Measuring the state of web privacy : Adoption of Google Tag Manager in E-Commerce websites and its Implications on Privacy

Vishaka Iyengar

Master of Science
School of Informatics
University of Edinburgh
2025

Abstract

"This research project investigates the privacy implications of Google Tag Manager (GTM) implementation on e-commerce websites. By developing an automated parser that analyzes network traffic, consent mechanisms, and tracking behaviors, we analyze the adoption rate of GTM, its Consent Mode, third-party tracking prevalence, and data collection practices in the form of events. Drawing on established methodologies for privacy measurement, our study will provide empirical evidence regarding GTM's potential role in facilitating privacy-invasive practices across the e-commerce ecosystem. We aim to deliver a comprehensive research findings and also an open-source tool enabling ongoing privacy audits of e-commerce platforms. This research addresses critical gaps in understanding how tag management systems may fundamentally alter the privacy landscape of online shopping environments" [10].

Research Ethics Approval

Instructions: This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Vishaka Iyengar*)

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr. Daniel Woods, for his constant support from ideation to creation. Through his invaluable experience, prompt feedback, and patience he helped me create a realistic project and gave me room to explore and grow all the while providing guidance and unwavering support.

I would also like to thank Ryan Chadwick, from Coalition Inc., for providing insights from his combined expertise as a software developer, cyber security enthusiast, researcher and a member from the industry.

I thank my parents for believing in me when I didn't. Without their unwavering support, warmth, and patience this journey would have been unbelievably difficult.

Lastly, I'd like to thank my friends for being a constant source of emotional support, understanding and most often, a sounding board that helped me generate my best ideas.

Table of Contents

1 Introduction **1**

2 Background and Literature Review **5**

2.1 Privacy and Its Presentation in Commerce 5

2.2 E-commerce and Online Tracking 6

2.3 Privacy Conscious Consumers in the E-commerce Market 8

2.4 Google Tag Manager (GTM) 9

2.5 Privacy and Tracking in GTM-Enabled Websites 10

3 Methodology **12**

3.1 Introduction 12

3.1.1 Parameters 12

3.2 Data Collection 13

3.3 Web Scraping 14

3.3.1 High-Level System Architecture 14

3.3.2 Web Crawler 15

3.3.3 Ethical Considerations 18

3.4 Preprocessing 18

3.5 Statistical Analysis 19

3.6 Methodology Limitations 20

4 Results and Discussion **21**

4.1 GTM Detection 21

4.1.1 Adoption of GTM based on Popularity Rankings 22

4.1.2 Relationship between Number of Events Tracked and Popularity Ranking 23

4.2 Consent Mode + Event Detection 25

4.2.1 Adoption of Consent Mode based on Popularity Distribution . 25

4.2.2 Relationship between consent mode and event tracking 27
4.2.3 Relationship between Consent Mode and Third-Party Trackers 30
4.3 Summary 32

5 Conclusions **34**
5.1 Summary 34
5.2 Limitations and Future Work 35
5.3 Reflections 36

Bibliography **37**

A Technology **47**
A.1 Technology Stack 47

B Participants' information sheet **48**

C Participants' consent form **49**

Chapter 1

Introduction

As e-commerce continues to expand globally, consumer privacy concerns have emerged as a significant barrier to market participation. Research indicates that privacy-conscious users are particularly hesitant to engage in online shopping transactions, creating tension between commercial interests and the need for privacy protection [3, 85]. Although numerous studies have documented tracking practices throughout the wider web ecosystem [12, 27, 16, 24], a relatively small pool of research exists on how tag management systems influence tracking prevalence and behaviour [73].

Tag Management Systems (TMS) have fundamentally transformed the deployment of tracking technologies on websites. Google Tag Manager (GTM), introduced in 2012 [29], has become particularly influential by allowing non-technical staff to implement sophisticated tracking mechanisms without developer involvement. Despite its growing popularity, the research landscape of privacy and security in the implementation of tag management systems has yet to be explored [73]. The democratization of tracking implemented by the TMS raises essential questions about its impact on user privacy, particularly in data-intensive environments such as e-commerce. Thus, this project studies Google Tag Manager (GTM) and its implications for privacy in the e-commerce industry.

The e-commerce context presents unique privacy considerations that warrant a focused study. E-commerce sites routinely process sensitive personal and financial information during transactions [34]. Most e-commerce websites have mechanisms to save user data, such as cards and addresses, to their accounts. Further, customer journeys contain multiple high-value interaction points (product browsing, cart additions, checkout) that create incentives for extensive tracking [87]. The customers need to be satisfied that their data will not be leaked due to the organizations carelessness or

misused by being sold to third-party vendors [11, 3, 85]. Since organizations must comply with privacy regulations or face lofty fines, institutions face a conundrum. By acting in complete support of user privacy, they stand to gain user trust, loyalty, and confidence. They also stand to lose on an incredibly lucrative opportunity of studying their consumer base, customizing their advertising and products, and providing a much better customer experience. On a darker note, they also stand to lose on the enormous data they have access to, which many other organizations pay for to gain a competitive advantage. A widely accepted solution is the use of dark patterns to coerce users to obtain their consent to proceed with these profitable endeavours [44, 94]. Moreover, the complexity of e-commerce implementations often creates substantial gaps between stated privacy policies and actual data collection practices. With feeble enforcement of these regulations against dark patterns, e-commerce platforms consider this a tacit acceptance of their practices.

To further complicate matters, the regulatory landscape presents presents a labyrinthine challenge in which compliance obligations change dramatically across jurisdictional boundaries. While the GDPR and CCPA establishes comprehensive privacy protections, its extraterritorial reach creates ripple effects throughout global e-commerce ecosystems, compelling global commerce entities to restructure their data practices. Privacy requirements that may be systematically undermined by GTM's technical capabilities. This jurisdictional patchwork necessitates that digital merchants navigate not only these flagship regulations but also the nuanced privacy requirements of each operational territory, while simultaneously establishing clear delineations between data processor and controller responsibilities, and legitimate interests and their boundary with consent management. The play of consent management platforms, granular consent, and privacy by design within the sphere of commerce are nuanced territories. They are often places where small and medium-sized enterprises can go wrong.

Enterprises deploying products that involve third-party data collection must ensure compliance with GDPR and mitigate risks related to legal liability, data security, and privacy. This requires understanding the roles of data controllers and data processors, as compliance obligations differ. Accurate characterization of providers, along with thorough due diligence, is essential. The consumerization of IT, including employees using personal devices (BYOD), further complicates these decisions, as enterprises increasingly adopt consumer products where the provider acts as a controller [49, 62]. Similarly, the GDPR makes a distinction and provides two separate legal bases for processing between legitimate interest and data collection that requires obtaining user

consent. E-Shops must have a strict legal/contractual justification for collecting user data without consent. Ambiguity in how or why these data are being processed can be grounds for compliance breaches [28, 62]. In a similar vein lies the granular consent. Each data processing purpose must have separate, distinct consent and a general consent ”for any personal data processing operation” is not legally valid. Granular consent dictates that individuals have the right and means to select to which purposes they consent and reject others individually [35].

Recent research by Mertens et al. (2024) concluded that 42% of the top 1 million websites adopted the use of Google Tag Manager (GTM) [73]. They also identified concerns about privacy within GTM, including data leaks and potential legal violations. Their groundbreaking work established that GTM-enabled websites frequently collect personal data without adequate disclosure and may continue tracking even when users decline to give their consent. The popularity of this tool makes it urgent to investigate its compliance and usage. Building on their methodological foundation, our research explicitly examines how these issues manifest in the e-commerce sector, where the stakes for consumer privacy are particularly high.

The primary research questions we address are:

1. Can we reliably create a tool to detect the adoption of GTM and its Consent mode?

2. What proportion of e-commerce websites adopt GTM, and how does GTM adoption correlate with website popularity rankings?

3. What is the distribution of websites, based on popularity ranking, that use the Consent mode of GTM? The adoption of GTM Consent mode increases significantly in websites with a popularity ranking of greater than 10k, with the adoption rate crossing 90%.

4. Do websites with consent mode enabled track different types of events than those without, and does the use of Consent mode impact the number of events tracked by a website?

5. Is there a threshold effect for tracker deployment - do websites with consent mode suddenly increase tracker usage after reaching a certain popularity level?

Our findings will have significant implications for multiple stakeholders:

- Website operators will gain insights into how their implementation choices affect user privacy
- Privacy advocates will receive evidence-based assessments of current practices
- Consumers will obtain greater transparency about data collection during online shopping

Our analysis examines how technical infrastructure shapes privacy outcomes independently of user intentions or regulatory frameworks. GTM represents a critical case study in how ostensibly neutral technologies can systematically undermine privacy protections through design choices that prioritize commercial utility over user control. Moreover, by releasing our GTM parser as an open-source tool, we will enable ongoing monitoring and verification of e-commerce privacy practices beyond the scope of this initial study. This contribution addresses a critical need for accessible technical tools to evaluate privacy compliance in increasingly complex web environments.

The remainder of this dissertation is organized as follows:

Chapter 2 provides an in-depth review of the backdrop of privacy in the e-commerce setting and users' concerns. Then we explore the technical aspects of Google Tag Manager (GTM) and its connection to web tracking and privacy. We get a thorough understanding of the privacy landscape in e-commerce forums that use GTM.

Chapter 3 explains the entire process of performing the project, the parameters, system architecture, the methodology, considerations, limitations, and the statistical tests run to ensure that the results we generate are relevant beyond the bubble of data we have collected.

Chapter 4 discusses the results of the statistical tests we have run, delivering the results and exploring them in the backdrop of relevant literature.

Chapter 5 summarizes the project and answers the research questions in brief, highlighting the contribution.

Chapter 6 reflects on the limitations of the project and provides direction for future work in this field.

Chapter 2

Background and Literature Review

2.1 Privacy and Its Presentation in Commerce

In 1948, privacy was recognized as a fundamental Human Right by the United Nations, without ever clarifying what it precisely meant [21]. This motif found its place throughout literature, the definition of privacy remained generic, ranging from "the right to be left alone" [99] to "Privacy is the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others." [101]. Regardless, it is indisputable that user privacy must be safeguarded.

The value of privacy is clear from a policy perspective; however, the actual implementation is often completely skewed. With the increase in digitization of mundane activities, all the way from socializing, to paying taxes, to buying life insurance or even socks, this shift of commerce to the web has resulted in unprecedented amounts of user data being stored on servers, processed by Machine Learning algorithms, studied using data analytics - all to weave a story and understand the user better than they possibly know themselves [82, 100]. Using these data lakes, a new "big data economy" emerged - one that is fueled with statistics, plotting trends, recognizing behaviors, and piecing together identities via IP addresses [7]. The age of "On the Internet, no one knows you're a dog." is long gone [86]. Commerce clearly has a financial incentive to capitalize on this opportunity, using all the tools in its arsenal to track and sell to the consumer at every given opportunity.

