we injected an HTML form into the webpage with several PII fields; namely email, phone number, first and last name, city, state, and zip code. Our extension filled out the form fields with placeholder data and then triggered the submit button. The form was attached to the first `<div>` or

<span> node found on the page. Our extension only injected into the top document of the page and not into nested iframes. The intent of this action was to trigger a form data collection event if a tracker on the page was configured to do so.

We opted to use our own form instead of one present on the page, as we were interested in measuring the tracker configuration, which is independent of the current HTML elements on the page. This method enabled us to measure the configurations of all websites, regardless of whether they currently have a PII form on the landing page. Consequently, we are not measuring acutal observed PII leaks but rather whether a website is configured in a way that could potentially cause PII leaks.

Download Action. On each website, we downloaded all the page source files, including the JavaScript and HTML files. We also downloaded all network traffic detected during our website visit, both through the network HAR file and using Chrome's manifest v2 webNavigation feature [36]. This ensured we captured any form data collection events potentially triggered by our form injection action and the JavaScript configuration files of any *Meta Pixel* and *Google Tag* tracker files loaded by the website.

Dataset Collection Retries. After our initial scraping, we identified the subset of websites that either had no instances of *Google Tag* or *Meta Pixel* tracker installations, or loaded a *Google Tag* or *Meta Pixel* but had no detected instances of form data collection, using a methodology described in the subsequent subsection. To reduce possible false negative detections, we then performed another round of website visits, including form injection, within this subset of websites. We repeated this process until every website had been visited at least three times or had a detected instance of form data collection. The diminishing returns are enumerated in Section 4.2.

We analyzed the data across all visits with the following parsing logic.

4.1.2 Data Parsing.

After each scrape, we processed all collected files (i.e., network traffic and downloaded source code) through our parsing pipeline. Our parsing logic allowed us to answer three questions: (1) Did this website have a *Meta Pixel* and/or *Google Tag* installed? (2) Did the *Meta Pixel* and/or *Google Tag* perform form data collection with our placeholder form data? (3) Was the *Meta Pixel* configuration file configured to collect form data?

Detecting Tracker Installation. We used network GET requests to determine if a website had a *Meta Pixel* or *Google Tag* installed. The specific URLs used to load each tracker were identified during the reverse engineering process. We determined if a website had a *Meta Pixel* installed by parsing *Meta Pixel* configuration files, loaded by a GET request to *connect.facebook.net/signals/config/*. An example is shown in Figure 6 in the appendix.

Similarly, we determined if a website had *Google Tag* installed by parsing *Google Tag* configuration files, loaded by a GET request to *googletagmanager.com* (see Figure 8 in the appendix). There are four types of tags that feed data to different parts of the Google ecosystem but are all considered valid *Google Tag* installations. They can be identified by their tag ID prefix: `AW` (Google Ads), `DC` (Google Floodlight), `G`/`GT` (Google Analytics 4), and `UA` (Universal Analytics).

In addition, Google supports first-party mode [32], which allows website administrators to install infrastructure that sits between a website and Google's servers. In these cases, *Google Tag* will be loaded from a domain chosen by the website administrator instead of Google, but still retain the *googletagmanager* file name. We consider these valid installations if they match the structure of a standard *Google Tag* configuration file.

Detecting Form Data. To measure instances of form data collection, we performed both a dynamic and a static analysis. The dynamic analysis allowed us to determine if *Google Tags* and *Meta Pixels* collected form data, and the static analysis allowed us to determine if and how *Meta Pixels* were configured for form data collection (unfortunately, we were unable to directly determine how *Google Tags* were configured).

In our **dynamic analysis**, we leveraged the output produced by our form injection action; we parsed captured network traffic and detected form data collection through known tracker URLs with query parameter values containing the (hashed) placeholder data we submitted through the form injection task. These URLs were identified in the initial reverse engineering process, and are listed in Table 2. We identified our data by hashing it according to the standard for each tracker (SHA-256 for *Meta Pixel* and SHA-256 on a base64 encoded string for *Google Tag*) and then searching network events for a matching string hash.

If we found any instance of our hashed PII data in a request to the known tracker URLs, we labeled the *Google Tag* or *Meta Pixel* tracker as a website with form data collection (FDC). An example of a *Meta Pixel* form data collection network event and *Google Tag* form data collection network event are presented in Appendix Figures 7 and 9 respectively. The highlighted portion is a hashed version of the placeholder email address submitted through our form and detected by our parser.

In our **static analysis**, we parsed the *Meta Pixel* tracker files collected during the download task to identify if the tracker was configured for data collection. Through our reverse engineering process, we were able to identify where the user interface configuration choices, described in Section 3, appeared in the JavaScript files loaded by the browser. This analysis can determine not only whether a *Meta Pixel* is configured to collect data but also which of the 11 supported PII fields it was configured to collect, all of which are selected by default. Screenshot Figure 5 in the appendix illustrates how this configuration appears in the JavaScript code, identified as *selectedMatchKeys*.

Unfortunately, the Google ecosystem is more complex, involves more components, and has been intentionally obfuscated, making static analysis more challenging. In addition, *Google Tag* form data collection is limited to email; all other PII field collection requires further manual customization from website administrators. Therefore, we leave static analysis of *Google Tag* code for future work.

Form Data Collection: Automatic vs. Manual. As mentioned in Section 2, *Meta Pixel* and *Google Tag* support two modes: manual (when a website administrator writes JavaScript to send PII fields to the third party explicitly) and automatic (when a website administrator authorizes the third party to auto-detect PII fields through the user interface).

Third Party	Data Collection URLs
Meta	facebook.com/privacy_sandbox/register/trigger facebook.com/tr
Google	googleadservices.com/pagead/conversion google.com/ccm/form-data/ analytics.google.com/g/collect google.com/pagead/form-data/

Table 2: Tracker URLs that receive form data from *Meta Pixel* and *Google Tag* installations on websites.

We were able to differentiate between *Meta Pixel* automatic and manual form data collection by the keys specified in the network event, determined through trial and error during the reverse engineering process. Automatic form data collection uses the key *udff*, followed by an abbreviation of the PII type. For example, email configured through the user interface will appear in a url as *udff[em]*, followed by the hashed email, as demonstrated by the highlighted text in Appendix Figure 7. Manual form data collection uses the key *ud* but otherwise appears the same way, i.e. with the PII abbreviation (or unabbreviated *external_id*) in brackets immediately following the identifier. The majority of websites with detected trackers (90.92%) exclusively used automatic mode. The remaining websites used a combination of the two and, when manual mode was used, it was primarily for the external ID field (92.82%).[2]

We were not able to make the same distinction for *Google Tag*, as the automatic and manual network events appeared identical. However, this distinction is less meaningful for *Google Tag* than it is for *Meta Pixel*, as both manual and automatic modes require some interaction with the *Google Tag* configuration UI, and in some cases, *Google Tag* will default to automatic mode if manual data is not detected [2]. Therefore, we do not distinguish between the automatic and manual modes of *Google Tag* in our analysis.

In the rest of the paper, we refer to tracker installations for which we detected a PII collection event as **Form Data Collection,** and to *Meta Pixels* for which our static analysis identified JavaScript data collection configuration as **FDC Configuration**. We use our form data collection analysis to report comparative measures on the differences between *Meta Pixel* and *Google Tag*. We use our FDC configuration analysis to explore the specific PII data fields that *Meta Pixels* are configured to collect.

4.2 Dataset Overview and Validation

In this section, we briefly introduce our results and then proceed with an overview of our dataset validation process. A more in-depth discussion of our results is reserved for Section 4.3.

Out of the 42,481 websites visited, we visited 40,150 (94.51%) successfully. Unsuccessful visits can be attributed to several factors, including inaccessible domains, bot detection, and general unreliability. To reduce the likelihood of these errors, we implemented retry logic, re-visiting each website until a form data collection event was detected or after the third visit. Google form data collection and Meta form data collection were considered independent events, i.e. even if a website had an instance of form data collection from one of the two trackers on the first try, we still made additional attempts for the other tracker. All of our further analysis is based on only the 40,150 websites that we were able to successfully visit. Our analysis considers whether a website ever had a tracker installation or form data collection in *any* of the scrapes over the three-month period of data collection. There is a possibility that a tracker configuration was altered between multiple visits, but we do not make such a distinction in our analysis.