Further, there's the growing pain of increasingly aware users, who notice how often their searches for hair products yield targeted advertisements for hair treatments, medications, pills, and a hundred related products sent to their browsers, on shopping

websites, and even in their texts. The Privacy Report of 2019 by Kaspersky found that 56% of the surveyed audience believed that it was impossible to keep data truly private online and that 46% had experienced at least one data breach [58]. Privacy-conscious individuals are frequently confronted with the dilemma of either seeking a privacy-preserving approach to achieve seemingly simple tasks, such as reading a blog on hair loss, or compromising their privacy values to complete the task at hand. This everyday battle to defend their right to privacy leads to disengagement from privacy-preserving choices, causing them to perform actions that do not align with their original intent of preserving privacy. This is the privacy paradox, and its result is privacy fatigue [98, 48]. The loss of control experienced adversely affects their customer experience, indicates signs of calculated coercion, which undermines trust in the brand and the overarching system, and ultimately sacrifices customer loyalty [97].

While extensive research has documented the privacy paradox and fatigue in digital commerce [98, 48], the specific role of tag management systems like GTM in facilitating or obfuscating tracking practices remains underexplored [73]. This research gap has significant implications for both users and regulators. Privacy-conscious consumers may unknowingly experience extensive tracking despite explicit consent preferences, while regulatory enforcement becomes tedious when compliance violations occur through dynamic, post-consent script loading that circumvents traditional monitoring approaches.

2.2 E-commerce and Online Tracking

E-commerce started in mid-1960s when some transport-based companies tried to go paperless [92]. Research spanning from 1999 [3] to 2007 [84] to 2021 highlights the trajectory of growth of the e-commerce domain and that customers' privacy concerns, including risk and benefit perceptions, negatively impact their attitude and adoption of e-commerce platforms [66]. As the user reveals more information about themselves, they enjoy a more personalized experience. Better discounts at checkout, reminders on when your bills are due, credit card offers tailored to your needs, or showing relevant search results. In the same breath as these benefits, sharing this information gives way to price-discrimination [5], potential leaks of Personally Identifiable Information or financial data, or even identity theft [4, 78].

The personalization of data ranges from e-commerce websites using metadata like the routing MacOS users to more expensive hotel suites (price steering) to using past

search history to artificially inflate prices of flights (price discrimination) to a complete profiling and increasing the cost of health insurance due to unhealthy lifestyles [45]. This metadata is captured via tag management systems like GTM and then passed on to the end point applications like Google Analytics [36]. Conducting online behavioral analysis and then specifically targeting users with advertisements, coupled with price discrimination by individualizing the offers to customers makes the e-commerce forum highly lucrative [74, 13]. As Cambridge Analytica's co-founder-cum-whistleblower Christopher Wylie said, "We exploited Facebook to harvest millions of people's profiles and built models to exploit what we knew about them and target their inner demons. That was the basis the entire company was built on." [6] Dark Pattern designs use UI to coerce, deceive, and manipulate the users into performing actions that benefit the organization instead of the user [67] and numerous studies have demonstrated this, right from teenagers being targeted with personalized adverts during times of vulnerability to Uber drivers being coerced into serving longer hours for the cheapest ride rate to maintain Uber's competitive advantage [88].

Numerous studies conducted over the years document the efficacy of various kinds of browser fingerprinting, using first and third party cookies, site origin policies (SOP), Cookie Synchronisation, cross device tracking, and even using benign features such as browser dimensions and fonts to formulate a unique fingerprint for each device across phones, tablets, and computers [25, 80, 15, 63, 2]. These elaborate measures gather data that is then not just used for analytics by the parent site, but also sold to data markets and advertisers, giving the bidders/owners a competitive advantage [80]. This interconnection of data now weaves a story. A consumer searches about a mundane health condition - say low stamina and fatigue, gets influenced into worrying about it, starts receiving subliminal messaging via social media about health hacks and"influencer life hacks", starts seeing advertisements about health supplements and lifestyle changes to overturn their life, and possibly being displayed a higher insurance base rate than 6 months ago, due to AI being used without appropriate training leading to proxy discrimination [76]. All of this is possible because of a curious search, data being collected and transferred across multiple advertising markets, a need being identified, a vision/dream being sold, resulting in purchases and payments over the months. A lead is converted into a sale - and we have the water-like flow of data across Internet, with the users "consent" of it to thank.

Despite extensive documentation of tracking mechanisms, there is a significant gap in privacy research on the impact of the dynamic nature of tag management systems,

particularly GTM's ability to load trackers conditionally based on consent states. While studies document tracker prevalence, they fail to examine how modern tag management systems may enable more sophisticated consent circumvention strategies.

2.3 Privacy Conscious Consumers in the E-commerce Market

The 2020 edition of the Cost of a Data Breach Report by the Ponemon Institute covering 524 organizations across 17 countries and regions, and 17 industries to provides a global average. They found that 80% of the breached organizations reported that customer Personally Identifiable Information (PII) was compromised and the average cost of a stolen record was only $150 [53]. That is the value of a customer's privacy online.

Studies have shown that users who value their privacy more are less likely to engage in e-commerce transactions. The study also outlines how different personality traits influence people's ability to gauge online shopping privacy risks [11]. With privacy concerns as a barrier to e-commerce, let us explore how people get tracked.

Most websites track user data, including searches, geolocation, products bought, etc, mainly for targeted advertising. However, when shared with third parties, this data can also influence the prices of things (think flight prices after you've searched for them), insurance costs and coverage, identity theft, or surveillance. Third-party tracking, Log File Data, metadata collection, first-party and third-party cookies, website fingerprinting, inappropriately long session tokens, and more are regularly used to create a user profile [17].

Research on privacy-conscious consumers has primarily focused on individual behavioral responses to tracking [11], overlooking the technical infrastructure that enables or prevents effective privacy protection. Studies examine user consent behavior and privacy paradoxes [48], but fail to investigate how tag management systems may undermine user privacy choices through technical implementation gaps. This represents a critical oversight, as privacy-conscious consumers may unknowingly be tracked despite their explicit consent preferences when websites use dynamic tag loading systems.

2.4 Google Tag Manager (GTM)

Tagman introduced the first-ever Tag Management System (TMS) in 2007 [54]. By 2012, Google Tag Manager (GTM) entered the market [29].

A page tag is a piece of JavaScript code that is configured to trigger at the occurrence of an event and send data from the client to the analytics program. This code can be embedded into the source code of the webpage and needs a developer to embed it, or it can be added by a non-technical data analyst or marketer as well, using a tool like GTM [9]. GTM provides numerous facilities :

- Update and add tags without touching the source code, thus reducing the dependency on developers for an analyst's needs.
- Easy to track tags to avoid duplication, which can skew the data for business analysts.
- Includes debugging, testing, version control, and preview tools.
- Can implement access control
- . . . And all that is needed to start is a Google account! [61]

Now, there is a distinction to note here. GTM itself is a script management system, but the actual trackers it loads (Facebook, GA, etc.) are the tracking technologies. GTM acts as a hub for data collection and distribution to multiple endpoints. Furthermore, exporting and importing tags is a simple process that does not require a deep coding background. There is ample documentation to facilitate the sharing of tag codes [20, 29]. Thus, using GTM, websites can load third-party trackers dynamically, avoid detection if scripts load after consent is given or via obfuscated tags, and have less transparency unless a full script and traffic analysis is conducted. Thus, GTM can facilitate third-party tracking and website fingerprinting.

GTM's technical capabilities and implementation methodologies are well-documented in industry literature, academia, and online forums, however academic research has largely overlooked its privacy implications. The system's ability to dynamically load tracking scripts, modify tracking behavior post-consent, and operate with minimal transparency on what happens to the collected data creates significant gaps in current privacy protection frameworks.

2.5 Privacy and Tracking in GTM-Enabled Websites

We have already established via multiple studies that third-party tracking is omnipresent on the internet. Research demonstrates the extensive presence of tracking across the internet, with Mayer finding that 86% of the top 500 websites deploy third-party cookies, some connecting to over 100 third-party domains [69]. This tracking ecosystem has grown increasingly sophisticated, including concerning practices like using Flash cookies to respawn deleted HTTP cookies on 107 of the top 10,000 sites [1]. The challenge for privacy-conscious users intensifies as tracking methods evolve. Research revealed fingerprinting on 404 of the Alexa top million sites, with 95% of instances originating from just three companies [26]. More concerning still is the shift toward server-side tracking, which occurs on web servers rather than in browsers, making it nearly invisible to users and circumventing most privacy tools through direct server-to-server communication [32]. This evolution from client-side to server-side tracking represents a significant challenge for privacy advocates, as these technologies can persist despite user privacy efforts, potentially deterring e-commerce participation and raising concerns about data monopolization.

The pervasiveness of tracking is further complicated by dark patterns in consent mechanisms. Research by Gray et al. demonstrates how consent banners employ manipulative design elements that undermine genuine user consent, with 95% of analyzed banners using at least one dark pattern [44]. Similarly, research also found that Consent Management Platforms (CMPs) often manipulate website publishers through complex implementation processes that favour privacy-invasive defaults [94]. These practices exist within a regulatory landscape established by frameworks like GDPR and CCPA, yet compliance remains problematic. Urban et al. revealed that 92.6% of websites failed to provide valid revocation mechanisms for previously given consent [57], while Trevisan et al. documented that 54% of websites load trackers before obtaining explicit consent [56]. This compliance gap is particularly concerning in the context of Google Tag Manager (GTM), as Sanchez-Rola et al.'s research demonstrates that websites using GTM frequently load tracking scripts before obtaining user consent, with 82.3% of analyzed websites loading privacy-invasive elements during the initial page load. Their study further revealed that GTM implementations enable widespread fingerprinting techniques on 65.7% of sites and that websites using GTM deployed an average of 3.4 times more trackers than those without it [73]. The dual role of CMPs as both processors and potential controllers creates additional regulatory concerns [81], especially when

Field studies by Utz et al. demonstrate the ineffectiveness of consent notices, with position and design choices significantly affecting user interaction and the alarming finding that dark patterns in consent interfaces can increase consent rates by up to 23 percentage points [96].

While these studies provide valuable insights into tracking prevalence and consent manipulation, they treat GTM primarily as another tracking vector rather than examining its unique role as a tracking orchestration platform. The studies mentioned above collectively indicate a significant gap between regulatory intent and implementation reality in online privacy protections, particularly for GTM-enabled websites. It is crucial to explore this gap given GTM's widespread adoption [73] and its potential to fundamentally alter how privacy regulations are implemented and circumvented.

Chapter 3

Methodology

3.1 Introduction

This project aims to assess the impact of using Google Tag Manager (GTM) in e-commerce websites on user privacy. We undertake an Experimental and Statistical Analysis approach to analyze the gathered data and draw inferences; thus, our methodology is Quantitative.
We sample 4500 websites from the top 10,000 websites based on the Top-N approach using a web parsing bot. This approach is ideal for this project, as it allows us to study the web at large by sampling websites from different popularity rankings from the top 10k e-commerce websites. We detect and tabulate whether the sampled websites use GTM, if the websites that do use it also employ its Consent mode, events that GTM tracks, which specific trackers are deployed by GTM, and which third-party domains are installed within it. We first explain the parameters that we detect and why they matter.

3.1.1 Parameters

- **GTM Detection**- we first categorize websites on the basis of whether or not they use GTM.

- **Consent** - we identify whether or not websites that use GTM employ its Consent Mode. Google Tag Manager provides Consent mode as a feature using a Consent Initialization Trigger [43]. This trigger is put in place to ensure that user consent is honoured, and is used for tags that update or set the user consent state for a website. However, this trigger is not applicable on tags that fire before GTM

loads on the page.

- **GTM Events** - Events are actions or occurrences on a website. Examples of events include - loading a page, scrolling on a page, clicking on a link/product, adding to a basket, making a purchase, or even noting a page crash [40].
- **Third-Party Trackers** - refers to the trackers from another website that have been integrated into the website that is being visited. These trackers are usually present for the purpose of advertising or analytics [31].
- **Third-party domains** - refers to the domains that have been integrated with the GTM container to send/receive data. These domains are not inherently a part of the parent website, but are called by the tags that have been created for the purpose of analytics, tracking, advertising etc.
- **E-commerce Platform** - It is a software or a service that used to create and run stores on the internet. Like cloud services, they have many models. Some are complete packages that will host your store for you, while some are software platforms that must be hosted and run individually [89].
- **Best Popularity Rank** - It is a the popularity rank that has been provided by the Chrome User Experience Report [90]. The report is a dataset that endeavours to reflect how users experience different websites on the Internet, and ranks the websites accordingly [18].
- **Google URLs count** - it is a cumulation of the number of Google services and URLs that are called by a website upon the homepage loading. It hints at the depth of the website integration with Google services.

3.2 Data Collection

Given the projects focus on the e-commerce domain, we required a systematic way to categorize websites as e-commerce and create a working, ground-truth dataset. We could have sampled URLs directly from Tranco, an academically well-established and used forum [64]. However, the Tranco list contains domains from all categories, and accurately determining which were e-commerce websites would be a difficult task [77]. To address this, we use the HTTP Archive (via Web Almanac/BigQuery) as our seed sample.

The HTTP Archive is an open-source project and a sub-project of the Internet Archive. The project has been systematically collecting data on millions of websites since 2010 and have made it publicly accessible through the Google BigQuery database. The HTTP Archive also publishes an annual state-of-the-web report called the Web Almanac. The Almanac documents its collection methodology and published metrics [46]. Using the HTTP Archive as our seed ensured that our starting sample was both comprehensive and well-documented.

From this seed dataset, we extracted only e-commerce websites by querying the "httparchive.crawl.pages" table in BigQuery. We adapted an existing SQL query provided by prior researchers, available on their GitHub project, to specifically select e-commerce URLs, their associated e-commerce platforms, and their popularity rank [51, 47].

The Web Almanac integrates Wappalyzer to detect the technology stack used in each URL and the Chrome User Experience Report to gather their best popularity rank [90]. Using these metrics, we compiled a refined list of the top 10,000 e-commerce websites ranked by popularity. Finally, we sampled from this list in batches of 1,500 URLs, selecting entries from the beginning, middle, and end of the ranking distribution. This served as the working dataset for subsequent analysis, including Google Tag Manager detection and related measurements.

3.3 Web Scraping

3.3.1 High-Level System Architecture

The GTM Parser is a project hosted on Docker containing 3 Python programs that work in tandem with one another. The simple_detector.py file contains all the detector modules and is integrated with a Gostery Tracker DB to store verified valid trackers. The progress_manager.py is a file that handles batch processing during the data collection phase. It offers flexibility in the batch size and the number of batches that need to be run, and outputs the data consolidated in a single CSV file. The program writes to that CSV file after each batch has been executed, this in case of a crash or unexpected error, the data upto the previous batch is still saved and available to use. Lastly, there is the main.py file that contains the code to run the program. It also contains a few test batch scenarios and has custom commands for the same. The output of this program is a single flat-file artifact - CSV file.

3.3.2 Web Crawler

We addresses the implementation of the web scraper in `simple_detector.py`. The simple detector is divided into modules through which we detect the parameters discussed in 3.1.1. We document the logic, the singularities, and the considerations that went into designing and implementing each of these modules.

3.3.2.1 GTM Detector

We detect the implementation of GTM by parsing through the users browser's Performance API to get network requests. The detector parses through Network Requests and identifies dataLayer objects that contain data that needs to be forwarded to GTM [37]. It is mandatory to create a dataLayer for to use GTM on web pages, making it the ideal object to detect GTM. As the Google documentation highlights the dependency of GTM on the existence of the dataLayer, the possibility of having False Positives with this mode of detection is null[39].
This dataLayer object contains 3 main events, namely gtm.js - for the container's script loading on the page, gtm.dom - the page being DOM ready, and gtm.load - for the page loading DOM events, in this order [93]. Thus, the parser identifies URL's containing common GTM string matches like 'googletagmanager.com', 'gtm.js', 'gtag/js', '/gtm?id=', and 'gtm-' [37]. Further, we also detect GTM Container IDs by matching string patterns in the network requests.

3.3.2.2 Consent Mode Detector

The Consent Mode detector module checks the `dataLayer` for HTTP request parameters such as `&gcs=`, `&consent=`, `&gcd=`, `&npa=`, `&pscdl=`, `consent%3d`, `gcs%3d`, `gcd%3d`, `npa%3d`, `pscdl%3d`, `consent_state`, `analytics_storage`, and `ad_storage`. Regardless of whether consent mode is enabled on the webpage, the `gcd` parameter is always sent to Google services as it contains user consent choices' data [38].

3.3.2.3 GTM Event Detection

As the dataLayer contains all GTM events [38], the detector filters the dataLayer for objects that contain an event, maps it to the event value, and returns a list.

```
const events = window.dataLayer
    .filter(item => item && item.event)
```

```
.map(item => item.event);
```

3.3.2.4 Third-Party Trackers

The third-party tracker detector architecture is split into 6 major categories.

The Detector. It tests if the Ghostery Tracker DB is loaded. It uses analyzes network usage to detect the use of GTM, and if yes, it times when GTM was loaded. The detector also detects the different trackers from network requests. Using timing correlation between when GTM was loaded and when the trackers were loaded, the detector generates a score for each tracker, quantifying the likelihood that the tracker was called by GTM. Then, it uses the Ghostery Tracker DB that had been loaded to identify legitimate trackers. Lastly, it returns both tracker names and domain names as two separate lists.

Ghostery TrackerDB Integration. To ensure that the Ghostery tracker DB is accessible at all times, we design the following 3 layered approach:

1. GitHub Releases API - access the tracker online and download the latest TrackerDB

2. Local Cache - once downloaded, the tracker is considered as a valid cache for 24 hours. If the 24 hours have expired, then a new TrackerDB file will be downloaded. In case the download fails, the cache files will be used as a fallback. After 1 week, the cache files are deleted.

3. Docker Volume Mounted ZIP - as a backup, in case of a missing cache file and the tracker can not be accessed online.

Caching System. The program uses the latest available cache to legitimize and map the trackers. In case of an expired cache, a new tracker will be downloaded from GitHub. If the download fails, the old cache is used for upto a week. After 1 week, the cache is deleted. In case there is no cache available and the download fails as well, the program falls back to a local ZIP file that contains the TrackerDB.

Time-Based GTM Attribution System. This system was designed to understand which trackers could be attributed to GTM vs being called by the parent website itself. The tracker system is as follows :

TrackerDB Data Structure and Parsing. A JSON structure is expected and a look up table is constructed. It handles domain to tracker mappings efficiently and handles

Time Window	Scoring	Confidence
Before GTM load	0.0	Impossible
0–5 seconds after GTM	0.9	Very high confidence
5–15 seconds after GTM	0.8	High confidence
15–30 seconds after GTM	0.6	Medium confidence
30+ seconds after GTM	0.3	Low confidence

Table 3.1: Scoring and confidence levels based on time elapsed since GTM load.

multiple domains per tracker.

Tracker Identification Algorithm. We've designed a multi-level domain matching system that handles exact domain matches, subdomain matches, and partial matches such as analytics.facebook.com contains facebook as well.

3.3.2.5 Testing

We test the code by manually reviewing the Network Traffic and tabulating the data. We then cross verify the results by using the Wappalyzer chrome extension to validate the use of Google Tag Manager. We then compare the results of the manual tests with the results generated by the web scraper. This manual verification is done on a list of randomly selected 30 websites based on daily usage. There are some test commands that are hard coded into the main.py file, such as :

- python main.py –test # Original 4 test URLs
- python main.py –comprehensive # Comprehensive 13 URLs
- python main.py –batch-test # 300 URLs (batches 1-3)
- python main.py –full-ecommerce # Full 10k e-commerce analysis

3.3.2.6 End-to-End Workflow

The project runs in a Docker container with mounted data. The web scraping pipeline starts with using Async Playwright in stealth mode with a Chromium browser. The stealth mode helps avoid some cases of anti-scraping measures. We also introduce random delays (2-5s) between requests and realistic user agents and headers.

The analysis pipeline first loads the URLS from the "2025-06-01 10k Unique e-commerce websites" file. It then initializes the Ghostery Tracker DB. Within that

module, it checks if there is a recent version of the DB that has been cached. The cache lasts for 24 hours before being deleted. The websites are sequentially processed, in batch sizes that have to be specified at run time. We detect GTM, Consent Mode, Third-Party Trackers and Domains and Google URL counts via Network Monitoring. Once third-party trackers have been detected, they are identified by matching with the Ghostery Tracker DB patterns. Finally, the result is stored in a CSV format.

3.3.3 Ethical Considerations

In conducting this research, we employed ethical web scraping methodologies, including verification of `robots.txt` compliance, to ensure that data collection was performed in a respectful manner. To further minimize disruption and avoid anti-scraping measures, we configured Playwright in Stealth Mode. We also introduced human-like delays of a few seconds and simulated natural interactions such as page scrolling. These human-like movements not only reduce the likelihood of detection but also help trigger Google Tag Manager (GTM), which on some websites is activated only after specific user interactions.

3.4 Preprocessing

Since the project has been developed and run on a personal computer, there are RAM limitations. Thus, to ensure that the scraping occurs smoothly, and to debug and not lose progress in case of a failure, we create a simple progress manager that allows us to divide the entire data set into customizable batch sizes, and run for a customisable number of batches. We run the code for batches of 100, 50, and 25 URLs each as needed. The result is a miminum of 15 csv files containing data for each set.

These csv files contain the data obtained from scraping websites. These parameters are :

url, gtm_detected, consent_mode, gtm_events, third_party_trackers,
third_party_domains_count, third_party_domains_list,
trackerdb_patterns_count, trackerdb_data_source, status,
google_urls_count, analysis_time, timestamp, and raw_urls.