We observed the following marginal gains from each retry. For *Google Tag* tracker installations, the first visit discovered 85.4% of our total tracker installation count, the second visit found an additional 13.67%, and the third visit identified less than 1%. We found a similar pattern for *Meta Pixel* tracker installations. The first visit discovered 91.28%, the second visit 8.09%, and the third visit less than 1% of new observed tracker installations. For *Google Tag* form data collection, we found that the first visit contributed 85.35% of our total, the second visit 11.91%, and the third visit 2.75%. For *Meta Pixel* form data collection, we found the first visit contributed 89.64%, the second 7.48%, and the third 2.88%. Therefore, we believe further retries would not have led to a meaningful increase in either tracker installation or form data collection numbers.

Even with retries, by the nature of large-scale data scraping, many factors may have impacted the accuracy of our measurements. To quantify this impact, we performed a number of manual validations that address the following questions: (i) *How many tracker installations and form data collections are we missing?* and (ii) *How accurate is our form data collection detection methodology as a proxy for measuring FDC configuration?*

Additionally, as mentioned in Section 4.1, we chose not to interact with cookie consent banners. In order to understand the impact this had, we performed an additional validation to answer the question: (iii) *How many more tracker installations and form data collections would we detect if we accepted or rejected cookies?*

We answered these validation questions by manually investigating samples of websites. Unlike our automated scraping, all validation was done on local machines by the paper's authors. The sample sizes used in our validation are based on population counts presented in Table 3. Every sample referenced in this section was selected uniformly at random from the relevant population, and its size was calculated using a population proportion $\hat{p} = 0.5$ (the worst case) to achieve a 95% confidence interval.

4.2.1 *How many tracker installations and form data collections are we missing?*

To measure the extent to which we under-counted the number of tracker installations, we performed manual validation on two samples: websites with no detected *Meta Pixel* tracker installations and websites with no detected *Google Tag* tracker installations. Note that we considered *Meta Pixel* and *Google Tag* independent of each other, so each sample may have included instances of the other. We did not explicitly measure false positives. Given that we identified FDC configuration and form data collection from parsed network events, we considered false positives to be unlikely.

For *Google Tags*, we visited a sample of 372 out of the 11,013 websites with no detected *Google Tag* tracker installation. Of the 345 reachable websites, 91.6% (316) were true negatives, and 8.4% were false negatives. Similarly, out of the 28,841 websites without

[2] The External ID represents users in an advertising system [63].

	Tracker Installation		**FDC Configuration**			**Form Data Collection**		
	Websites	*All Websites*	*Websites*	*All Websites*	*Tracker Installation*	*Websites*	*All Websites*	*Tracker Installation*
Google	29,137	72.6%	—	—	—	3,377	8.4%	11.6%
Meta	11,309	28.2%	7,849	19.5%	69.4%	7,049	17.6%	62.3%
Google ∪ Meta	29,363	73.1%	—	—	—	8714	21.7%	29.7%

Table 3: Overview of all tracker installations, and tracker installations with FDC configuration and form data collection for Google, Meta, and websites that have either (Google ∪ Meta). The denominator for "All Websites" percentages is 40,150.

a detected *Meta Pixel* tracker installation, we manually visited a sample of 380 websites. We successfully reached 368, of which 95.7% (352) were true negatives and 4.3% (16) were false negatives.

Although we do not know the exact reason for each missed tracker installation, we were able to identify some possible reasons through our manual validation. We observed instances where trackers loaded too slowly for our data collection infrastructure (i.e., our automation closed the website before the tracker had a chance to load) and instances where we were detected as a bot. It is also possible that some websites added tracker installations after we performed the data collection but before we performed manual validation.

We validated missed form data collection events in a similar manner. For each website in the selected sample, we manually visited each website and injected and submitted a filled PII form. Although this task was similar to our form simulation task described in the methodology, it was performed manually on local machines, and thus, human judgment was used to determine how best to inject the form.

To validate form data collection events from *Google Tag,* we selected a sample of 379 websites from the 25,760 websites with *Google Tag* installation but no form data collection. We successfully reached 364 websites but found that 13 of them did not have a *Google Tag* at the time of our validation, so we excluded them from our analysis. Of the remaining 351 websites, only 5 (1.4%) were false negatives.

Similarly, for Meta, out of the 4,260 websites with *Meta Pixel* but no detected form data collection, we investigated a sample of 353 websites. Of the 343 websites successfully visited, 19 websites did not have a *Meta Pixel* at the time of validation. Out of the remaining 324 websites, 31 (9.6%) were false negatives.

We notice a higher percentage of missed form data collection with *Meta Pixel* than *Google Tag.* This is possibly explained by the fact that far more *Meta Pixels* in general are configured to collect data than *Google Tags.* Although we cannot be precisely sure why we are under-counting, it could be attributed to the same reasons we observed missed tracker installations. In addition, we note that automated form injections can be challenging and, given the diversity of website designs, there is the potential that some of our injections failed, thus contributing to our under-counting.

We generally found that a subset of the false negatives could be attributed to cookie consent or other dialog boxes that blocked page load (17.3%), bot detection (8.6%), and websites we had not successfully visited (4.9%).

4.2.2 Form Data Collection as a proxy for FDC Configuration. As described in Section 4.1.2, we computed two different measures for *Meta Pixels*: form data **configuration** and form data **collection**. We proceed by explaining the relationship between those two measures and show that collection is a suitable proxy for configuration.

For each website with detected FDC configuration, we looked at corresponding instances of detected form data collection to validate that websites that we categorized as configured to collect data did have a measurable instance of data collection. Of the 7,849 websites with FDC configuration, 89.8% had *Meta Pixel* form data collection, detected by inspecting network events; the remaining 10.2% were false negatives. We further validated that specific PII fields (email, phone number, etc.) in the FDC configuration were the same fields identified in the form data collection network events.

To understand why 10.2% of websites had FDC configuration and no form data collection, we took a uniformly random sample of 260 websites from the population of websites where our FDC configuration and form data collection results disagreed. After removing websites that either no longer had a *Meta Pixel* or no longer had FDC configuration, we evaluated a sample of 219 websites. We found that 36.1% of those websites did in fact have *Meta Pixel* form data collection. We also observed that 17.8% of our sample websites had multiple *Meta Pixels* installed, where at least two trackers had FDC configuration discrepancy (i.e. at least one *Meta Pixel* had FDC configuration and at least one did not). As mentioned in Section 3, this may lead to unpredictable form data collection behavior, which could have contributed to our error rate. Assuming independent samples, we believe that for around a quarter of the websites that had FDC configuration but no form data collection, the error can be attributed to errors in our form data collection methodology described above, rather than errors in our FDC configuration analysis.

In the opposite direction, we found that only 2 websites with detected *Meta Pixel* form data collection did not have a detected FDC configuration (0.28%).

We conclude that form data collection detected based on injected forms is highly correlated with FDC configuration and thus can be used as a suitable proxy to draw conclusions about website tracker configurations. In the following analysis, we will use form data collection as a proxy for both *Meta Pixel* and *Google Tag* FDC configuration, enabling us to compare between the two since we do not have FDC configurations for *Google Tag*.

4.2.3 Effect of Accepting or Rejecting Cookies on Tracker Installation and Form Data Collection. Our decision not to interact with cookie consent banners means that we were unable to measure trackers that do not load until a visitor accepts cookies. In order to determine the impact this might have had on our measurements, we took a random sample of 371 websites from our dataset with no detected

Meta Pixel or *Google Tag* installation. From this sample, we excluded any websites that did not load or had been erroneously classified and identified 146 websites with cookie consent banners. We accepted all cookies on these websites and ran the same process of data collection and analysis described above.

We found that 43.8% of websites had *Meta Pixel* tracker installations after accepting cookies, 24.0% had *Meta Pixel* FDC configurations and 20.5% had *Meta Pixel* form data collection. These percentages are higher for tracker installation, and only slightly higher for FDC configuration and form data collection compared to our overall dataset. We found *Google Tag* tracker installations 71.9% of the time and *Google Tag* form data collection 8.2% of the time, which closely matches what we saw in our overall dataset, which is more thoroughly discussed in Section 4.3.

We used the same sample to measure what would have happened had we rejected cookies. After excluding websites that either did not allow cookies to be rejected or required a subscription in the absence of cookies, we were left with 129 websites to test. We rejected all cookies and found that none of the websites had *Meta Pixel* FDC configuration or *Meta Pixel* form data collection. We found *Google Tag* trackers on 9.3% of websites but no instances of *Google Tag* form data collection.

Since this sample was uniformly chosen, we believe with 95% confidence (assuming independence) that these percentages would apply to all websites with neither tracker (10,787) had we accepted or rejected cookie consent banners.

4.2.4 Validation Takeaway. Based on our validation, we conclude that our measurements likely provide a lower bound on the number of tracker installations, FDC configurations, and form data collections that exist on websites we visited. While all validations had an error rate no higher than 10% (except websites that had FDC configuration and no form data collection), the bound is tightest for *Google Tag* tracker installations, which had the lowest rate of miss-classification. We believe over-counting is unlikely, as we only classify a website as having form data collection if PII is identified in one of the dynamic URLs in Table 2, which specifically route to Meta and Google domains. However, it is possible websites have modified tracker configurations to no longer collect data since our initial visit.

4.3 Analysis

We visited popular websites to measure the configuration and form data collection of tracker installations. We guide our analysis of these measurements with the following research questions:

RQ3. How prevalent is form data collection for *Google Tags* and *Meta Pixels*?
RQ4. Do tracker installations in health and finance verticals have different incidences of form data collection than non-sensitive verticals?
RQ5. What types of PII are tracker installations configured to collect?

4.3.1 How prevalent is form data collection for Google Tags and Meta Pixels? Table 3 shows a breakdown of tracker installations and form data collections for Meta and Google.

We detected a *Google Tag* installation on 72.6% of websites, and a *Meta Pixel* on 28.2% of websites. There was considerable overlap between the two. When there was a *Meta Pixel*, 98% of the time there also was a *Google Tag*; only 0.6% of the websites we studied have only a *Meta Pixel*. In other words, *Google Tag* was by far the most present tracker, with a large drop-off to *Meta Pixel*.

However, form data collection was much more prevalent on *Meta Pixel* than *Google Tag*. We detected a *Meta Pixel* form data collection event on 62.3% of websites with a *Meta Pixel* (17.6% of all websites). In contrast, only 11.6% of websites with *Google Tag* exhibited form data collection (8.4% of all websites). That is, when a *Meta Pixel* is present, it is much more likely to perform form data collection than in the case of *Google Tag*. This observation is consistent with our analysis in Section 3, which shows that Meta recommends the use of automatic data collection and provides a configuration UI flow that requires the website administrator to actively accept or decline the configuration of this feature. By contrast, while Google also recommends the use of automatic data collection, the required setup flow does not include a prompt to configure data collection, which is off by default.

We also observe that when there is a *Meta Pixel* on a website, *Google Tag* is more likely to collect form data. Table 5 shows that out of the 18,054 websites that have *Google Tag* but no *Meta Pixel*, only 5.4% (976) perform Google form data collection, while out of the 11,083 websites that have both trackers installed, 21.7% (2,401) have *Google Tag* form data collection. This absolute as well as relative increase indicates a correlation between websites that install *Meta Pixel* and their configuration of *Google Tag* to collect form data.

We also find this correlation in the additional Logistic Regression analysis we performed, found in Appendix Table 7, which suggests much higher odds of having Google form data collection when there is a *Meta Pixel* (4.839), as well as moderately higher odds of having Meta form data collection when there is a *Google Tag* (1.903).

4.3.2 Do tracker installations in health and finance verticals have different incidences of form data collection than non-sensitive verticals? Table 4 breaks down the Google and Meta tracker installations and form data collection across health, finance, and all other non-sensitive verticals. Although other categories may be considered sensitive, such as adult or gambling websites, we differentiate between health and finance specifically as they are subject to special legal restrictions in the United States (HIPPA and Gramm-Leach-Bliley Act, respectively) and are restricted from enabling automated form data collection by Meta and Google.

We notice that *Google Tag* and *Meta Pixel* tracker installations are consistent across verticals, with *Google Tag* found on 67.5% to 75.3% of each vertical's websites, and *Meta Pixel* between 27.7% and 31.6%, respectively. *Google Tag* form data collection is consistent across verticals, ranging from 11.5% of tracker installations to 13.4%. However, there is a big difference in form data collection for *Meta Pixel*. While 68% of *Meta Pixels* in non-sensitive verticals collected form data, only 30.8% of health and 20.3% of finance websites with *Meta Pixel* installations did. This suggests that Meta is somewhat effective at preventing form data collection in these sensitive verticals. These results are further corroborated by our Logistic Regression Analysis presented in Appendix B, which shows an inverse relation between health/finance verticals for *Meta Pixel* form data collection,

		Google			Meta		
Vertical	**Websites**	**Tracker Installation**		**Form Data Collection**	**Tracker Installation**		**Form Data Collection**
		Websites	Vertical Websites	Tracker Installation	Websites	Vertical Websites	Tracker Installation
Non-Sensitive	35,113	25,471	72.5%	11.5%	9,731	27.7%	68.0%
Health	3,406	2,565	75.3%	11.6%	1,075	31.6%	30.8%
Finance	1,633	1,103	67.5%	13.4%	503	30.8%	20.3%
Total	40,150	29,137	72.6%	11.6%	11,309	28.2%	62.3%

Table 4: Breakdown of tracker installations and form data collection for Google and Meta on different verticals.

Subset	**Websites**	**Form Data Collection**	
		Google	**Meta**
Google Tag* ∩ *Meta Pixel	11,083	21.7%	62.7%
Google Tag* ∩ ¬*Meta Pixel	18,054	5.4%	—
¬*Google Tag* ∩ *Meta Pixel*	226	—	44.2%

Table 5: Breakdown of form data collection for different websites with both trackers, websites with only *Google Tag*, and websites with only *Meta Pixel*.

while it shows no statistical significance between the same verticals and *Google Tag* form data collection.

As discussed in Section 3, both Meta and Google prohibit form data collection on websites that identify as a health or finance website. However, that identification is at the discretion of the website administrator and, as discussed in Section 3, not clearly explained by either party. When reviewing our data, we found several websites configured for form data collection that clearly belonged to health or finance. Therefore, for reasons that we did not investigate, these websites likely have an incorrect vertical classification that circumvented the third party's technical restrictions. We provide a few examples of such websites by vertical and third party.

Health Meta Pixel. (i) *Avenues Recovery* [6], a network of drug and alcohol rehabilitation centers, (ii) *Benefits Checkup* [9], an information service that connects seniors to various food, housing, and medical assistance programs, and (iii) *Nugg MD* [70], a medial marijuana card provider.

Health Google Tag. (i) *National Alliance on Mental Illness (NAMI)*, a mental health organization with local support services [67], (ii) *Banner Health*, a non-profit healthcare system [7], and (iii) *Cross River Therapy*, a therapy service for children with Autism [16].

Finance Meta Pixel. (i) *After Pay* [3], a buy now pay later loan company, (ii) *Patriot Software* [71], a small business payroll and accounting software company, and (iii) *KB Card* [47], a credit card service company.

Finance Google Tag. (i) *Equifax* [23], a credit reporting agency, (ii) *Capital One* [14], a bank holding company, and (iii) *Nationwide* [68], an insurance and financial services company.

Although Meta and Google's technical restrictions are a step in the right direction, they are clearly insufficient at preventing all websites in health and finance verticals from using automatic data collection features and ensuring compliance with Google's and Meta's policies.