We then stitch the files together by running a code that will append the rows. Thus we create 3 files of 4500 URLs total. We then add the parameters of ecommerce platform and best popularity rank from the original data set to each of these 3 files, to

create the final 3 sets.

Lastly, we combine set 1, 2, 3 to create the Final Combined Data Set of 4500 URLs.

3.5 Statistical Analysis

1. Binomial test : A Binomial Test is a null hypothesis test that is used to achieve a binary result by weighing the probabilities of an event happening [70].

2. Chi-square Test of Independence : This test produces outputs that are independent of data distribution that analyzes the difference between the dependent variable and the independent variable. Chi-square can be used in studies where ”parametric assumptions cannot be met”. Cramer’s V is the most common strength test within a Chi-square test, where a higher number is a better fit [71].

3. Chi-square Goodness-of-Fit Test : This test that determines how well sample and theoretical distributions match. It is flexible as it can be used for any distribution. It assesses the discrepancy between observed and expected frequencies in categorical data and helps determine if the observed data is likely to have come from a population following a specific distribution [95, 19].

4. Fischer’s Test : This test also evaluates the relationship between two variables when the parameters being compared are independent.

   The chi-square test produces better results using large samples, while Fischer’s exact test produces results for small sample sizes [60].

5. Logistic Regression : It is a linear model that is used to study the relationship between independent and dependent variables when the output is segregated into categories and not continuous. ”Logistic regression identifies a curve to map input values to a probability.” We use statistical tests such as the t-test and ANOVA to assess which variables have the most significant impact on the results. A p-value ≤ 0.5 means that the variable has a significant impact on the results [50, 52].

6. Mann-Whitney U Test : Also known as a Wilcoxon Rank Sum Test. It is a nonparametric method for comparing two independent groups on a single ordinal variable without assuming any specific distribution. The null hypothesis for the Mann-Whitney U test states that both samples originate from the same population. To evaluate this, data from the two groups are pooled, ranked, and then split

back into their respective groups. Each group's ranks are summed, and these totals (along with sample sizes) are used to calculate the U statistic. The test accommodates tied ranks by assigning adjusted values, ensuring that all forms of ranking can be handled consistently [72].

3.6 Methodology Limitations

The data set is created using a larger set of data collected by the HTTP Archive on 1 June 2025. However, the HTTP Archive themselves acknowledge certain limitations in recognizing e-commerce sites. Since they rely on Wappalyzer for detection, the tool can only identify e-commerce platforms if they are explicitly recognized. Moreover, because the analysis is restricted to the home page, websites where the e-commerce functionality resides in subpages are excluded [91]. As a result, our data set omits some of the largest e-commerce platforms such as *amazon.com*, *ebay.com*, and *etsy.com*, which often rely on custom-built platforms not detected by Wappalyzer.

A second limitation concerns the sample selection method. We use a Top-N approach, which naturally creates a bias toward the most popular websites. While this bias aligns with our research focus on highly influential e-commerce websites, it reduces the representativeness of the broader ecosystem. Future studies could instead consider a more uniformly distributed sampling strategy to capture a wider variety of websites.

Finally, when testing the most popular e-commerce websites such as Amazon, the automated code failed to detect the presence of GTM, despite manual testing confirming its use. This shortcoming stems from the strong anti-bot measures employed by such platforms, which could not be circumvented even with Playwright's Stealth Mode.

Chapter 4

Results and Discussion

Having gathered the data, we conduct statistical tests to understand the impact of GTM on user privacy by comparing our results with established literature. This chapter is divided into the following categories to understand the individual and combined impact of these parameters. Section 4.1 explores the adoption of GTM in the e-commerce industry. Section 4.2 explores the adoption of consent mode based on website popularity and its impact on the number of events tracked. Section 4.3 explores the relationship between third-party trackers and consent mode. In Section 4.4 we summarize the results.

4.1 GTM Detection

Figure 4.1 shows that 3,192 out of 4,500 e-commerce websites (70.9%) were detected to use GTM, while 1,308 websites (29.1%) did not use GTM. To test whether this observed proportion differs significantly from random adoption, we conduct a binomial test with two possible outcomes: ”uses GTM” vs ”does not use GTM”. With a fixed sample size of 4,500, each website represents an independent trial, being unrelated to the detection of GTM in another URL. Considering a null hypothesis of 50% adoption rate, we find the observed 70.9% adoption rate is statistically significant ($p < 0.001$), indicating that the observed proportion is highly unlikely to occur by chance alone. This suggests a clear preference among e-commerce websites to adopt GTM as a tag manager, rather than random adoption patterns. This finding aligns with prior research by Mertens et al., who established that 42% of the top 1 million websites had adopted the use of GTM [73], though their data includes websites from all categories, suggesting that e-commerce sites show significantly higher GTM adoption rates than the general web.

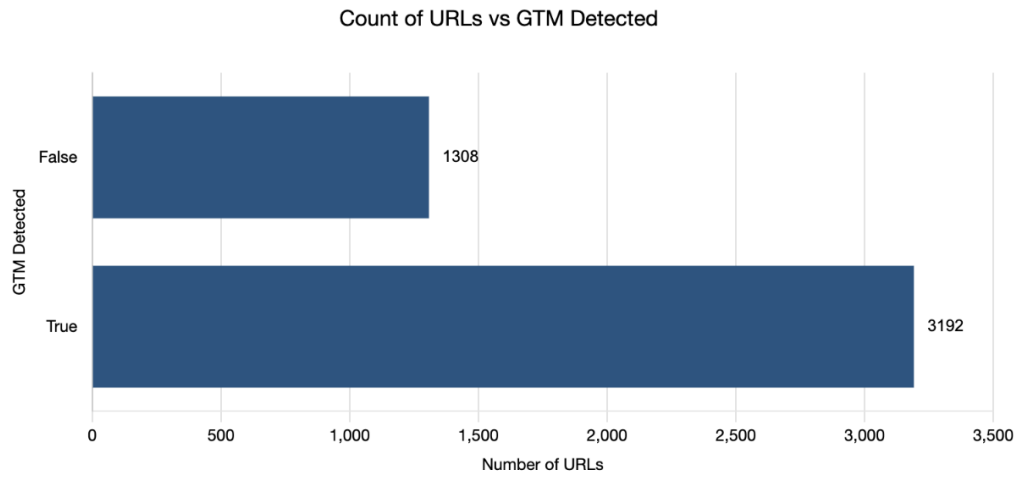

Figure 4.1: Number of URLs using GTM

4.1.1 Adoption of GTM based on Popularity Rankings

The next logical question that arises is whether the popularity rank of websites has any impact on their likelihood of adopting GTM. To understand the statistically significance, with results transcending just the output obtained from our current data, we conduct a Chi-square Test of Independence. We have two categorical variables:
Variable 1: Website Popularity (categorical: popularity rankings - 1-1k, 1k-5k, 5k-10k, 10k-50k, 50k-100k, 100k-500k).
Variable 2: GTM Usage (categorical: Uses GTM/Does not Use GTM)
The Chi-square tests if there is a statistical correlation between GTM adoption and website popularity ranking, thus making the test of Independence appropriate for this question/null hypothesis.

As described in Fig 4.2, there is a systematic relationship between size/popularity and technology choices. We find the p-value is < 0.001 denoting high statistical significance and Cramér's V value of 0.061 which implies a small effect. Further, less popular websites use GTM more than popular ones.

The result and the Cramer's V value suggest that popular sites may have resources for custom analytics solutions, while smaller sites rely on GTM as an accessible, plug-and-play solution. Large corporations, the likes of Amazon and Capital One compete for competitive advantage by deploying "industrial-strength analytics". They invest in developing custom technology solutions, adopting highly customizable tools, hiring experts in niche fields, integrating strategy in a top-down approach, and as a result churn out profits [22]. Thus it is likely that highly popular websites use Google Tag Manager

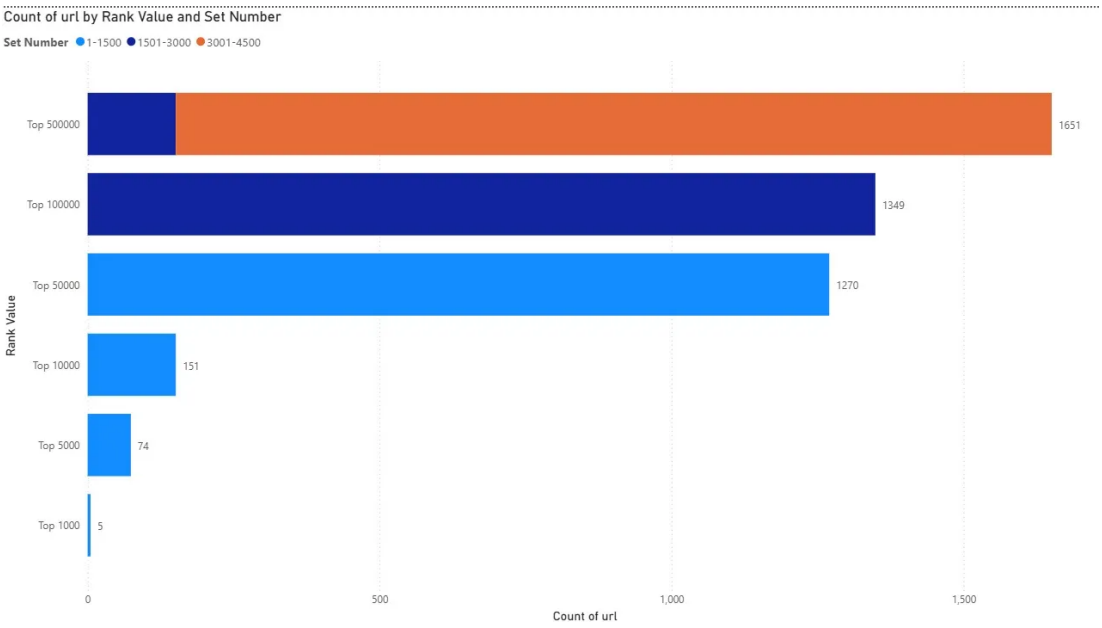

Figure 4.2: URL Popularity Ranking vs GTM Adaptation

as a solution less often than less popular websites.

Conversely, smaller e-commerce sites demonstrate greater reliance on GTM as an accessible solution. Digital technologies, including digital artifacts, platforms, and infrastructure, are instrumental in enhancing the accessibility of entrepreneurship and business ventures to a broader demographic [75]. Several determinants influence the adoption of technology, such as the perceived relative advantage of its implementation, compatibility with existing backend systems, the endorsement and support from top management, prior experience among personnel in utilizing the technology, features that facilitate operational efficiency, industry standards, supplier initiatives, and external computing support [8]. The Infrastructure, e-commerce platforms, data analytics, and even tag management systems as-a-Service model are easier to integrate and implement than developing a solution from scratch, which is a cost, resource and Info-Tech intensive job. Consequently, SME's favor plug-and-play solutions, exemplified by the adoption of Google Tag Manager (GTM) for tag management purposes.