4.3.3 What kind of PII are tracker installations configured to collect?

When analyzing *Meta Pixel* configurations, we were able to identify the specific types of PII that *Meta Pixels* are configured to collect. We omit *Google Tag* because it can only be configured to collect email addresses automatically and requires specifying CSS or JavaScript selectors for other PII fields, which is a more technically involved task.

Meta Pixel supports the following eleven PII types (some of which are grouped together): email, phone number, first and last name, city, state and ZIP code, gender, country, date of birth and external (advertising) ID. Figure 4 in Appendix A demonstrates Meta's PII field collection interface in the setup flow.

In Section 3, we found that Meta's default pre-selection of all 11 PII fields to be a case of the *Least Private Defaults* pattern. Here, we analyze how often website administrators change these default settings based on the FDC configurations we have collected and reverse-engineered for *Meta Pixel*. We limit our results to websites with only one *Meta Pixel* installed (86.5% of all websites with *Meta Pixel*), as multiple *Meta Pixels* with different configurations can have unexpected behavior and cause inaccurate tracking, as reported by Meta [62].

We found over half of these websites (51.3%) used the default configuration, which enables a *Meta Pixel* to detect all PII fields. Table 6 shows the percentage of websites that modified the default configuration to collect specific fields.[3] When website administrators customize their configuration, they most frequently exclude the external id, date of birth, and country fields (found in only 5.4%, 5.0%, and 4.6% of custom configurations, respectively). In turn, email address, first and last name, and phone number were the most likely to be collected (48.2%, 42.4%, and 42.2%). This is aligned with Cui et al.'s finding that email was the most commonly collected PII field from website forms across categories [17].

Considering both default and custom configurations, the overwhelming majority of *Meta Pixels* with FDC configuration are configured to collect email, name, and phone number (99.5%, 93.7%, and 93.5%, respectively). We note that these three fields, especially combined, can likely identify a specific individual. Recall that Meta documentation recommends website administrators at least toggle ON collection for email and phone numbers (Table 1). Again, we observe consistency between real-world configurations of tracker installations and the instructions in the documentation.

[3] Our analysis of PII field selection by websites in the health and finance verticals roughly follows a similar pattern to the entire dataset.

Field		% FDC Configuration Websites
Default (All Fields)		51.3%
Custom	Email	48.2 %
	First and Last Name	42.4%
	Phone Number	42.2%
	City, State, and ZIP Code	38.5%
	Gender	37.0%
	External ID	5.4%
	Date of Birth	5.0%
	Country	4.6%

Table 6: Breakdown of PII field collection configurations of *Meta Pixels*. Percentages split by default and custom configurations, and relative to all websites with exactly one *Meta Pixel* configured for form data collection. 51.3% of these websites use the default configuration, which collects all fields.

4.4 Limitations

Geography. All data scraping in this study was done from Google Cloud Linux machines in the United States. Therefore, any legal protections provided in other countries, such as the European Union's GDPR, are not taken into account in this study.

Trackers. This study focused on two trackers, *Meta Pixel* and *Google Tag*. Future work could investigate similar configurations on other popular trackers.

Privacy Policies. This study did not look at privacy policies and thus did not measure the disclosure of PII form data collection to website visitors in those policies.

Landing Page Visits. This study only visited the landing pages of websites. For some websites [43], tracker installations might be less frequent on their landing pages compared to their other pages.

Webpage Forms. As demonstrated in this study, PII form data collection requires a form with at least one PII field. Our methodology used a generated form to measure whether data collection was *enabled*. We did not measure if a website actually had a PII form or if a website had real PII data leaks.

5 Ethical Considerations

Data Collection. The PII used to submit forms on each website was placeholder data generated by the authors of this paper. No real PII was collected. Further, since our form was not tied to any website infrastructure, our placeholder data is much less likely to pollute any website's real advertising ecosystem or even be sent to the website directly.

Disclosures. We disclosed observed form data collection to Meta, Google, and 119 websites we determined to be in health and finance verticals. We clarified in our disclosure letter (Appendix Section C) that we did not collect or observe any real user data but simply the potential for data collection.

6 Discussion

PII form data collection primarily exists to enable the identification of a specific individual for cross-device or, more broadly, cross-context tracking. Since people often own multiple devices, cross-context tracking is essential for creating a more complete profile of a person's activity, which often increases the value of a website visitor in targeted advertising auctions across third-party marketing tools. Our methodology, which analyzed PII form data collection from multiple perspectives, offers a deeper understanding of the mechanics behind this PII data collection than prior work.

We found that website administrators face several challenges when configuring trackers. First, the tracking documentation provided by third parties to website administrators serves a dual role as a product guide and marketing material. As a result, it contains instances of marketing language that focus on the benefits of the data collection feature at the expense of adequately explaining the details of how it works. For example, Meta introduces automatic data collection as something that can "help you optimize your Meta ads to drive better results" [57]. Similarly, Google states that "data collected helps customers understand their users' needs" [34]. Even more concerning, we uncover that both Meta and Google assert hashing PII as a legitimate privacy technique to website administrators. Although, from an economic lens it is understandable why Google and Meta continue to treat hashing as an adequate privacy solution, both the US FTC and the research community have repeatedly debunked hashing as ineffective for privacy.

Our website measurements indicate that there is likely a divergence in the frequency at which website administrators configure trackers to collect PII form data. Website administrators appear more likely to configure *Meta Pixel* to collect PII form data than *Google Tag*. This aligns with our analysis that Meta's configuration UI flow forces the website administrator to enable or disable automated PII form data collection and recommends enabling this feature, whereas it is possible to complete the *Google Tag* configuration flow without encountering a PII form data collection prompt.

As we demonstrated, and to Google and Meta's credit, they block automated PII form data collection on websites self-declared as belonging to the health or finance vertical. However, based on our website measurements, we find that websites clearly in these categories, such as drug and alcohol rehab centers and banks, have in fact installed *Meta Pixels* and *Google Tags* to collect PII form data. This indicates that some websites in regulated industries may be incorrectly declaring their verticals. We contacted 119 websites that fell into this category to notify them of a PII form data collection configuration (although we are not explicitly verifying actual collection of customer data). Four websites have followed up with an intention to review website configurations.

There are several tactics that may reduce form data collection, particularly on sensitive websites. Website administrators may need to remain vigilant about how trackers are deployed, especially when relying on an external contractor or other third party to manage tracker configurations. Additionally, there are technical defenses available to website visitors, like ad blockers, that can help prevent form data collection. However, they may degrade website functionality by triggering ad-blocker detection that withholds website content until trackers are enabled or break parts of the website by accidentally blocking key scripts. As an alternative, it is possible to proactively warn website visitors that a tracker present on the page has been configured to collect form data. We created a proof-of-concept Chrome extension, included in the paper's artifacts, that can analyze *Meta Pixel* tracking code loaded by a website and notify website visitors which PII fields, if any, will likely be sent to Meta before they fill out any form. We open-source this extension with the hopes that the community can expand it to include notifications for other popular trackers, such as *Google Tag*. Unfortunately, this mitigation is imperfect since it both places a burden on the website visitor, who may be unable to access services without interacting with a web form, and does not detect any server-side PII data collection mechanisms.

Ultimately, decreasing the use of PII form data collection requires a diverse and comprehensive set of interventions from industry, government, and the research community. While data regulation does exist in the United States, this research demonstrates that it is insufficient without further enforcement actions, particularly in improving the documentation and configuration interfaces provided by Meta and Google. Further, additional regulation in the United States is likely necessary to enforce alternatives to hashing that provide strong privacy, which would likely reduce ad revenue by making it challenging to re-identify a person. Although this study exclusively visited websites from the United States, it has applications to other jurisdictions. For example, health data is also a protected category under the European Union's GDPR [24]. Beyond regulation, we need solutions that enable privacy-preserving methods of profile linking and targeting to lessen the economic impact on advertisers and advertising networks while weening them off their fire hose of PII.

Artifacts

Artifacts used in this study, including data collection and parsing code, Google/Meta documentation, and a prototype of the Meta configuration Chrome extension are available at: https://github.com/CybersecurityForDemocracy/trackers-not-equal.

Acknowledgments

The authors used Grammarly [41] to revise all parts of the text to correct typos, grammatical errors, and awkward phrasing. We thank Surya Mattu for discussions at the start that helped guide our study and Steven Englehardt for sharing his wisdom on large-scale tracker data collection. This work was supported by the National Science Foundation (grant numbers 2151837 and 2247516), Democracy Fund, Google Cloud Computing credits gift, and a gift from Reset.

References

[1] Ghostery . 2024 . Ghostery WhoTracks.Me . https://www.ghostery.com/whotracksme. Accessed: 2024-11-30.

[2] Google . 2024. [GA4] User-provided data collection . https://support.google.com/analytics/answer/14077171?sjid=1386089463089068573-NA. Accessed: 2024-11-29.

[3] Afterpay. n.d.. Buy Now Pay Later with Afterpay. https://www.afterpay.com/en-US. Accessed: 2024-11-30.

[4] Emad Aghajani, Csaba Nagy, Olga Lucero Vega-Márquez, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, and Michele Lanza. 2019. Software Documentation Issues Unveiled. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. 1199–1210. https://doi.org/10.1109/ICSE.2019.00122

[5] Noura Alomar and Serge Egelman. 2022. Developers say the darnedest things: Privacy compliance processes followed by developers of child-directed apps. *Proceedings on Privacy Enhancing Technologies* 2022 (2022), 250–273. https://doi.org/10.1145/3543507.3583311

[6] Avenues Recovery. n.d.. Drug Rehab & Alcohol Rehab | Detox and Inpatient | Avenues Recovery. https://www.avenuesrecovery.com/. Accessed: 2024-11-30.

[7] Banner Health. n.d.. Banner Health | Health Care Made Easier in AZ, CO, WY, NE, NV, CA. https://www.bannerhealth.com/. Accessed: 2024-11-30.

[8] Paschalis Bekos, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2023. The Hitchhiker's Guide to Facebook Web Tracking with Invisible Pixels and Click IDs. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2132–2143. https://doi.org/10.1145/3543507.3583311

[9] Benefits Checkup. n.d.. Worry Less and Age Better with BenefitsCheckUp. https://benefitscheckup.org/. Accessed: 2024-11-30.

[10] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies* (2016), 237–254. Issue 4. https://doi.org/10.1515/popets-2016-0038

[11] H Brignull, M Leiser, C Santos, and K Doshi. 2023. Types of deceptive pattern. https://www.deceptive.design/types. Accessed: 2024-08-05.

[12] Chris Brown and Chris Parnin. 2021. Dark Patterns for Influencing Developer Behavior. In *Position Papers of CHI'22 "What Can CHI Do About Dark Patterns?"*.

[13] Tomasz Bujlow, Valentín Carela-Español, Josep Sole-Pareta, and Pere Barlet-Ros. 2017. A survey on web tracking: Mechanisms, implications, and defenses. *Proc. IEEE* 105, 8 (2017), 1476–1510.

[14] Capital One. n.d.. Capital One | Credit Cards, Checking, Savings & Auto Loans. https://www.capitalone.com/. Accessed: 2024-11-30.

[15] Manolis Chatzimpyrros, Konstantinos Solomos, and Sotiris Ioannidis. 2019. You shall not register! detecting privacy leaks across registration forms. In *International Workshop on Information and Operational Technology Security Systems*. Springer, 91–104.

[16] Cross River Therapy. n.d.. Life-Changing ABA Therapy - Cross River Therapy. https://crossrivertherapy.com/. Accessed: 2024-11-30.

[17] Hao Cui, Rahmadi Trimananda, and Athina Markopoulou. 2024. Understanding Privacy Norms through Web Forms. *arXiv preprint arXiv:2408.16304* (2024).

[18] Ha Dao and Kensuke Fukuda. 2021. Alternative to third-party cookies: investigating persistent PII leakage-based web tracking. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*. 223–229.

[19] Levent Demir, Amrit Kumar, Mathieu Cunche, and Cédric Lauradoux. 2017. The pitfalls of hashing for privacy. *IEEE Communications Surveys & Tutorials* 20, 1 (2017), 551–565.

[20] Andrea Downing and Eric Perakslis. 2022. Health advertising on Facebook: Privacy and policy considerations. *Patterns* 3, 9 (2022). https://doi.org/10.1016/j.patter.2022.100561

[21] Steven Englehardt, Jeffrey Han, and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies* 2018 (2018), 109 – 126. https://doi.org/10.1515/popets-2018-0006

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

[23] Equifax. n.d.. Equifax | Credit Bureau | Check Your Credit Report & Credit Score. https://www.equifax.com/. Accessed: 2024-11-30.

[24] European Data Protection Board. 2025. EDPB Data Protection Guide for SME. https://www.edpb.europa.eu/sme-data-protection-guide/data-protection-basics_en. Accessed: 2025-02-07.

[25] Todd Feathers. 2022. Facebook Is Receiving Sensitive Medical Information from Hospital Websites. =https://web.archive.org/web/20241214193606/https://themarkup.org/pixel-hunt/2022/06/16/facebook-is-receiving-sensitive-medical-information-from-hospital-websites. *The Markup* (2022).

[26] Ed Felten. 2012. Does Hashing Make Data "Anonymous"? https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2012/04/does-hashing-make-data-anonymous. Accessed: 2024-09-24.

[27] FTC. 2024. No, hashing still doesn't make your data anonymous. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous. Accessed: 2024-09-13.

[28] FTC. n.d.. Gramm-Leach-Bliley Act . https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act. Accessed: 2024-08-22.

[29] Google. 2017. About enhanced conversions - Google Ads Help. http://web.archive.org/web/20240222135103/https://support.google.com/google-ads/answer/9888656. Accessed: 2025-03-03.

[30] Google. 2017. Set up your Google tag. http://web.archive.org/web/20240612205518/https://support.google.com/analytics/answer/12002338. Accessed: 2025-03-03.

[31] Google. 2024. Data sharing settings - Analytics Help . http://web.archive.org/web/20240516105538/https://support.google.com/analytics/answer/1011397?hl=en#zippy=%2Cin-this-article. Accessed: 2025-03-03.

[32] Google. 2024. Set up First-party mode. https://developers.google.com/tag-platform/tag-manager/first-party/setup-guide?setup=manual. Accessed: 2024-11-27.

[33] Google. 2024. [GA4] User-provided data collection. https://support.google.com/analytics/answer/14077171?hl=en. Accessed: 2024-07-08.

[34] Google. n.d.. About advanced matching for web. http://web.archive.org/web/20240511004924/https://support.google.com/analytics/answer/6004245?hl=en. Accessed: 2025-03-03.

[35] Google. n.d.. Best practices to avoid sending Personally Identifiable Information (PII) - Analytics Help. https://web.archive.org/web/20250216094419/https://support.google.com/analytics/answer/6366371?hl=en#zippy=%2Cin-this-article. Accessed: 2025-03-03.

[36] Google. n.d.. chrome.webNavigation. https://developer.chrome.com/docs/extensions/reference/api/webNavigation. Accessed 2024-11-02.

[37] Google. n.d.. [GA4] Set up Analytics for a website and/or app. https://support.google.com/analytics/answer/9304153?hl=en. Accessed: 2024-08-01.

[38] Google. n.d.. Set up enhanced conversions for leads with the Google tag - Google Ads Help. http://web.archive.org/web/20250120112601/https://support.google.com/google-ads/answer/11021502. Accessed: 2025-03-03.

[39] Google. n.d.. Set up enhanced conversions for web using Google Tag Manager - Google Ads Help. https://web.archive.org/web/20240416212855/https://support.google.com/google-ads/answer/13258081?hl=en. Accessed: 2025-03-04.