4.1.2 Relationship between Number of Events Tracked and Popularity Ranking

To understand the relationship between the mentioned parameters, we consider the data as we have collected it - in 3 sets. We then calculate the average number of events per website:

- Set 1 (first 1500): 3.81 events
- Set 2 (middle 1500): 4.56 events ← HIGHEST
- Set 3 (last 1500): 2.99 events ← LOWEST

So the ranking is: $Set2 > Set1 > Set3$.
However, this alone is not sufficient to draw conclusions as our distribution of data of websites by popularity rank are not evenly spread across all 3 sets. Thus while the sets are organized by decreasing popularity rank, not each set has the same number of URLs of each rank.

Thus, we conduct a Welch ANOVA test to handle unequal variances in the 3 sets. It handles continuous dependent variable (number of events tracked) without forcing artificial categorization. It is more likely to detect true differences between groups compared to non-parametric alternatives when dealing with reasonably large samples.

The p-value for our study is less than 0.001, showing that our findings are extremely statistically significant. Since our websites were organized in terms of rankings, this shows that there is no evidence that more popular websites are likely to track more events through GTM.

On surveying the literature on event logs from websites, we find numerous uses for event tracking and analysis. A 2025 study highlights how small and medium businesses can perform process mining by extracting events triggered and captured with timestamps to reverse engineer business processes that also conform with industry standards. The authors prove that process mining increases efficiency, permitting real-time auditing that gives SME owners a competitive advantage [83].

Another study explores using Complex Event Processing (CEP) for SME's so that they don't have to interpret complex web analytics themselves. They simply have to collect events that contain real-time user interaction data and send it to the analytics platform. The platform then performs automated actions that can range from sending a notification to triggering changes dynamically onto the web page, based on user interaction. Such interactive e-shops enhance sales [79].

Such articles can reasonably lead us to believe/speculate the following reasons for the results we have achieved :

- Websites in set 2 track more events to gain a competitive advantage and expand their business.

- Websites in set 1 might have a more sophisticated event triggering mechanism, firing only on certain user activities.

Due to the inconclusive reasoning for the results, further studies must be carried out on a larger scale, and also with a focus on captured events and their types and implications.

4.2 Consent Mode + Event Detection

Consent Mode is designed to gather alternative signals from visitors who decline the use of personal data or browser storage for tracking. Google then applies these signals to estimate conversions (in tools like Google Ads and Floodlight) and to model user activity in Google Analytics 4 [41]. This raises some interesting questions about the adoption of Consent Mode in e-commerce websites. Thus, we explore the the proportion of websites that use AND do not use Consent mode of GTM.

Based on the data we have scraped, disregarding popularity ranking for a moment, we discover the following:

Parameter	**Group A (GTM=True, Consent=True)**	**Group B (GTM=True, Consent=False)**	**Group C (GTM=False)**
Total websites	2876	316	1308
Percentage of total	63.91%	7.02%	29.07%

Table 4.1: Distribution of websites by GTM and consent mode

4.2.1 Adoption of Consent Mode based on Popularity Distribution

A consequent line of inquiry that opens up is - are e-commerce websites with a higher popularity rank more likely to use GTM consent mode? We explore this question using a Logistic Regression as the outcome is binary (GTM Consent Mode = True/False). Since websites have popularity ranks, they can be treated as ordinal predictors to test for a linear trend across the popularity levels.

We obtain an Odds Ratio of 1.29, which is much higher than 1, indicating a positive association between the predictor and the outcome. Meanwhile the Pseudo R^2 value is

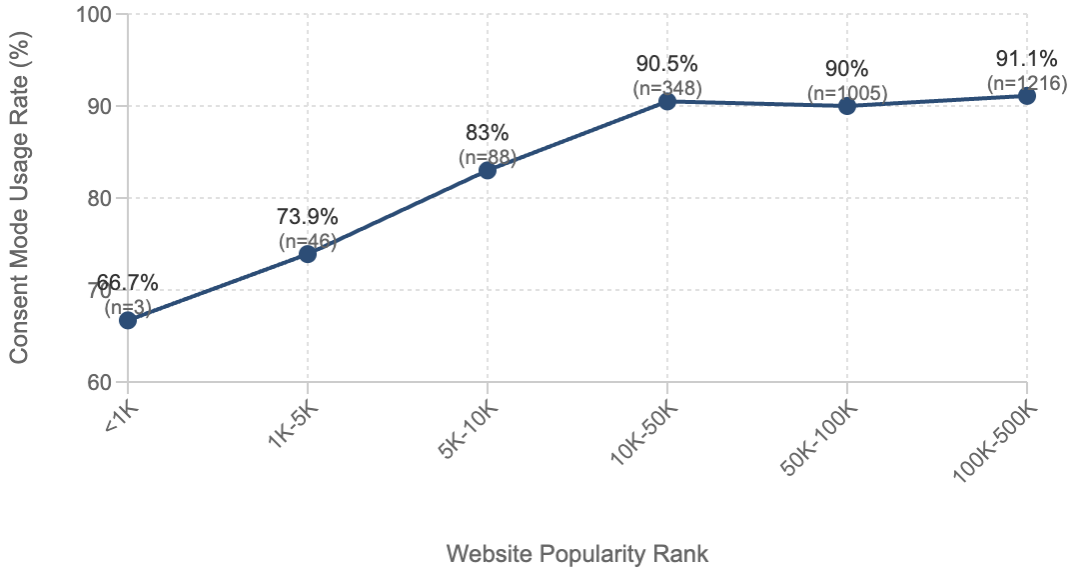

Figure 4.3: Consent Mode Adoption Rate by Website Popularity Rank

0.0024 which is a very poor fit as it explains only 0.24% of the results. These results determine that websites with lower popularity rank are more likely to use Consent mode.

As seen in Figure 4.3, the adoption of consent mode increases with decreasing popularity ranking. More than 90% of websites with with a popularity rank of 10k and above opt to use Consent Mode. This is quite interesting, as Consent mode is not the default setting for GTM [42].

The following arguments can be made to support this finding. Firstly, Google Tag Manager makes it increasingly easy to adopt the use of Consent mode. There is ample documentation provided by Google [42], there exists a sea of research on the use GTM to boost analytic prowess, and given its widespread use, there are numerous aides available online. Secondly, websites understand that if they adopt Consent mode, they can enable higher rates of event tracking and third-party tracking. This can be attributed to the way consent mode works. If a user declines consent, then the cookie data is simply not shared with the website, thus abiding by privacy laws. The users feel safeguarded and assured that their privacy is valued by the performance in the Consent Theater [30]. However, this is not always the case. There exists a gap between the user consent status that is received and mechanisms in place to honor it and to work around it [68]. Thirdly, websites can employ dark patterns to increase the likelihood of receiving user consent, by confusing them or obfuscating the option to decline or by simply not displaying it or by making declining a tedious process [44, 94, 67, 65].

Lastly, websites with a lower popularity rank, or SME's generally have a limited budget to spend on technology. Hence, we saw that plug-and-play solutions like GTM were increasingly popular among them. We also understand that this lack of budget means they have to comply with Privacy Laws such as the GDPR and other local laws while having little to invest in information technology. Thus, we can assume that one possible reason for this result could be that the organizations simply misunderstand the Consent Mode to be a "privacy compliance checkbox" and a legal cover [65].

4.2.2 Relationship between consent mode and event tracking

To understand the relationship between websites with consent mode enabled and the events being tracked by them, we first sort our data to identify the GTM events that are most popular in websites that use consent mode and those that do not.

Statistic	**Group A (With Consent Mode)**	**Group B (Without Consent Mode)**
Websites with events	2876	316
Total event occurrences	15605	1433
Unique events	1434	247
Average events per website	5.43	4.53

Table 4.2: Comparison of event statistics between websites with and without Consent Mode.

Fig 4.4 and Fig 4.5 highlight the top 10 events by count and frequency in websites that use and do not use consent mode.

Based on this data, we try to gauge how the relative frequencies of events differ between consent mode categories. The most appropriate test to explore this difference is the Chi-Square Goodness-of-fit Test. It checks if the distribution of event types in the "WITHOUT Consent Mode" group follows the same pattern as the "WITH Consent Mode" group. The test handles unequal sample sizes appropriately considering that we attempt to compare distributions.

As visible from Fig 4.6, there is a statistically significant difference in event tracking patterns between consent groups ($p = 0.018$). While most individual events occur at similar rates, the overall distribution differs, primarily due to increased scroll depth tracking (2.85% more) in websites with consent mode.