[40] Google. n.d.. Tag your website using Google Ads. https://support.google.com/google-ads/answer/2476688?hl=en. Accessed: 2024-10-08.

[41] Grammarly. n.d.. Grammarly. https://www.grammarly.com/. Accessed: March 11, 2025.

[42] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*.

[43] Mingjia Huo, Maxwell Bland, and Kirill Levchenko. 2022. All eyes on me: Inside third party trackers' exfiltration of phi from healthcare providers' online systems. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society*. 197–211.

[44] Hiroki Inayoshi, Shohei Kakei, and Shoichi Saito. 2024. Detection of Inconsistencies between Guidance Pages and Actual Data Collection of Third-party SDKs in Android Apps. In *Proceedings of the IEEE/ACM 11th International Conference on Mobile Software Engineering and Systems*. 43–53.

[45] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

[46] Katie Palmer. 2024. https://themarkup.org/pixel-hunt/2024/04/19/ftc-cracks-down-on-telehealth-addiction-service-monument-for-sharing-health-data. Accessed: 2024-11-30.

[47] KB Card. n.d.. KB Kookmin Card, the people's happy life partner. https://card.kbcard.com. Accessed: 2024-11-30.

[48] Simon Koch, Manuel Karl, Robin Kirchner, Malte Wessels, Anne Paschke, and Martin Johns. 2025. The Impact of Default Mobile SDK Usage on Privacy and Data Protection. *Proceedings on Privacy Enhancing Technologies* 2025 (2025), 808–823. Issue 1. https://doi.org/10.56553/popets-2025-0042

[49] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. https://doi.org/10.14722/ndss.2019.23386

[50] Colin Lecher and Jon Keegan. 2023. Suicide Hotlines Promise Anonymity. Dozens of Their Websites Send Sensitive Data to Facebook. https://web.archive.org/web/20241231101538/https://themarkup.org/pixel-hunt/2023/06/13/suicide-hotlines-promise-anonymity-dozens-of-their-websites-send-sensitive-data-to-facebook. Accessed: 2025-02-11.

[51] Colin Lecher and Ross Teixeira. 2023. Facebook Watches Teens Online As They Prep For College. https://web.archive.org/web/20241204215341/https://themarkup.org/pixel-hunt/2023/11/22/facebook-watches-teens-online-as-they-prep-for-college. Accessed: 2025-02-26.

[52] Colin Lecher and Ross Teixeira. 2024. Suicide Hotlines Promise Anonymity. Dozens of Their Websites Send Sensitive Data to Facebook. https://themarkup.org/pixel-hunt/2024/05/15/mortgage-brokers-sent-peoples-estimated-credit-address-and-veteran-status-to-facebook. Accessed: 2024-07-01.

[53] Ada Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*.

[54] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. 2021. Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2489–2506. https://www.usenix.org/conference/usenixsecurity21/presentation/maass

[55] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–32.

[56] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What makes a dark pattern... dark? design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.

[57] Meta. 2017. About advanced matching for web | Meta Business Help Center. https://web.archive.org/web/20240628094145/https://www.facebook.com/business/help/611774685654668?id=1205376682832142. Accessed: 2025-03-03.

[58] Meta. 2017. Accurate Event Tracking with Multiple Pixels. https://web.archive.org/web/20240831034606/https://developers.facebook.com/ads/blog/post/v2/2017/11/28/event-tracking-with-multiple-pixels-tracksingle/. Accessed: 2025-03-03.

[59] Meta. 2017. Advanced Matching - Meta Pixel. https://web.archive.org/web/20240615150146/https://developers.facebook.com/docs/meta-pixel/advanced/advanced-matching. Accessed: 2025-03-03.

[60] Meta. 2017. Best Practices for Advanced Matching for Web | Meta Business Help Center. https://www.facebook.com/business/help/930861050579797?id=1205376682832142. Accessed: 2025-03-03.

[61] Meta. 2017. Set up automatic advanced matching in Meta Events Manager | Meta Business Help Center. https://www.facebook.com/business/help/1993001664341800?id=1205376682832142. Accessed: 2024-08-09.

[62] Meta. 2024. Accurate Event Tracking with Multiple Pixels . https://developers.facebook.com/ads/blog/post/v2/2017/11/28/event-tracking-with-multiple-pixels-tracksingle/. Accessed: 2024-11-27.

[63] Meta. 2024. External ID. https://web.archive.org/web/20240224042839/https://developers.facebook.com/docs/marketing-api/conversions-api/parameters/external-id/. Accessed: 2024-11-28.

[64] Meta. n.d.. About advanced matching for web. https://web.archive.org/web/20240628094145/https://www.facebook.com/business/help/611774685654668?id=1205376682832142. Accessed: 2025-03-03.

[65] Meta. n.d.. Advanced - Tracking clicks on Buttons. https://web.archive.org/web/20240911074640/https://developers.facebook.com/docs/meta-pixel/advanced. Accessed: 2025-03-03.

[66] Meta. n.d.. Set up automatic advanced matching in Meta Events Manager . https://www.facebook.com/business/help/1993001664341800?id=1205376682832142. Accessed: 2024-07-08.

[67] National Alliance on Mental Hillness. n.d.. NAMI | National Alliance on Mental Illness. https://www.nami.org/. Accessed: 2024-11-30.

[68] Nationwide. n.d.. Insurance and Financial Services Company – Nationwide. https://www.nationwide.com/. Accessed: 2024-11-30.

[69] Nick Nikiforakis, Luca Invernizzi, Alexandros Kapravelos, Steven Van Acker, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2012. You are what you include: large-scale evaluation of remote javascript inclusions. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (Raleigh, North Carolina, USA) *(CCS '12)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/2382196.2382274

[70] Nugg MD. n.d.. Get Your Medical Marijuana Card Online | NuggMD. https://www.nuggmd.com/. Accessed: 2024-11-30.

[71] Patriot Software. n.d.. Simple Payroll - Patriot Software. https://www.patriotsoftware.com. Accessed: 2024-11-30.

[72] Robert Robinson. 2018. Prevalence of web trackers on hospital websites in Illinois. arXiv:1805.01392 [cs.CY] https://arxiv.org/abs/1805.01392

[73] David Rodriguez, Joseph A Calandrino, Jose M Del Alamo, and Norman Sadeh. 2025. Privacy Settings of Third-Party Libraries in Android Apps: A Study of Facebook SDKs. *Proceedings on Privacy Enhancing Technologies* 2025 (2025), 173–187. Issue 2. https://doi.org/10.56553/popets-2025-0056

[74] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. 2022. Leaky forms: A study of email and password exfiltration before form submission. In *31st USENIX Security Symposium (USENIX Security 22)*. 1813–1830.

[75] Similarweb. n.d.. Similarweb Digital Intelligence: Unlock Your Digital Growth. https://www.similarweb.com. Accessed: 2024-11-14.

[76] Oleksii Starov, Phillipa Gill, and Nick Nikiforakis. 2016. Are you sure you want to contact us? quantifying the leakage of PII via website contact forms. *Proceedings on Privacy Enhancing Technologies* (2016), 20–33. Issue 1. https://doi.org/10.1515/popets-2015-0028

[77] Tawatchai Suksida and Lalita Santiworarak. 2017. A study of website content in webometrics ranking of world university by using similar web tool. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. 480–483. https://doi.org/10.1109/SIPROCESS.2017.8124588

[78] The FTC Office of Technology. 2023. Lurking Beneath the Surface: Hidden Impacts of Pixel Tracking. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/03/lurking-beneath-surface-hidden-impacts-pixel-tracking. Accessed: 2024-07-08.

[79] Michael Toth, Nataliia Bielova, and Vincent Roca. 2022. On dark patterns and manipulation of website publishers by CMPs. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2022, 3 (2022), 478–497. https://doi.org/10.56553/popets-2022-0082

[80] Christoph Treude, Justin Middleton, and Thushari Atapattu. 2020. Beyond Accuracy: Assessing Software Documentation Quality. arXiv:2007.10744 [cs.SE] https://arxiv.org/abs/2007.10744

[81] U.S. Department of Health and Human Services. n.d.. Health Information Privacy. https://www.hhs.gov/hipaa/index.html. Accessed: 2024-08-22.