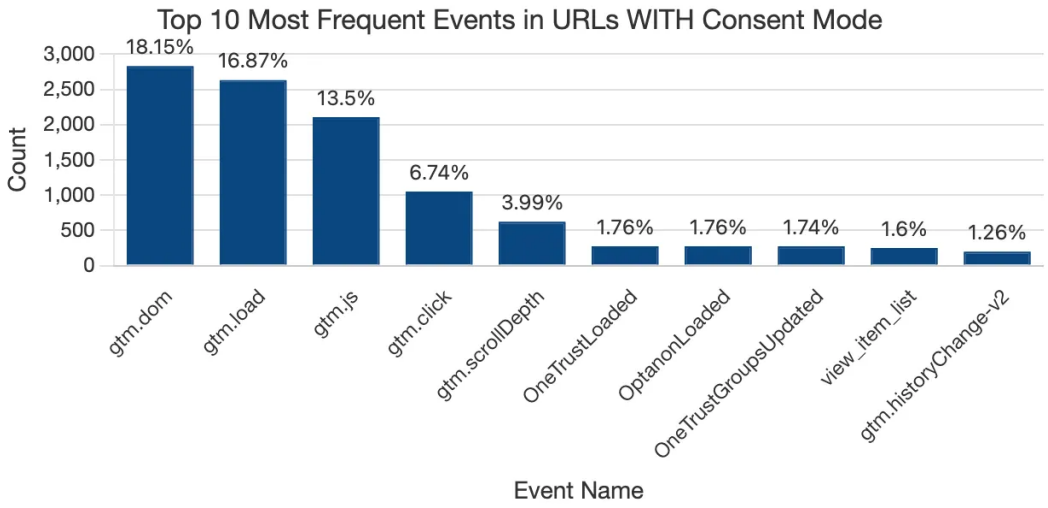

Figure 4.4: Top 10 Frequent Events in URLs With Consent Mode

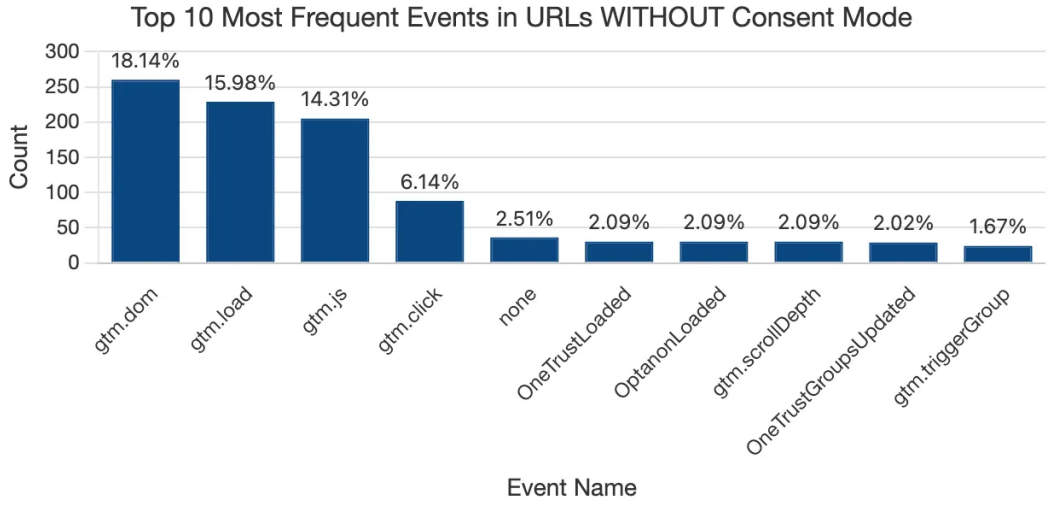

Figure 4.5: Top 10 Frequent Events in URLs Without Consent Mode

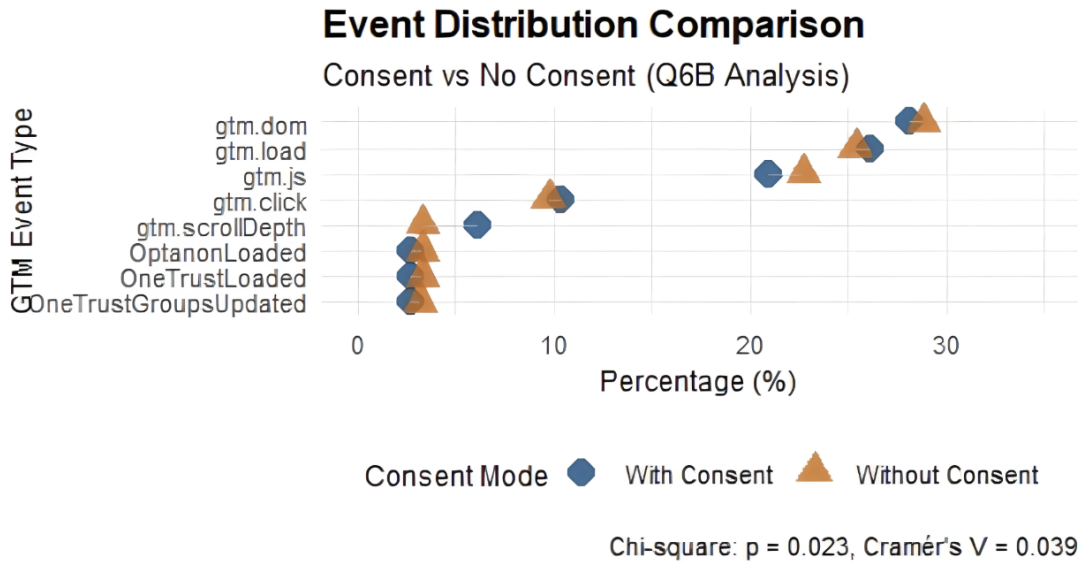

Figure 4.6: Event Distribution Comparison : Consent Mode True vs False

The statistical test confirms the overall pattern is different, even though not every single event is dramatically different. Although the difference is statistically significant it is possible that it has been affected heavily due to outliers, namely the gtm.scrollDepth event and the gtm.js event. Further studies need to be conducted with a larger sample size to see if this result is replicated.

With Consent Mode	**Without Consent Mode**
`gtm.dom` - 2833 times (18.15%)	`gtm.dom` - 260 times (18.14%)
`gtm.load` - 2632 times (16.87%)	`gtm.load` - 229 times (15.98%)
`gtm.js` - 2106 times (13.5%)	`gtm.js` - 205 times (14.31%)
`gtm.click` - 1051 times (6.74%)	`gtm.click` - 88 times (6.14%)
`gtm.scrollDepth` - 622 times (3.99%)	none - 36 times (2.51%)
`OneTrustLoaded` - 274 times (1.76%)	`OneTrustLoaded` - 30 times (2.09%)
`OptanonLoaded` - 274 times (1.76%)	`OptanonLoaded` - 30 times (2.09%)
`OneTrustGroupsUpdated` - 272 times (1.74%)	`gtm.scrollDepth` - 30 times (2.09%)
`view_item_list` - 249 times (1.6%)	`OneTrustGroupsUpdated` - 29 times (2.02%)
`gtm.historyChange-v2` - 197 times (1.26%)	`gtm.triggerGroup` - 24 times (1.67%)

Table 4.3: Event counts with and without Consent Mode

Lastly, we attempt to understand the impact of using Consent mode on the number of events tracked by a website. We use the Mann-Whitney U Test as it is robust against outliers, and compares medians rather than means, which is more appropriate for non-normal data.

While the sample size of websites with consent mode (2876) vs without (316) may skew this number, it is still statistically significant as the p-value is less than 0.001 and r-value is 0.091. We find that websites with consent mode collect data on more events than websites without consent mode.

This counterintuitive finding suggests that consent mode implementation may legitimize rather than restrict tracking activities, representing a 20% increase in data collection despite user expectations of enhanced privacy protection. The result challenges assumptions about privacy compliance tools, indicating they may enable ”privacy theater” where consent mechanisms provide legal cover for expanded rather than reduced tracking [14].

4.2.3 Relationship between Consent Mode and Third-Party Trackers

We attempt to quantify the likelihood of consent mode being deployed based on the number of trackers called by GTM on a website. To quantify this, we use Logistic Regression as the outcome is binary (GTM Consent Mode - True/False). The relationship between number of trackers and consent mode probability is non-linear. As tracker count increases, the probability of using consent mode doesn't increase linearly. Lastly, the predictor variable is the number of trackers which is easily handled by logistic regression.

Parameters	**Value Found**	**General Range**	**Meaning**
P-value	0.001	low value < 0.05	High statistical significance
Z-value	6.656	high value > 2.0	High statistical significance
AUC (Area Under Curve)	0.674	$1 > AUC > 0.5$, higher is better	High statistical significance
OR (Odds Ratio)	1.068	OR $> 1 \rightarrow$ predictor increases odds	High statistical significance
95% CI (Confidence Interval)	1.048–1.090	> 1	Statistical significance

Table 4.4: Statistical parameters and their interpretation.

The logistic regression's p-value, z-value, and AOC reveals a highly significant relationship between third-party tracker count and consent mode. Each additional tracker increases the odds of consent mode implementation by 6.8%, with websites having 10+ trackers showing 90% probability of consent mode deployment. While the model demonstrates significant predictive capability above random chance, the moderate AUC (0.7) and low explained variance (Pseudo $R^2 = 0.026$) indicate that tracker count provides useful but limited prediction accuracy, reinforcing the counterintuitive pattern where privacy compliance tools correlate with expanded rather than restricted tracking

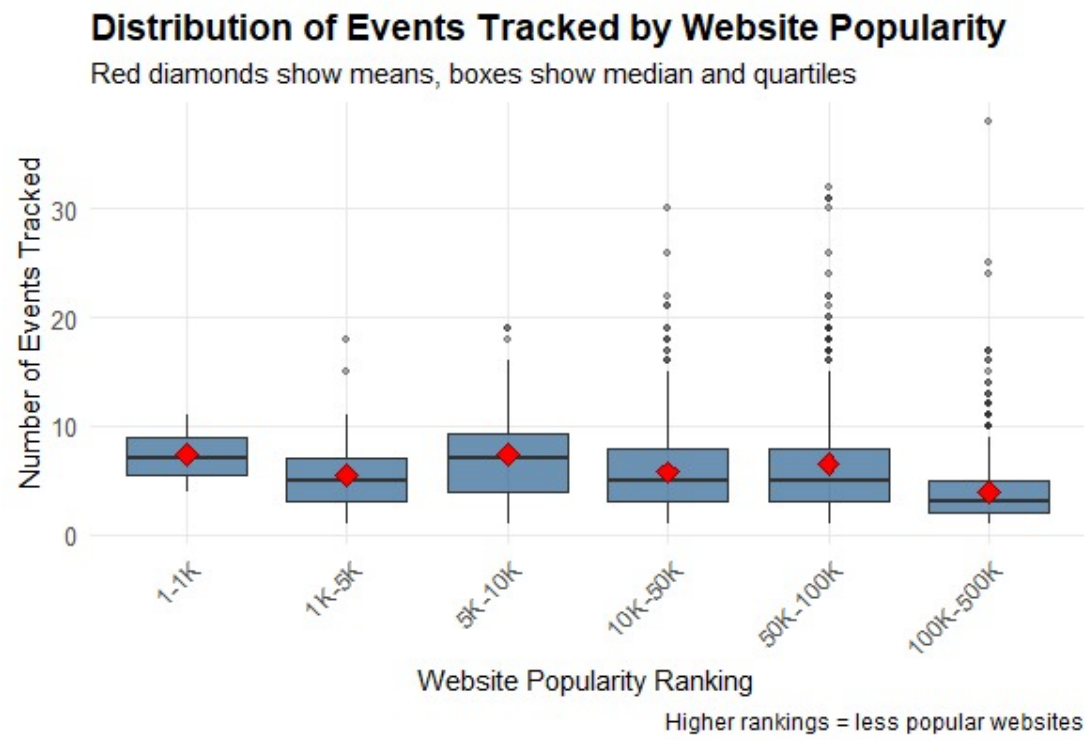

Figure 4.7: Distribution of events by Website Popularity

infrastructure.

The likelihood of the use of Consent mode increases with the number of trackers detected. This follows the trend that we discovered in section 4.2, 4.2.1, and 4.2.2.

We then question if there exists a threshold effect for tracker deployment - do websites with consent mode suddenly increase tracker usage after reaching a certain popularity level?

Fig 4.7 displays the distribution of trackers based on the popularity ranking of websites. No clear trend can be observed as the means and medians fluctuate across different rankings. With a larger data set with a more even distribution of the number of websites across each popularity ranking set, further studies can be conducted to identify if a significant link exists between the number of events tracked and website popularity ranking.

To enhance our understanding of the consent mode, we analyzed 2876 websites that used consent mode and 316 that did not. Our goal was to determine whether there is a difference in the number of trackers activated on these websites. As evident from Fig 4.5 and 4.6, among the top 11 trackers identified, 9 were common to both groups. However, websites utilizing consent mode triggered these trackers much more frequently, with an average usage rate of 56.8% for each tracker in the consent mode group (Group

A), compared to 35.2% in the non-consent mode group (Group B). This results in a difference of 21.6% higher usage for the consent mode group. Thus, we conclude that websites employing consent mode activate significantly more third-party trackers.