[82] Henrique Xavier. 2024. The Web Unpacked: A Quantitative Analysis of Global Web Usage. In *Proceedings of the 20th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 183–190. https://doi.org/10.5220/0012905900003825

[83] Tetiana Zavalii and Vladyslav Shakhrai. 2023. SimilarWeb as a digital tool of competitive intelligence: Ukrainian realities. *Marketing and Digital Technologies* 7 (2023), 86–104. https://doi.org/10.15276/mdt.7.2.2023.7

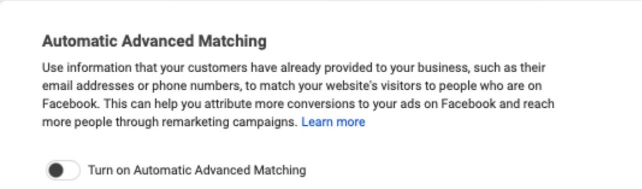
[84] Yifan Zhang, Zhaojie Hu, Xueqiang Wang, Yuhui Hong, Yuhong Nan, XiaoFeng Wang, Jiatao Cheng, and Luyi Xing. 2024. Navigating the Privacy Compliance Maze: Understanding Risks with Privacy-Configurable Mobile SDKs. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA, 6543–6560. https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-yifan

[85] Alexander R Zheutlin, Joshua D Niforatos, and Jeremy B Sussman. 2021. Data-tracking on government, non-profit, and commercial health-related websites. *Journal of general internal medicine* (2021), 1–3.

A UI and Network Screenshots

Here we include figures that demonstrate the configuration interfaces and network events discussed in Section 4.

Figures 3 and 4 illustrate how a website administrator is presented with form data collection in the setup flow of *Meta Pixel,* first with a UI prompt to toggle on, and then with the resulting auto-selection of all available PII fields after taking the prompt. Figure 5 provides a snapshot of where these options are reflected in the generated configuration code of *Meta Pixel.*

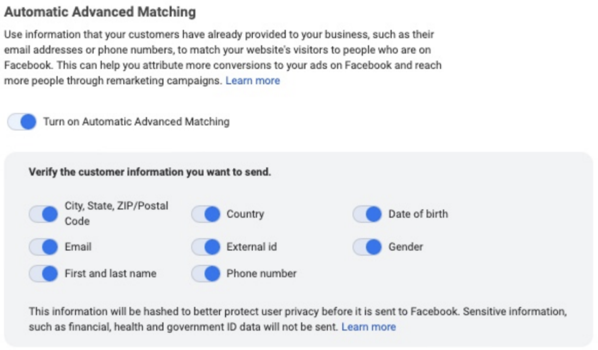


Automatic Advanced Matching

Use information that your customers have already provided to your business, such as their email addresses or phone numbers, to match your website's visitors to people who are on Facebook. This can help you attribute more conversions to your ads on Facebook and reach more people through remarketing campaigns. Learn more

○ Turn on Automatic Advanced Matching

Figure 3: Meta UI - Data Collection Prompt



Automatic Advanced Matching

Use information that your customers have already provided to your business, such as their email addresses or phone numbers, to match your website's visitors to people who are on Facebook. This can help you attribute more conversions to your ads on Facebook and reach more people through remarketing campaigns. Learn more

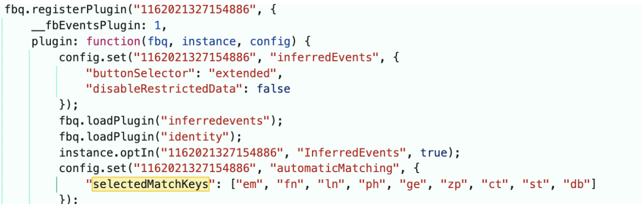
● Turn on Automatic Advanced Matching

Verify the customer information you want to send.

● City, State, ZIP/Postal Code ● Country ● Date of birth
● Email ● External id ● Gender
● First and last name ● Phone number

This information will be hashed to better protect user privacy before it is sent to Facebook. Sensitive information, such as financial, health and government ID data will not be sent. Learn more

Figure 4: Meta UI - Data Collection Fields



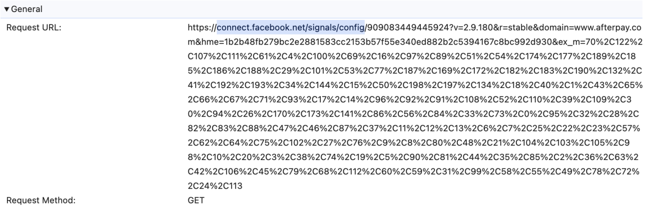
```
fbq.registerPlugin("1162021327154886", {
    __fbEventsPlugin: 1,
    plugin: function(fbq, instance, config) {
        config.set("1162021327154886", "inferredEvents", {
            "buttonSelector": "extended",
            "disableRestrictedData": false
        });
        fbq.loadPlugin("inferredevents");
        fbq.loadPlugin("identity");
        instance.optIn("1162021327154886", "InferredEvents", true);
        config.set("1162021327154886", "automaticMatching", {
            "selectedMatchKeys": ["em", "fn", "ln", "ph", "ge", "zp", "ct", "st", "db"]
        });
```

Figure 5: Meta Data Collection in the Code Configuration

Figures 6, 7, 8, and 9 illustrate the URLs and hashed PII identified by our parser to label website installations and form data collection.

B Regression Analysis Tasks

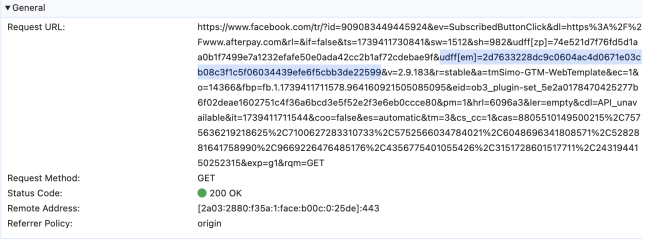
To assist us with providing these insights, we performed two separate Logistic Regression analysis tasks. Table 7 provides an overview of both of them; in the first, we trained a model to predict whether a website performs form data collection through Meta. We trained



▼General

```
Request URL:      https://connect.facebook.net/signals/config/909083449445924?v=2.9.180&r=stable&domain=www.afterpay.co
                  m&hme=1b2b48fb279bc2e2881583cc2153b57f55e340ed882b2c5394167c8bc992d930&ex_m=70%2C122%2
                  C107%2C111%2C61%2C4%2C100%2C69%2C16%2C97%2C89%2C51%2C54%2C174%2C177%2C189%2C18
                  5%2C186%2C188%2C29%2C101%2C53%2C77%2C187%2C169%2C172%2C182%2C183%2C190%2C132%2C
                  41%2C192%2C193%2C34%2C144%2C15%2C50%2C198%2C197%2C134%2C18%2C40%2C1%2C43%2C65%
                  2C66%2C67%2C71%2C93%2C17%2C14%2C96%2C92%2C91%2C108%2C52%2C110%2C39%2C109%2C3
                  0%2C94%2C26%2C170%2C173%2C141%2C86%2C56%2C84%2C33%2C73%2C0%2C95%2C32%2C28%2C
                  82%2C83%2C88%2C47%2C46%2C87%2C37%2C11%2C12%2C13%2C6%2C7%2C25%2C22%2C23%2C57%
                  2C62%2C64%2C75%2C102%2C27%2C76%2C9%2C8%2C80%2C48%2C21%2C104%2C103%2C105%2C9
                  8%2C10%2C20%2C3%2C38%2C74%2C19%2C5%2C90%2C81%2C44%2C35%2C85%2C2%2C36%2C63%2
                  C42%2C106%2C45%2C79%2C68%2C112%2C60%2C59%2C31%2C99%2C58%2C55%2C49%2C78%2C72%
                  2C24%2C113
Request Method:   GET
```

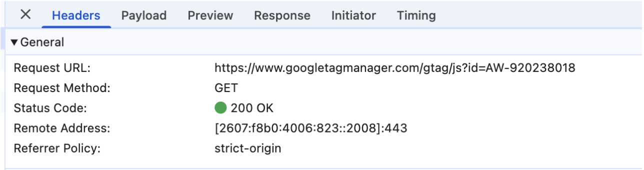
Figure 6: GET request for Meta Pixel



▼General

```
Request URL:      https://www.facebook.com/tr/?id=909083449445924&ev=SubscribedButtonClick&dl=https%3A%2F%2
                  Fwww.afterpay.com&rl=&if=false&ts=1739411730841&sw=1512&sh=982&udff[zp]=74e521d7f76fd5d1a
                  a0b1f7499e7a1232efafe50e0ada42cc2b1af72cdebae9f&udff[em]=2d7633228dc9c0604ac4d0671e03c
                  b08c3f1c5f06034439efe6f5cbb3de22599&v=2.9.183&r=stable&a=tmSimo-GTM-WebTemplate&ec=1&
                  o=14366&fbp=fb.1.1739411711578.964160921505085095&eid=ob3_plugin-set_5e2a0178470425277b
                  6f02deae1602751c4f36a6bcd3e5f52e2f3e6eb0ccce80&pm=1&hrl=6096a3&ler=empty&cdl=API_unav
                  ailable&it=1739411711544&coo=false&es=automatic&tm=3&cs_cc=1&cas=8805510149500215%2C757
                  5636219218625%2C7100627283310733%2C5752566034784021%2C6048696341808571%2C52828
                  81641758990%2C9669226476485176%2C4356775401055426%2C3151728601517711%2C24319441
                  50252315&exp=g1&rqm=GET
Request Method:   GET
Status Code:      ● 200 OK
Remote Address:   [2a03:2880:f35a:1:face:b00c:0:25de]:443
Referrer Policy:  origin
```

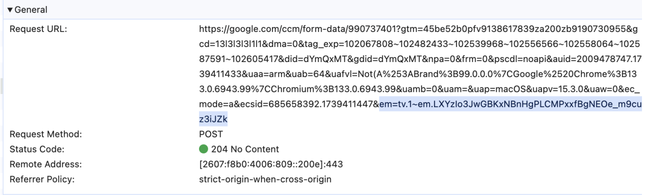
Figure 7: Form Data Collection Event for Meta Pixel



▼General

```
Request URL:      https://www.googletagmanager.com/gtag/js?id=AW-920238018
Request Method:   GET
Status Code:      ● 200 OK
Remote Address:   [2607:f8b0:4006:823::2008]:443
Referrer Policy:  strict-origin
```

Figure 8: GET request for Google Tag



▼General

```
Request URL:      https://google.com/ccm/form-data/990737401?gtm=45be52b0pfv9138617839za200zb9190730955&g
                  cd=13l3l3l3l1l1&dma=0&tag_exp=102067808~102482433~102539968~102556566~102558064~1025
                  87591~102605417&did=dYmQxMT&gdid=dYmQxMT&npa=0&frm=0&pscdl=noapi&auid=2009478747.17
                  39411433&uaa=arm&uab=64&uafvl=Not(A%253ABrand%3B99.0.0.0%7CGoogle%2520Chrome%3B13
                  3.0.6943.99%7CChromium%3B133.0.6943.99&uamb=0&uam=&uap=macOS&uapv=15.3.0&uaw=0&ec_
                  mode=a&ecsid=685658392.1739411447&em=tv.1~em.LXYzIo3JwGBKxNBnHgPLCMPxxfBgNEOe_m9cu
                  z3iJZk
Request Method:   POST
Status Code:      ● 204 No Content
Remote Address:   [2607:f8b0:4006:809::200e]:443
Referrer Policy:  strict-origin-when-cross-origin
```

Figure 9: Form Data Collection Event for Google Tag

this dataset on only websites that already have a *Meta Pixel*. We used as features four boolean variables; the first is true when a website has a *Google Tag*, the second when a website has Google form data collection, the third when a website is in the Health vertical, and the fourth when it is in the Finance vertical.

The second model was trained on only websites that have a *Google Tag*, and trained on three features; the first feature is true when the website has a *Meta Pixel*, and once more the two features that signify whether the website belongs in the sensitive Health or Finance vertical, respectively. We do not use for this model a feature to signify Meta form data collection because there was a very high correlation between this feature and the existence of a *Meta Pixel* (0.71 Pearson's correlation coefficient). We kept the feature that was a better predictor.

It is important to acknowledge that our models are a relatively weak fit; the former has a pseudo R-squared of 0.0721 and the

	Feature	OR	p-value	CI
Meta	Has *Google Tag*	1.903	0.000	[1.443, 2.512]
	Google Form Data Collection	1.699	0.000	[1.533, 1.885]
	Is Health	0.206	0.000	[0.180, 0.237]
	Is Finance	0.118	0.000	[0.094, 0.147]
Google	Has *Meta Pixel*	4.839	0.000	[4.473, 5.233]
	Is Health	0.952	0.457	[0.835, 1.084]
	Is Finance	1.086	0.376	[0.905, 1.305]

Table 7: Odds Ratios (OR), p-values, and Confidence Intervals (CI) from our Logistic Regression Analyses:
(i) With *Meta Pixel* form data collection as dependent variable; trained on all websites that have *Meta Pixel* (Pseudo R-squared: 0.0721). Results suggest that a website is more likely to have *Meta Pixel* form data collection when it has *Google Tag* and *Google Tag* form data collection, and less likely if it belongs to Health or Finance verticals.
(ii) With *Google Tag* form data collection as dependent variable; trained on all websites that have *Google Tag* (Pseudo R-squared: 0.0825). Results suggest that a website is more likely to have *Google Tag* form data collection when it has a *Meta Pixel.*
All features are boolean variables – True if a website has the property.

latter has a pseudo R-squared of 0.0825. This means that their explanatory power is limited. However, we can still draw some useful insights from them especially if we can combine them with our measurements.

B.1 Meta Form Data Collection Model Results

The upper half of Table 7 presents the results for our Meta form data collection model. When a website has a *Google Tag* then it is 1.903 times more likely, and when it has Google form data collection it is 1.699 times more likely to have Meta form data collection. We also see that the odds ratios for Health (0.206) and for Finance (0.118) suggest that it is 79.4% less likely for a website to have Meta form data collection when it belongs in the Health vertical, and 88.2% less likely when it belongs to the Finance vertical.

B.2 Google Form Data Collection Model Results

Regarding Google form data collection, we see in the lower half of Table 7 that according to our model, when a website in our training dataset has a *Meta Pixel*, it is 4.839 times more likely to have Google form data collection. In addition, we note that belonging to either Health or Finance is not a statistically significant feature for predicting Google form data collection.

C Notification Template

We identified 119 websites that had form data collection configured and we conservatively believed to be in a Health or Finance vertical. We contacted them through emails found, in order of preference, as the technical email associated with the domain, an IT support contact, a general contact, or a PR contact email. We used the following template for notification.

To Whom It May Concern,

I am [name and affiliation]. Our research team is studying third-party tracker configurations on [health | finance] websites. Based on the products and services detailed on your website, we believe you can be classified as a [health | finance] website.

We believe that your website includes a [Google | Meta] tracker (with ID [TRACKER ID]) that is configured to collect [emails | visitor data, including emails] and send them to [Google | Meta]. This configuration may be set up through the [Google | Meta] user interface. However, websites categorized as [health | finance] are not permitted to use this collection feature by [Google | Meta]. Therefore, we believe trackers included in your website have been mis-categorized during setup.

As it is currently configured, the tracker(s) may collect data and share it with [Meta | Google] when a user visits your website and fills out a form, such as a login, contact, or subscription form. We did not specifically verify that your website has a form requesting this information from visitors, and all of our testing was done with fabricated data; we did not view any real customer information. Testing was completed between September and November 2024.

Sincerely, [name] [contact email]