Rank	Tracker	Websites Using	Percentage of Tracker Users	Percentage of Group
1	Google	2460	85.57	85.54
2	Google Marketing Platform	2330	81.04	81.02
3	Facebook	2034	70.75	70.72
4	Google Static	1695	58.96	58.94
5	Google Analytics	1590	55.3	55.29
6	Google Fonts	1458	50.71	50.7
7	X Corp.	1131	39.34	39.33
8	Microsoft Clarity	1018	35.41	35.4
9	Shopify CDN	993	34.54	34.53
10	TikTok Analytics	940	32.7	32.68

Table 4.5: Top 10 Tracker Usage Statistics in Websites Using Consent Mode

Performing a Mann-Whitney U test, we obtain a p-value of 0.011, indicating high statistic significance, with an r-value of 0.602 signifying a large effect. Additionally, a paired t-test, which compared the means of two related sets of measurements, confirms a p-value of 0.002. Thus, we conclude that websites employing consent mode activate significantly more third-party trackers.

Rank	Tracker	Websites Using	Percentage of Tracker Users	Percentage of Group
1	Google Static	149	48.38	47.15
2	Google	147	47.73	46.52
3	Google Fonts	127	41.23	40.19
4	Facebook	122	39.61	38.61
5	Google Marketing Platform	111	36.04	35.13
6	Google Analytics	106	34.42	33.54
7	X Corp.	101	32.79	31.96
8	Shopify CDN	78	25.32	24.68
9	Microsoft Clarity	60	19.48	18.99
10	Google APIs	55	17.86	17.41

Table 4.6: Top 10 Tracker Usage Statistics in Websites Without Consent Mode

4.3 Summary

By studying a dataset of 4500 websites varied by popularity rank in the e-commerce domain, our research finds that GTM adoption in our e-commerce sample (70.9%)

significantly exceeds the general web adoption rate of 42% [73]. We also find that websites with lower popularity rankings typically have a higher GTM adoption rate which can be attributed to a preference of commercially available solutions instead of creating custom solutions [75, 8, 22].

Upon investigating the use of Consent mode, of the total websites that adopted GTM, 90.1% of the websites opted to use Consent mode. Of these, the adoption rate increased from 66.7% in websites with a popularity ranking of under 1k to 90% and above in websites with a popularity ranking greater than 10k, as visible in Fig 4.3 Also, it is interesting to note that both event tracking and third-party domains tracking increased significantly with an increase in consent mode adoption. Consent mode anonymizes data when the users decline consent, and thus technically respect user privacy and comply with privacy policies [41]. However, we already have numerous studies that have proved that simple anonymization does not provide any privacy and the users still get tracked and can even be identified [23, 59, 55, 33]. This leaves scope for future work honing into the nuances of consent mode to verify if using it truly protects users or leaves them vulnerable to these secondary forms of tracking.

Chapter 5

Conclusions

5.1 Summary

We have developed a Chromium-based web parser called GTM Parser that successfully detects the adoption of GTM, GTM Consent Mode, the events tracked, the third-party domains and trackers called, and the number of Google URLs fired. We have made the code, data sets used, and results available. Further, we conducted various statistical tests to ensure the validity and significance of the results, observed trends within the GTM and Consent mode adoption based on popularity ranking, event tracking and third-party domain tracking. We can summarize our work by addressing the research objectives.

RQ1: We have successfully implemented the GTM Parser that reliably detects the adoption of GTM. One major limitation to this is its success rate in websites that deploy industrial scale anti-bot / scraping measures, and when GTM is triggered on specific user actions.

RQ2 : 70.9% of websites adopt GTM as a tag management system of their preference in the e-commerce industry. The adoption of GTM is found to be inversely proportional to the websites popularity rankings.

RQ3: The adoption of GTM Consent mode increases significantly in websites with a popularity ranking of greater than 10k, with the adoption rate crossing 90%.

RQ4: The type of events tracked regardless of whether or not consent mode is adopted is largely the same (9 of 11 most commonly occurring events) but the number of events tracked increases with the use of Consent mode.

RQ5: There is no obvious threshold effect that impacts tracker deployment based on websites using consent mode reaching a certain popularity rank. However to ascertain this result, further statistical tests must be conducted on a more evenly distributed

sample size across website popularity rankings.

Our contribution is that we have very specifically studied GTM in the e-commerce domain where it is apparent from our results that GTM is a very popular solution. We also see, grounded in our research and in literature, that consent mode is supposed to anonymize user data when the user declines consent, however the events and trackers and other forms of metadata can still be tracked. We also know from prior research that users can be identified and tracked despite anonymizing data [23, 59, 55, 33]. This provides an interesting avenue to explore further to understand how exactly Consent mode of GTM aligns with the privacy laws and how it can be circumvented to still track data and users.

5.2 Limitations and Future Work

There are several limitations in our methodology.

Sample Data : Currently, this sample does not depend on geography. In further research, we could have separate data sets with URLs from zones with their own major privacy laws such as the EU with GDPR, the USA, and the UK with ICO. We can also increase the scope of the project by further categorizing the e-commerce industry based on industry, industry size, and whether the business has a larger presence online or in retail stores. There is also scope to study the difference in GTM and Consent mode adoption, and tracker behaviors between purely online stores and "bricks-and-clicks businesses".

Parameters : Within the parameters explored, there can be more data gathered on the Google domains that have been called to answer more interesting questions on the use of GTM and the depth of integration of websites within the Google ecosystem. Within the list of trackers that have been gathered, having a ground truth categorization of each of them would open avenues to explore the relationship between the type of trackers used and the use of Consent mode, and possibly the type of e-commerce website it is. We can also deep dive into the different types of events, trackers and domains that have been called by GTM. We must have a ground truth data set that we categorize our collected trackers and domains into such as analytics, advertising etc. Given more time, we can explore more complex relationships between parameters and study if there is any interlinking between the mentioned parameters and if there is then what.

Measurement and Statistical Testing : The time based GTM attribution scoring in GTM is not the most accurate format to determine which trackers can be attributed to GTM. This may miss out cached resources that get loaded instantly which might appear as trackers. Further, some trackers might coincidentally load when the GTM loads but not be caused by it. In depth analysis needs to be conducted to determine the appropriate means to detect which trackers were called by GTM.

Further, it was beyond the scope of this dissertation to ascertain which events that were recorded had privacy implications vs ones that did not. Future work could answer the question - are more events necessarily worse for privacy?

5.3 Reflections

This project was my first experience creating a tool, and a web parser at that. Through this project, I gained experience of the complete Software Development Life Cycle, from design to development to testing to deployment and to maintenance. I'm satisfied with the tool that has been created, however having worked on this project for such a duration, I am acutely aware of numerous limitations and avenues to improve this parser, its outputs and have various ideas as outlined above to improve on statistical tests and accuracy as well. I am extremely grateful for this opportunity as my research skills, analytical thinking, and programming skills have significantly improved. These are invaluable skills to my career.

Bibliography

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*, pages 674–689, 2014.

[2] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, and Bart Preneel. Fpdetective: dusting the web for fingerprinters. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1129–1140, 2013.

[3] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 1–8, 1999.

[4] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

[5] Alessandro Acquisti and Hal R. Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3):367–381, 2005.

[6] Al Jazeera. Cambridge analytica and facebook: The scandal so far. https://www.aljazeera.com/news/2018/3/28/cambridge-analytica-and-facebook-the-scandal-so-far, 2018. Accessed: 2025-08-16.

[7] Anita L Allen. Protecting one's own privacy in a big data economy. *Harv. L. Rev. F.*, 130:71, 2016.

[8] Yazn Alshamaila, Savvas Papagiannidis, and Feng Li. Cloud computing adoption by smes in the north east of england: A multi-perspective framework. *Journal of Enterprise Information Management*, 26(3):250–275, 2016.

[9] American Library Association. Chapter 2: Getting to know web analytics. In *Library Technology Reports*. n.d. Accessed: 2025-08-15.

[10] B260521. Measuring the state of web privacy - building an open source google tag manager parser. Unpublished internal report, School of Informatics, University of Edinburgh, January 2025.

[11] Jošt Bartol, Vasja Vehovar, Michael Bosnjak, and Andraž Petrovčič. Privacy concerns and self-efficacy in e-commerce: Testing an extended apco model in a prototypical eu country. *Electronic Commerce Research and Applications*, 60:101289, 2023.

[12] Reuben Binns et al. Tracking on the web, mobile and the internet of things. *Foundations and Trends® in Web Science*, 8(1–2):1–113, 2022.

[13] Sophie C Boerman, Sanne Kruikemeier, and Frederik J Zuiderveen Borgesius. Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, 46(3):363–376, 2017.

[14] Christopher G Bradley. Privacy theater in the bankruptcy courts. *Hastings LJ*, 74:607, 2022.

[15] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*, 2017(2):133–148, 2017.

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

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

[18] Chrome Developers. Overview of crux. https://developer.chrome.com/docs/crux, 2025. Accessed: 2025-08-13.

[19] W. G. Cochran. The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3):315–345, 1952.

[20] Suzanna Conrad. Using google tag manager and google analytics to track dspace metadata fields as custom dimensions. *The Code4Lib Journal*, (27), Jan 2015.

[21] Marco Cremonini, Chiara Braghin, and Claudio Agostino Ardagna. Chapter 42 - privacy on the internet. In John R. Vacca, editor, *Computer and Information Security Handbook (Second Edition)*, pages 739–753. Morgan Kaufmann, Boston, second edition edition, 2013.

[22] Thomas Davenport. Competing on analytics. *Harvard business review*, 84:98–107, 134, 02 2006.

[23] Clemens Deußer, Steffen Passmann, and Thorsten Strufe. Browsing unicity: On the limits of anonymizing web tracking data. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 777–790, 2020.

[24] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1388–1401, New York, NY, USA, 2016. Association for Computing Machinery.

[25] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1388–1401, 2016.

[26] Steven Englehardt, Dillon Reisman, and Arvind Narayanan. Detecting and defending against third-party tracking on the web. *Technical Report*, 2014.

[27] Tatiana Ermakova, Benjamin Fabian, Benedict Bender, and Kerstin Klimek. Web tracking-a literature review on the state of research. 2018.

[28] European Commission. What does 'grounds of legitimate interest' mean? https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/grounds-processing/what-does-grounds-legitimate-interest-mean$_e n$, 2025. *Accessed* : 2025 – 08 – 23.

[29] Tabatha Farney. Designing shareable tags: Using google tag manager to share code. *Journal of Library Technology Reports*, 2021. Accessed: 2025-08-15.

[30] Matthias Fassl, Lea Theresa Gröber, and Katharina Krombholz. Stop the consent theater. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[31] Federal Trade Commission. How websites and apps collect and use your information. https://consumer.ftc.gov/articles/how-websites-and-apps-collect-and-use-your-information, n.d. Accessed: 2025-08-17.

[32] Imane Fouad, Cristiana Santos, and Pierre Laperdrix. The devil is in the details: Detection, measurement and lawfulness of server-side tracking on the web. *Proceedings on Privacy Enhancing Technologies*, 2024:450–465, 10 2024.

[33] Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre De Montjoye. Anonymization: The imperfect science of using data while preserving privacy. *Science advances*, 10(29):eadn7053, 2024.

[34] Norjihan Abdul Ghani and Zailani Mohamed Sidek. Personal information privacy protection in e-commerce. *WSEAS Transactions on Information Science and Applications*, 6(3):407–416, 2009.

[35] Juanita Goicovici. Granularity and specificity of consent and implications thereof for the data controller in the light of the principle of 'purpose limitation '. *InterEULawEast: Journal for the international and european law, economics and market integrations*, 9(2):43–69, 2022.

[36] Google Analytics. Use a data layer with event handlers. https://developers.google.com/tag-platform/tag-manager/datalayeruse$_{a}d$ata$_{l}$ayer$_{w}$ith$_{e}$vent$_{h}$andlers, 2025. *Accessed* : 2025 – 08 – 23.

[37] Google Developers. Check if a web page uses analytics. https://support.google.com/analytics/answer/1032399, 2025. Accessed: 2025-08-13.

[38] Google Developers. Consent mode http parameters. https://developers.google.com/tag-platform/security/concepts/consent-modeconsent$_{m}$ode$_{h}$ttp$_{p}$arameters, 2025. *Accessed* : 2025 – 08 – 13.

[39] Google Developers. The data layer. https://developers.google.com/tag-platform/tag-manager/datalayerinstallation, 2025. Accessed: 2025-08-20.

[40] Google Developers. [ga4] about events. https://support.google.com/analytics/answer/9322688?hl=enzip Accessed: 2025-08-13.

[41] Google Developers. How consent mode works. https://support.google.com/google-ads/answer/10000067?hl=em, 2025. Accessed: 2025-08-20.

[42] Google Developers. Set the default consent state. https://developers.google.com/tag-platform/security/guides/consent?consentmode=advanced, 2025. Accessed: 2025-08-19.

[43] Google Developers. Tag manager consent mode support. https://support.google.com/tagmanager/answer/10718549?hl=enconsent-initialization-trigger, 2025. Accessed: 2025-08-13.

[44] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[45] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*, pages 305–318, 2014.

[46] HAR.fyi. Getting started accessing the http archive with bigquery. https://har.fyi/guides/getting-started/, 2025. Accessed: 2025-08-13.

[47] HAR.fyi. httparchive.crawl.pages table schema. https://har.fyi/reference/tables/pages/, 2025. Accessed: 2025-08-13.

[48] E. Hargittai and A. Marwick. "What Can I Really Do?" Explaining the Privacy Paradox with Online Apathy. *International Journal of Communication*, 10:3737–3757, 2016. Accessed: 2025-08-14.

[49] Mike Hintze. Data controllers, data processors, and the growing use of connected products in the enterprise: Managing risks, understanding benefits, and complying with the gdpr. *Journal of Internet Law (Wolters Kluwer), August*, 2018.

[50] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[51] HTTP Archive Contributors. Top ecommerce sql query — 2024 web almanac. https://github.com/HTTPArchive/almanac.httparchive.org/blob/main/sql/2024/ecommerce/top$_e$*commer* $2025-08-13$.

[52] IBM. What is logistic regression? https://www.ibm.com/think/topics/logistic-regression, 2023. Accessed: 2025-08-16.

[53] IBM Security and Ponemon Institute. Cost of a data breach report 2020. *IBM Security*, 2020.

[54] InfoTrust. Tag management: What it means for your business, n.d. Accessed: 2025-08-15.

[55] Claudia Irti. Personal data, non-personal data, anonymised data, pseudonymised data, de-identified data. In *Privacy and Data Protection in Software Services*, pages 49–57. Springer, 2021.

[56] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. The internet with privacy policies: Measuring the web upon consent. *ACM Transactions on the Web (TWEB)*, 16(3):1–24, 2022.

[57] Gayatri Priyadarsini Kancherla, Nataliia Bielova, Cristiana Santos, and Abhishek Bichhawat. Measuring compliance of consent revocation on the web. *arXiv preprint arXiv:2411.15414*, 2024.

[58] Kaspersky Lab. The true value of digital privacy: Are consumers selling themselves short? https://www.kaspersky.com/blog/privacy-report-2019/, 2019. Kaspersky Global Privacy Report 2019.

[59] Katharine Kemp. 'a rose by any other unique identifier': Regulating consumer data tracking and anonymisation claims. 2022.

[60] Hyeong Yeol Kim. Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152–155, 2017.

[61] Jonathan Weber Kulkarni and Shiridhar Gadhe. *Practical Google Analytics and Google Tag Manager for Developers*. Apress, New York, NY, 2015.

[62] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J Biega. Investigating deceptive design in gdpr's

legitimate interest. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[63] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 878–894, 2016.

[64] lieven.desmet@kuleuven.be. Tranco rankings. https://tranco-list.eu/aboutus, 2025. Accessed: 2025-08-13.

[65] Dominique Machuletz and Rainer Böhme. Multiple purposes, multiple problems: A user study of consent dialogs after gdpr. *arXiv preprint arXiv:1908.10048*, 2019.

[66] Haroon Iqbal Maseeh, Charles Jebarajakirthy, Robin Pentecost, Denni Arli, Scott Weaven, and Md. Ashaduzzaman. Privacy concerns in e-commerce: A multilevel meta-analysis. *Psychology & Marketing*, April 2021.

[67] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

[68] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe's transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 791–809. IEEE, 2020.

[69] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE Symposium on Security and Privacy*, pages 413–427, 2012.

[70] Phil McAleer, Carolina E. Kuepper-Tetzel, and Helena Paterson. Nhst: Binomial test and one-sample t-test. https://psyteachr.github.io/analysis-v2/nhst-binomial-test-and-one-sample-t-test.html, 2024. Last built: June 11, 2024; Accessed: 2025-08-17.

[71] Mary L. McHugh. The chi-square test of independence. *Biochemia Medica*, 23(2):143–149, 2013.

[72] Patrick E. McKnight and Julius Najab. Mann-whitney u test. In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc., 2010.

[73] Gilles Mertens, Nataliia Bielova, Vincent Roca, and Cristiana Santos. Google tag manager: Privacy leaks and potential legal violations. *arXiv e-prints*, pages arXiv–2312, 2023.

[74] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting price and search discrimination on the internet. *Proceedings of the 11th ACM workshop on hot topics in networks*, pages 79–84, 2012.

[75] Satish Nambisan. Digital entrepreneurship: Toward a digital technology perspective of entrepreneurship. *Entrepreneurship Theory and Practice*, 41(6):1029–1055, 2017.

[76] Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019.

[77] Xiaoguang Qi and Brian D. Davison. Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2), February 2009.

[78] Sasha Romanosky. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135, 2016.

[79] Christopher Ruff, Florian Maier, Tobias Müller, and Holger Kett. Leveraging complex event processing for supporting small and medium-sized e-commerce enterprises. In *2015 Second International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM)*, pages 61–66, 2015.

[80] Iskander Sanchez-Rola, Igor Santos, and Davide Balzarotti. Cookie synchronization: Everything you always wanted to know but were afraid to ask. *The World Wide Web Conference*, pages 1432–1442, 2019.

[81] Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, and Vincent Roca. Consent management platforms under the gdpr: processors and/or controllers? In *Annual Privacy Forum*, pages 47–69. Springer, 2021.

[82] Bruce Schneier. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015.

[83] M Siek and RMG Mukti. Business process mining from e-commerce event web logs: Conformance checking and bottleneck identification. In *IOP Conference Series: Earth and Environmental Science*, volume 729, page 012133. IOP Publishing, 2021.

[84] R. Smith and J. Shao. Privacy and e-commerce: a consumer-centric perspective. *Electronic Commerce Research*, 7:89–116, June 2007.

[85] Yong Whi Song, Hayoung Sally Lim, and Jeeyun Oh. "we think you may like this": An investigation of electronic commerce personalization for privacy-conscious consumers. *Psychology & Marketing*, 38(10):1723–1740, 2021.

[86] Peter Steiner. On the internet, nobody knows you're a dog. *The New Yorker*, 69(20):61, 1993.

[87] K Shanmuga Sundaram. User monitoring behaviour in structured ecommerce web application. *Journal of Science, Computing and Engineering Research*, 7(6), 2024.

[88] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 4:1–45, 2019.

[89] The HTTP Archive. Web almanac 2020: Ecommerce. https://almanac.httparchive.org/en/2020/ecommerce, 2025. Accessed: 2025-08-13.

[90] The HTTP Archive. Web almanac 2021: Ecommerce. https://almanac.httparchive.org/en/2021/ecommerce, 2025. Accessed: 2025-08-13.

[91] The HTTP Archive. Web almanac 2024: Limitations. https://almanac.httparchive.org/en/2024/ecommercelimitations, 2025. Accessed: 2025-08-13.

[92] Yan Tian and Concetta Stewart. History of e-commerce. In *Encyclopedia of e-commerce, e-government, and mobile commerce*, pages 559–564. IGI Global Scientific Publishing, 2006.

[93] Abhishek Tiwari. Google tag manager: How it works. https://www.abhishek-tiwari.com/pdf/google-tag-manager-how-it-works.pdf, 2014. Published on: October 03, 2014.

[94] Michael Toth, Nataliia Bielova, and Vincent Roca. On dark patterns and manipulation of website publishers by cmps. *Proceedings on Privacy Enhancing Technologies*, 2022(3):478–497, 2022.

[95] Loughborough University. The χ^2 test of goodness of fit. Teaching resource, Mathematics Learning Support Centre, n.d. Accessed: 2025-08-18.

[96] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990, 2019.

[97] S. Wahab, A. S. Mohd Zahari, K. Al Momani, and N. A. Mohd Nor. The influence of perceived privacy on customer loyalty in mobile phone services: An empirical research in jordan. *International Journal of Computer Science Issues*, 8(2):45–52, 2011.

[98] W. Wang, Q. Wu, D. Li, and X. Tian. An exploration of the influencing factors of privacy fatigue among mobile social media users from the configuration perspective. *Scientific Reports*, 15:427, 2025. Accessed: 2025-08-14.

[99] Samuel D. Warren and Louis D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, 1890.

[100] Sarah Myers West. Data capitalism: Redefining the logics of surveillance and privacy. *Business & society*, 58(1):20–41, 2019.

[101] Alan F. Westin. *Privacy and Freedom*. Ig Publishing, New York, reprint ed. edition, 1987.

Appendix A

Technology

A.1 Technology Stack

We use the following technology stack :

Parameter	**Specification**
Language	Python 3.9+
Browser Automation	Playwright with Chromium
Network Monitoring	Playwright's onRequest
Data Storage	SQLite for logging
Deployment	Docker
Logging	Standard level + Debug mode enabled

Table A.1: System Configuration Details

Appendix B

Participants' information sheet

If you had human participants, include key information that they were given in an appendix, and point to it from the ethics declaration.

Appendix C

Participants' consent form

If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration.