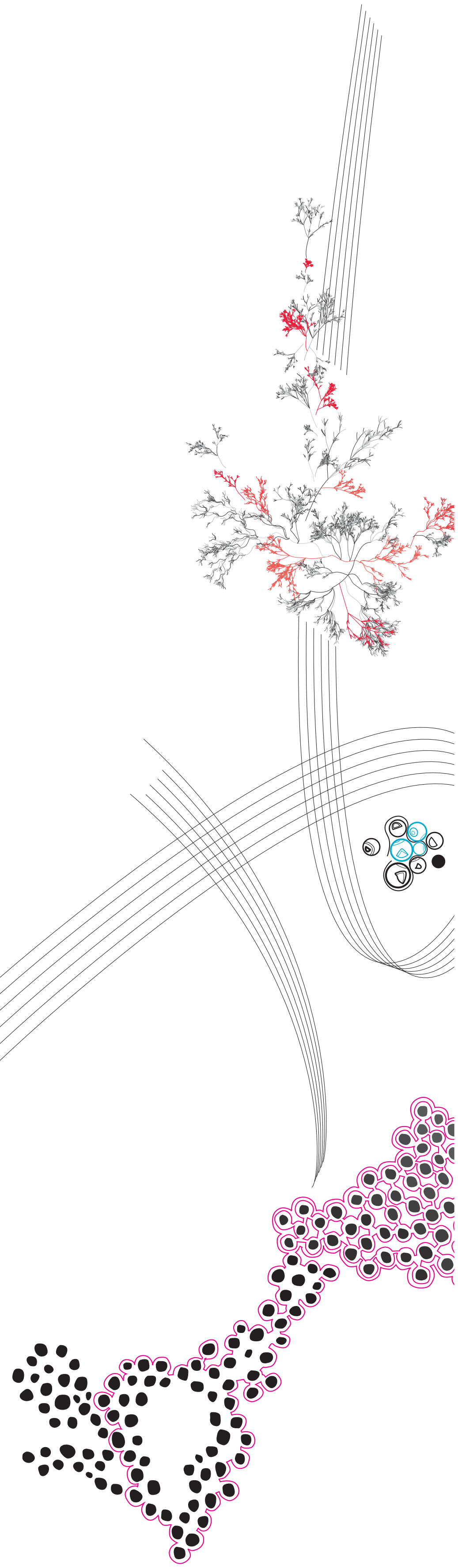


MSc Computer Science
Final Project

Understanding and Characterizing CDN Services and Paid Features

Youwei Xu

Supervisor:
Dr. R.M. van Rijswijk - Deij
Dr. A. Abhishta

August, 2024

Master Computer Science in Cyber Security
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Abstract

Content Delivery Networks (CDNs) play a crucial role in optimizing web performance and ensuring content security. This thesis provides a comprehensive analysis of CDN usage, focusing on market distribution, feature adoption, and security configurations. The study aims to fill the research gap in understanding the market dynamics of CDN services and their implications for security and privacy. By employing a robust methodology that includes data collection, statistical analysis, and automated detection methods, the research identifies major CDN providers and examines their paid features and Web Application Firewalls (WAF) across different website rankings and industry sectors. The findings reveal that 34% of the measured websites use detectable CDNs, with Cloudflare dominating the market, showing significant usage among both high-ranking and lower-ranking websites. Furthermore, 22% of the measured websites utilize at least one detectable paid feature, with Fastly and Amazon CloudFront showing higher proportions of paid feature usage among high-traffic websites. The study also uncovers the extensive use of WAF solutions, particularly among high-ranking websites, emphasizing the importance of security features in the CDN landscape. This research provides valuable insights for website operators and security professionals, aiding in informed decision-making regarding the selection and deployment of CDN services to enhance web performance, security, and user experience.

Keywords: CDN Identification, Measurement, Market Analysis, WAF

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In today's digital landscape, the importance of Content Delivery Networks (CDNs) is becoming increasingly evident. According to Cisco's report, by 2022, Content Delivery Networks (CDNs) will handle 72% of internet traffic, up from 56% in 2017 [3]. Additionally, the COVID-19 pandemic further highlighted the essential role of CDNs as online availability became increasingly important, with the rise in remote work, online education, and digital entertainment driving a unprecedented demand for reliable and fast content delivery[14]. The CDN market is expected to grow steadily during the forecast period, as the volume of data exchanged over the internet continues to increase with the ongoing rollout of high-speed networks [29]. According to recent market insights, the Content Delivery Network market size was valued at USD 18.6 billion in 2022, and is projected to grow from USD 21.67 billion in 2023 to USD 94.98 billion by 2030, with a compound annual growth rate of 23.5% during 2023-2030 [21]. CDNs play a critical role not only in cyber security but also directly impact user experience, they effectively distribute content globally, and help to defend against various cyber threats.

However, the CDN features and protection configurations used by backend servers are often opaque to end-users and may vary based on different plans and settings. Some CDN providers offer free options while charging for additional features or products, but data regarding the proportion of paying customers and the specific features or products they use typically remains obscure. This lack of transparency presents challenges for researchers and analysts aiming to understand the CDN market and its implications for security and privacy from a security economic perspective. Therefore, gaining a deeper understanding of the market distribution, usage patterns, and preferences of different websites for CDN services holds significant theoretical and practical value.

Current research on CDN services predominantly focuses on technical performance and single case studies, lacking systematic studies on overall market distribution and usage patterns. Additionally, the demands and usage patterns of CDNs vary significantly among different websites. This study will identify, analyze, and evaluate CDN service providers and their paid features, as well as protection configurations such as Web Application Firewalls (WAF) and bot management, aiming to fill this research gap.

Understanding the distribution and usage characteristics of CDN features and protection configurations is crucial. It enables website operators and security professionals to make informed decisions regarding the selection and deployment of CDN services and protective measures. By investigating websites across different ranking intervals and industry categories and analyzing the adoption rates and usage characteristics of the CDN features they utilize, organizations can optimize their defense strategies and resource allocation, thereby enhancing website performance, security, and overall user experience.

1.2 Research Objective

The primary objective of this research is to identify, analyze, and evaluate the Content Delivery Network (CDN) services and their paid features employed by websites. Specifically, the study aims to determine the most commonly used CDN service providers and assess the market distribution and usage patterns of these providers across different website ranking intervals and industry sectors. The research will investigate the various types of paid CDN features, such as Web Application Firewalls (WAF) and Bot Protection, and analyze their adoption rates among websites. Additionally, this study will explore the characteristics of websites utilizing CDN services, focusing on their business models and industry classifications. By correlating CDN usage with website categories, the research aims to uncover patterns and preferences in the adoption of CDN services and features.

Through detailed data collection, statistical analysis, and the development of automated detection methods, this research seeks to provide valuable insights into the CDN ecosystem. The findings will aid businesses in understanding the competitive landscape of CDN providers, the usage patterns of different CDN features, and the strategic decisions involved in selecting and optimizing CDN services. Ultimately, the study aims to help businesses enhance their web performance, security, and overall user experience through informed investment in CDN solutions.

1.3 Research Questions

Based on the research motivation, objectives, and current State-of-the-art outlined above, the key question of this study is: **How many websites are using paid features?** To better address this key question, multiple methods and steps need to be designed and executed. Therefore, we have divided it into several sub-questions:

1. How can we identify the CDN service providers used by the origin server?

   We begin by analyzing publicly available statistical data to identify the most popular CDN service providers, accounting for over 80% of the market share. Subsequently, we conduct a comprehensive literature review and refer to relevant materials and projects to design a methodology for accurately detecting these CDN service providers.

2. What are the usage distributions of different CDNs across most popular websites?

   Initially, we employ the devised methodology to detect CDN usage across the top 10K ranked websites and a random sample of 10K websites from the Tranco as test datasets. Then we refine the methodology based on initial findings, adjusting features as necessary. Finally, we perform large-scale measurements to analyze CDN provider distributions among the top 1 million Tranco-ranked websites.

3. What methods are available to profile the CDN features and protection configurations employed by backend servers?

   First, we create accounts of relevant CDN service providers based on requirements and refer to their technical documentations and specifications to understand the protection features they offer. Then, we categorize the features and products based on their types, and for each category, we refer to relevant materials and projects to design a method for detecting whether websites are using these features and products or have enabled corresponding protection configurations.

4. To what extent are these detection measures reliable? (How accurate are these detection methods?)

   We conduct sample testing and manually validate results in real-world scenarios, iteratively improving detection methods for enhanced reliability. Additionally, we may consider purchasing protection features from CDN service providers for self-hosted servers to establish a ground truth dataset, optimizing detection methods based on test results.

1.4 Thesis Organization

This thesis is structured into several chapters to systematically address the research objectives and questions outlined. The remainder of the thesis is organized as follows:

Chapter 2 delves into the foundational concepts related to CDNs, including their architecture, functionality, and significance. It also reviews related work in the field, highlighting the existing literature on CDN security, privacy, and identification methods. Chapter 3 first describes the research design and methods used to conduct the study. It details the approaches for CDN identification, feature identification, and the detection of Web Application Firewalls (WAF) and bot protections, as well as the identification of paid Cloudflare WAF and bot users. Then, it covers the practical steps taken to implement the research methods. It includes data collection and correlation from Tranco-ranked websites and domain categorization. Chapter 4 presents the findings of the research. It provides a comprehensive analysis of CDN usage distribution, feature adoption, and the implementation of security configurations across various websites and industry sectors. Chapter 5 summarizes the key findings of the research, discusses their implications, and suggests areas for future study to further enhance the understanding and management of CDN services.

Chapter 2

Background

2.1 CDN

2.1.1 CDN Introduction

Content Delivery Networks (CDNs) are a critical component of modern internet infrastructure, designed to enhance the speed and stability of content delivery and optimize user experience through their globally distributed network architecture and advanced caching technology.

The architecture of CDNs consists of globally distributed proxy servers that cache and distribute static content to Internet Data Center (IDC) servers located in various geographic locations. The caching strategies of CDN edge servers typically follow the HTTP protocol, using the "cache-control" header field in HTTP response packets to set cache durations and other configurations. When a user requests data, the CDN node determines whether the cached data has expired. If the cache is still valid, the cached data is returned directly; if expired, the server fetches the latest data from the origin, updates the local cache, and then returns the data to the user. This mechanism not only improves content delivery speed but also significantly enhances network reliability.

Static content distributed by CDNs refers to files stored on servers that remain unchanged with each user request. Typical static content includes images, videos, frontend resources (such as HTML, CSS, JavaScript), software packages, APK files, and compressed files. The fixed nature of static content allows caching technology to greatly improve transmission efficiency and reduce server load. By caching static content on globally distributed proxy servers, CDNs enable users to access these resources from nearby locations, thereby reducing network latency and bandwidth consumption. In contrast, dynamic content changes based on specific user factors (such as access time, location, and device), including website files (such as ASP, JSP, PHP, Perl, CGI), API requests, and database interactions. When a user accesses dynamic content, the CDN edge server fetches the request from the origin server, which dynamically generates real-time data and returns it to the edge server, which then forwards it to the user. Additionally, CDNs accelerate dynamic content access through edge computing technology. Data sent by users is processed in the cloud, and the results can be directly returned to the user, reducing response time and improving access efficiency. Edge computing, combined with CDN caching mechanisms, optimizes static content distribution and significantly enhances dynamic content processing capabilities[12].

Through its globally distributed network architecture and advanced caching and load balancing technologies, CDNs provide fast, stable, and efficient content delivery services to internet users. Firstly, CDNs significantly improve website load speeds. Caching technology and proximity access reduce transmission latency, enhancing the user experience. Secondly, CDNs effectively lower bandwidth costs. By distributing cached content, they reduce bandwidth consumption on origin servers, thus lowering overall bandwidth expenses. Additionally, CDNs increase content availability and redundancy. Their distributed architecture and caching mechanisms ensure that even if one node fails, users can still access content from other nodes, improving content availability and redundancy. Finally, CDNs enhance website security. With features such as Distributed Denial of Service (DDoS) attack protection and Web Application Firewall (WAF), they improve overall website security[11].

CDNs demonstrate broad applicability and significant advantages across various application scenarios. They are widely used in video-on-demand and audio streaming, live streaming, online marketplaces, government and healthcare websites, news forums, and blogs. By accelerating content delivery, reducing buffering time, improving access quality, and ensuring information security, CDNs greatly enhance user experience and service quality[27].

2.1.2 CDN WAF & Bot

CDN Web Application Firewall (WAF) and Bot Protection products play a crucial role in modern network security. These products are designed to protect web applications from various cyber-attacks, including SQL injection, Cross-Site Scripting (XSS), Distributed Denial of Service (DDoS) attacks, and malicious bot activities[2]. By leveraging globally distributed server nodes, CDNs provide users with fast and reliable content delivery while filtering and inspecting traffic before it reaches the origin server, thus blocking malicious traffic at the source[10].

CDN WAFs utilize multiple detection techniques to identify and block attack traffic, including signature-based detection, behavioral analysis, and machine learning. Signature-based detection matches traffic against predefined attack patterns, while behavioral analysis monitors traffic behavior to identify anomalies. Machine learning algorithms further enhance detection capabilities by continuously improving detection models through self-learning, allowing for the identification of unknown attack patterns. CDN Bot Protection products focus on identifying and managing bot traffic, distinguishing between benign bots (such as search engine crawlers) and malicious bots (such as content scrapers and credential stuffing attacks). These products employ techniques such as fingerprinting, behavioral analysis, and challenge-response tests (like CAPTCHA) to detect and block malicious bots. By utilizing a global network of edge servers, these WAF and Bot Protection products can monitor and filter web traffic in real time, effectively blocking various cyber-attacks and ensuring the security and stability of web applications.

CDN WAF products typically employ multi-layered protection mechanisms to address evolving cyber threats. Their core principles include the real-time analysis of HTTP/HTTPS traffic utilizing predefined security rules and policies. These rule sets, maintained and updated by security experts based on the latest threat intelligence and attack patterns, cover common attack types such as SQL injection, XSS, and Remote Code Execution (RCE). Combining machine learning and behavioral analysis allows CDN WAFs to detect abnormal traffic patterns[34]. For instance, by analyzing the differences between normal user behavior and anomalous activities, WAFs can identify malicious bot traffic and DDoS attacks. This approach allows for effective defense against even unknown attack methods through behavioral anomaly detection mechanisms. Moreover, CDN WAF products possess adaptive learning capabilities. Through continuous learning and optimization, WAFs can dynamically adjust their security strategies to address emerging threats, ensuring efficient protection in a constantly changing attack environment.

Despite the robust security features provided by CDN WAF products, attackers may still attempt to bypass these defenses. For example, attackers might fragment malicious requests into smaller pieces to evade rule-based detection. This can be achieved by inserting harmless characters or using obfuscation techniques to bypass specific security rules. Additionally, attackers can employ various encoding or encryption methods, such as Base64 encoding, URL encoding, or double encoding, to hide malicious payloads within seemingly harmless requests, evading WAF detection. Furthermore, attackers can simulate normal user behavior or use legitimate traffic patterns to mask their malicious activities. Techniques such as rate limiting and delaying requests can reduce the suspiciousness of their actions, thereby bypassing WAF defenses[26].

CDN Bot Protection systems, such as Cloudflare, employ a variety of sophisticated detection methods to distinguish between legitimate users and automated bots[9]. IP address analysis is a key method, assigning trust scores by distinguishing between residential, mobile, and data center IP addresses to detect the suspiciousness of high-frequency requests. TLS handshake fingerprinting (JA3) is used to identify non-browser clients through unique fingerprints generated during the handshake. HTTP header analysis detects anomalies or inconsistencies in HTTP headers (such as User-Agent, Accept-Language, and Cookie), indicating bot activity. JavaScript fingerprinting involves executing JavaScript challenges to gather information about the client's runtime environment, browser capabilities, and hardware, which helps identify automation tools. Behavioral analysis monitors browsing patterns and request frequency to detect abnormal behavior, where high-frequency page requests can lower the trust score, leading to potential blocking[30].

Attackers also employ various evasion techniques to bypass these detection methods. Using high-quality proxies is a common tactic, selecting high-trust residential or mobile proxies and rotating them regularly to evade IP-based detection. Mimicking the TLS handshake and HTTP headers of popular browsers reduces the likelihood of detection. In response to JavaScript fingerprinting, attackers utilize headless browsers (such as Puppeteer, Selenium) to automate browser operations and solve JavaScript challenges, using plugins like puppeteer-stealth to hide automation behavior. Session management techniques combine headless browsers with HTTP clients (such as FlareSolverr) to reuse session values, reducing the need to repeatedly solve JavaScript challenges. Additionally, attackers simulate natural browsing patterns, including randomizing time intervals between requests, changing viewport sizes, and mimicking user interactions to maintain a high trust score and reduce the risk of detection[36].

2.2 Related Work

2.2.1 CDN Security and Privacy

There is relatively little research on CDN security at present. Understanding the common security challenges faced by CDNs and potential countermeasures can help comprehend CDN security configurations and optimize defense strategies against emerging cyber threats.

Ghaznavi et al. [16] provide the first comprehensive survey on security challenges facing CDNs, along with their attack detection and mitigation approaches. The authors categorize CDN security challenges per CDN infrastructure component, discuss possible countermeasures and their effectiveness, and describe future research directions. This paper aims to highlight the state of CDN security and identify important research challenges in this area. In their work, the authors categorize CDN security challenges into three sections by CDN components: i.e., edge server, request routing, and origin server, respectively. See the figure below:

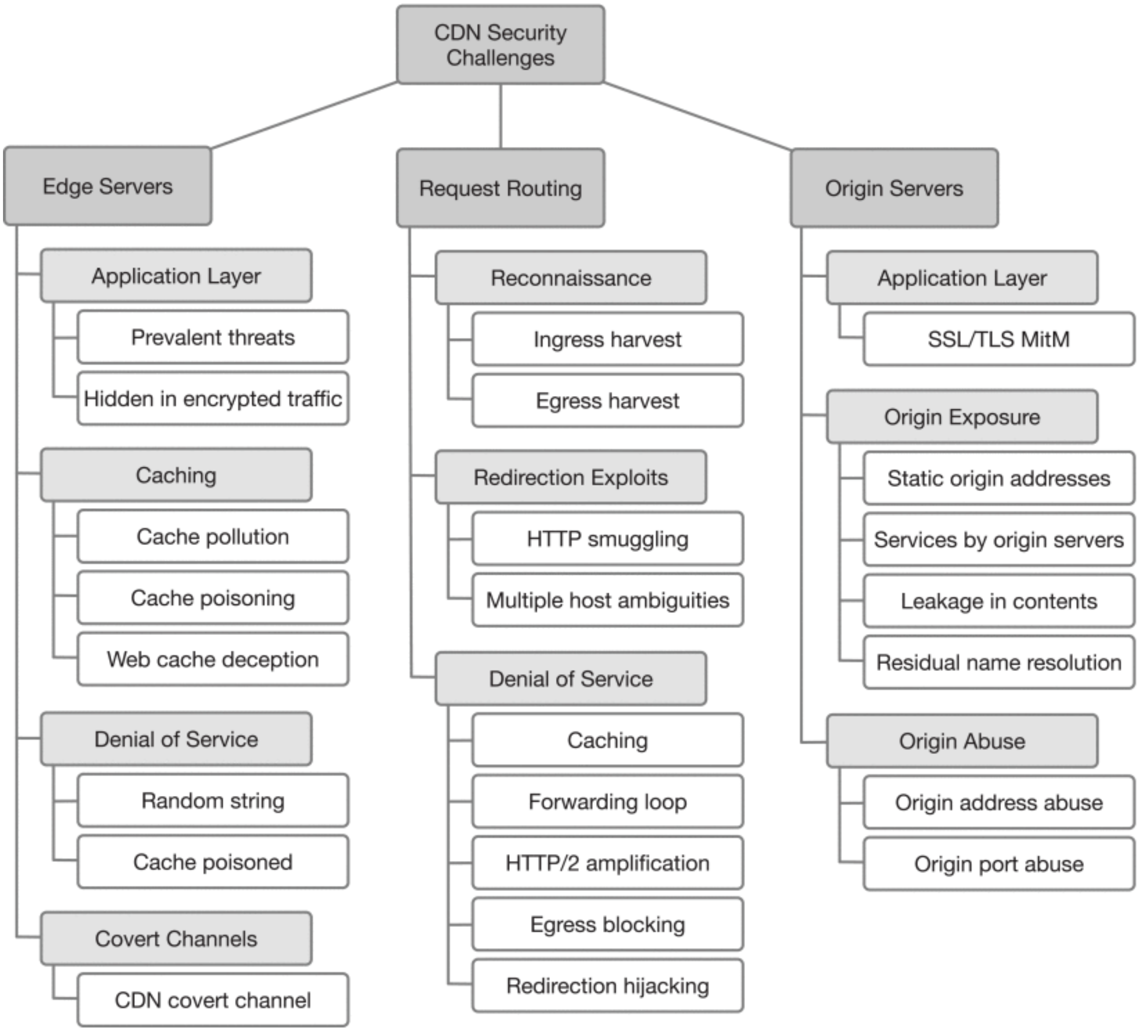


Figure 2.1: Categorization of CDN security challenges(Ghaznavi, et al, Content delivery network security: A survey., 2169)

Zolfaghari et al. [38] extract the lifecycle of a CDN as suggested by the existing research, and then investigate previous relevant works on each phase of the lifecycle to clarify where the research is currently located and headed. They summarize relevant research and categorize them into four domains: i.e., ownership claiming, intrusion detection, penetration testing and privacy. A comparison between security research in CDN and other technologies shows a clear need for more focus on CDN security. For example, protecting against Distributed Denial of Service (DDoS) attacks requires serious attention as CDNs closely interact with users. However, it is not just CDNs that need protection: Users and businesses should also be shielded from harmful CDNs. Privacy concerns, like managing user consent and safeguarding behavioral information, need more effort. Recent advances like homomorphic encryption offer promise in improving CDN privacy by allowing encrypted data to be processed as if it were decrypted, processed, and re-encrypted. Although homomorphic encryption is mature enough for some services like content search, integrating it into CDN privacy requires interaction between users and content owners/providers. Additionally, copyright measures such as watermarking, explored in fields like digital cinema, should be further researched and standardized to protect the business value of content shared over CDNs. Security considerations should be integrated into every phase of the CDN lifecycle.

2.2.2 CDN Identification

Detecting and identifying the CDNs used by websites is an important part of this research, they provide an important approach to measure and understand CDN market dynamics. Various methods and techniques have been proposed to pinpoint the CDN behind domains. However, to date, we have been unable to find studies focusing on detecting the protection configurations of CDNs or distinguishing between sites utilizing free and paid services. Hence, this study aims to shed light on this subject and provide insights into these aspects.

Zhou et al. [37] propose a new method "MultiFinder" for detecting CDNs, including multiple-CDN deployments, behind domains using a combination of DNS and HTTP-based measurements. Key elements of MultiFinder include sending DNS queries with EDNS0 Client Subnet prefixes to open DNS resolvers, analyzing CNAME, HTTP(S) headers, TLS certificates and RDAP(ipwhois) information.

Another method to discover CDNs behind the hostname is to convert IPs to AS and then perform a regular expression search on the name field in the AS2Org database to identify the CDNs. This method is widely used in academic works [32, 7].

Sosnowski et al. [33] propose an active measurement-based methodology for acquiring Transport Layer Security (TLS) metadata from servers and leverage it for their fingerprinting. The fingerprints capture the characteristic behavior of the TLS stack primarily caused by the implementation, configuration, and hardware support of the underlying server. Using an empirical optimization strategy that maximizes information gain from every handshake to minimize measurement costs, they generated 10 general-purpose Client Hellos used as scanning probes to create a large database of TLS configurations used for classifying servers. This approach can be used for CDN detection. CDNs provide large number of verifiable data samples. For this measurement, they combined Autonomous System (AS), HTTP header, and x509 certificate data to generate a ground truth. Then they assigned a server to a CDN if it had a previously observed fingerprint. This assignment was unambiguous because the fingerprints did not overlap.

Shobiri et al. [31] describe a methodology for identifying CDNs used by websites through feature extraction which is also used in previous studies [17, 18]. Key components of this process include reverse DNS, HTTP headers, CNAME records, and TLS certificates. Modifications and updates to previous feature sets are made due to changes in certificate extensions, privacy policies, and security concerns. They developed Python scripts to extract features from websites in the Alexa top 1 million, with features stored for analysis. Known CDN features are used to identify CDNs, while unknown CDNs are identified through feature clustering and manual verification. Various techniques, such as reverse DNS queries and examination of HTTP headers and CNAME records, are employed to detect CDN usage. However, TLS certificates are not utilized due to the widespread adoption of Server Name Indication (SNI) by most CDN providers, rendering them less useful for front-end scanning.

2.2.3 CDN Bot Protection

CDN Bot protection is also an essential component of CDN protection configurations. Detecting and identifying whether websites have implemented bot protection features or products is crucial for this study. Therefore, we conducted a literature review to understand the current research status and technological advancements in anti-bot services, as well as common evasion and bypass methods. However, to date, we have not found literature focusing on detecting the bot protection services used by websites.

Amin Azad, et al. [1] investigate the design and implementation details of commercial anti-bot services in an effort to understand how they operate and whether they can effectively identify and block malicious bots in practice. Since malicious bots can lie about their identity, prior research has proposed a number of methods for bot detection, including behavior-based detection (based on the premise that bots browse websites differently than real users [22, 25]), detection based on accessing content that is invisible for regular users [28, 35] and more recently, based on browser fingerprinting [5]. Once a visitor is suspected to be a bot, the website can request the solving of CAPTCHAs, rate-limit the user, or altogether block traffic from the offending IP address.

Li, et al. [24] outline methods for detecting and categorizing bots, including browser fingerprinting, TLS fingerprinting, behavioral analysis, and tracking browsing sessions. Browser fingerprinting involves identifying unique browser configurations and behaviors, while TLS fingerprinting detects discrepancies between declared and actual connection characteristics. Behavioral analysis examines factors such as request rates and interactions with critical endpoints, while tracking browsing sessions helps group requests from the same bot for analysis and identification purposes.

Chiapponi, et al. [6] discuss the evolution process of CDN bot protection in their work. Initially, techniques such as IP reputation, HTTP-based rate-limiting, and HTTP header anomaly detection were used. Browser fingerprinting became popular, followed by JavaScript and cookie challenges to deter bots. As bots evolved, methods like automated browser detection and human interaction checks were introduced, including CAPTCHAs and machine learning algorithms. Recently, strategies focus on wasting bot time and introducing crypto challenges. Anti-bot solutions now employ two main detection approaches: knowledge-based, which recognizes scraper fingerprints through HTTP headers and specific parameters, and behavior-based, which uses machine learning to detect outliers in HTTP headers and payloads, classifying them as bot traffic and implementing countermeasures. Correspondingly, bot evolution progresses from simple scripts to browser emulation frameworks like Scrapy and PhantomJS, employing automated browsers such as Selenium with cookies and JavaScript support. Bots now tackle CAPTCHAs by using infrastructures like captcha farms or redirecting them to unwitting users. Additionally, scrapers utilize Residential IP providers, offering anonymity and reputational advantages, eliminating the need for private distributed infrastructures.

Chapter 3

Methodology

3.1 Identification

This section outlines the method for identifying the CDN service providers used by origin servers. The process involves a combination of data collection, statistical analysis, and literature review to ensure the accuracy and comprehensiveness of CDN identification. The following steps detail our approach:

3.1.1 CDN Identification

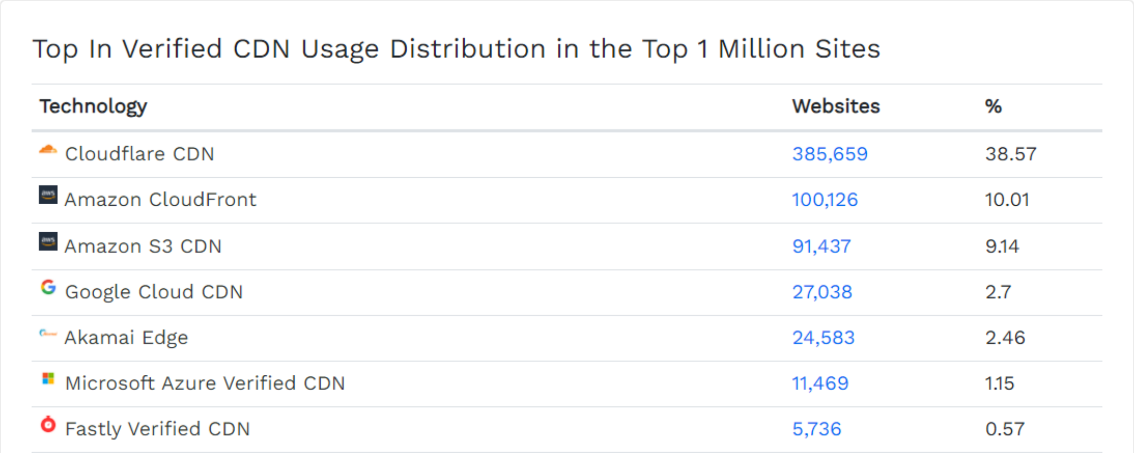
To identify the currently popular CDN service providers, we gathered industry reports, market research company data, reports from technical analysis platforms, and publicly available statistics to determine the major CDN service providers with over 80% market share. We mainly consulted three following sources:

- Gartner
- IDC
- Built With

Gartner defines a "Content Delivery Network" (CDN) as a highly distributed, edge-based cloud delivery platform that provides content acceleration, API caching, image optimization, streaming video delivery, web application and perimeter security, as well as edge computing and storage capabilities. We referred to the Gartner Global CDN Reviews and Ratings [15].

The IDC MarketScape report evaluates the market performance and capabilities of global commercial content delivery network service providers, categorizing them into different leadership tiers. We referred to the IDC MarketScape: Worldwide Commercial Content Delivery Network Services 2022 Vendor Assessment statistics [20].

Built With provides data on the distribution of CDN technologies used by the top one million websites globally and ranks CDN service providers by usage. We also referred to Built With's Verified CDN Usage Distribution in the Top 1 Million Sites publicly available statistics [4].



Top In Verified CDN Usage Distribution in the Top 1 Million Sites

Technology	Websites	%
Cloudflare CDN	385,659	38.57
Amazon CloudFront	100,126	10.01
Amazon S3 CDN	91,437	9.14
Google Cloud CDN	27,038	2.7
Akamai Edge	24,583	2.46
Microsoft Azure Verified CDN	11,469	1.15
Fastly Verified CDN	5,736	0.57

FIGURE 3.1: Top In Verified CDN Usage Distribution in the Top 1 Million Sites

TABLE 3.1 lists most popular CDNs presented in each data source. By synthesizing these three data sources, we also determined the CDN service providers we will include in this study, listed in alphabetical order as follows:

- Akamai
- Amazon CloudFront
- Cloudflare
- Fastly
- Google Cloud CDN
- Microsoft Azure CDN

TABLE 3.1: Most Popular CDNs Presented in Data Source

	Gartner	IDC	Built With
Cloudflare	Y	Y	Y
Amazon CloudFront	Y	Y	Y
Akamai	Y	Y	Y
Google Cloud CDN	Y	N	Y
Microsoft Azure CDN	Y	N	Y
Fastly	Y	Y	Y

To identify the CDN service providers used by websites, we conducted a literature review and referenced the CDN identification methods used in related research, making improvements to suit our study.

As mentioned in the related work, the currently popular CDN detection methods include DNS records (CNAME, PTR, etc.), HTTP Headers, and mapping IP addresses to ASNs. Some studies also used TLS certificates and TLS fingerprint data; however, these methods have certain limitations and are therefore less commonly applied.

Based on the identification methods utilized in related research, we have adopted, improved, and proposed the following identification methods:

- **CNAME**: Some CDN service providers use CNAME records to redirect requests from the origin domain to a CDN provider-specified (sub)domain. For instance, websites using Fastly typically have CNAME records ending in "fastly.net".

- **ASN**: Edge servers used by CDN providers are generally located within their corresponding Autonomous Systems (AS). Hence, domains using CDN services often have IP addresses associated with the AS Numbers(ASN)s of the CDN providers. For example, websites using Akamai typically have IP addresses associated with ASNs "16625" or "20940". **Note**: We did not use ASN to detect Amazon CloudFront, Google Cloud CDN, and Microsoft Azure CDN because these providers also offer cloud computing services. These services share the same ASN as the CDN edge servers, which could lead to inaccurate results.

- **PTR**: PTR records provide the domain information corresponding to an IP address. Some CDN providers choose to disclose reverse domain name resolution records to indicate the IP owner. For instance, websites using Amazon CloudFront often have reverse resolution records ending in "cloudfront.net".

- **HTTP Headers**: Many CDN providers return unique HTTP headers in responses. For example, websites using Cloudflare often include headers such as "Server: Cloudflare" and/or "CF-RAY".

More detailed identification methods are illustrated in TABLE 3.2.

Table 3.2: CDN Identifiers

CDN Provider	CNAME	ASN	PTR	Headers
Cloudflare	cdn.cloudflare.net	13335, 209242	/	Server: Cloudflare CF-Cache-Status CF-RAY
Amazon CloudFront	cloudfront.net	/	cloudfront.net	Server: CloudFront Via: CloudFront x-amz-cf-id x-amz-cf-pop
Akamai	edgekey.net edgesuite.net akamaized.net akamaihd.net	16625, 20940	akamaitechnologies.com	Server: AkamaiGHost X-Akamai-Transformed X-Akam-SW-Version Akamai-GRN X-Akamai-Request-ID Akamai-Mon-Iucid-Del Akamai-True-TTL
Google Cloud CDN	/	/	googleusercontent.com	/
Microsoft Azure CDN	azureedge.net azurefd.net	/	/	X-Azure-Ref x-fd-int-roxy-purgeid X-Azure-Ref-OriginShield
Fastly	fastly.net	54113	/	X-Fastly-Request-ID Fastly-Restarts Fastly-Client-IP fastly-request-id Fastly-Drupal-Html

3.1.2 CDN Features Identification

To identify the paid CDN features used by websites, we first visited the websites of various CDN service providers to understand the paid features each provider offers. Secondly, we referenced relevant documentation to understand their feature characteristics and proposed possible detection methods. We focused on the following attributes to identify the paid

CDN features used by websites. We explain some of used attributes below. For more detailed identification methods please refer to appendix A.

- **CNAME**: CNAME records can indicate the CDN features used by a website. For example, websites with CNAME records ending in "elb.amazonaws.com" suggest the use of Amazon Classic Load Balancer.
- **PTR**: PTR records can indicate the CDN features used by a website. For instance, websites with reverse domain name resolution results ending in "awsglobalaccelerator.com" suggest the use of Amazon Global Accelerator.
- **ASD / ASN** (Autonomous System Description / Number): Some CDN features might have specific ASD / ASN info. For instance, the IPs used by Cloudflare Spectrum corresponds to the AS information: CLOUDFLARESPECTRUM Cloudflare, Inc. / 209242.
- **Headers** (HTTP Headers): HTTP response Headers can also indicate the CDN features used by a website. For example, the Header "X-Azure-Ref-OriginShield" suggests the use of Azure Origin Shield.
- **NS**: NS records indicate the authoritative name servers used by a domain. Many CDN providers also offer paid DNS services. For instance, websites using Akamai Edge DNS typically have NS records containing "akam.net".
- **CERT-issuer** (TLS Certificate Issuer): Many CDN providers restrict the Certificate Authorities (CAs) from which a website can obtain TLS certificates. For example, only paid Fastly users can use CAs other than Let's Encrypt and Certainly.
- **CERT-san** (TLS Certificate Subject Alternative Name): SAN indicates the domains, IP addresses, and email addresses covered by a single TLS certificate. Many CDN providers restrict the scope of TLS certificates for websites but offer paid advanced certificates. For example, Amazon CloudFront provides paid dedicated IP certificates.

3.1.3 WAF&Bot Identification

Additionally, we also identified CDN WAF (Web Application Firewall) and Bot Protection features. However, it is important to note that the detectability of WAF and bot protection features may vary depending on website settings. We only made ethical modifications to HTTP requests and did not send any requests containing malicious code to the websites. This means that for some websites that prioritize user experience over security, or those using low-security configurations, our probing might not be detected as malicious behavior by the WAF, thus not triggering protective measures. Consequently, the identification results represent a lower bound, and it is possible that some websites may be using these features without our detection. We explain some of used attributes below, and identification methods are shown in TABLE 3.3-3.5.

Note: Since the functionality of CDN WAF and Bot Management features sometimes overlap and cannot be completely distinguished from the client side, we have unified the terminology in this study, referring to both features collectively as WAF.

- **HTTP Title**: Many CDN providers' WAF and Bot features return default error pages. For instance, websites using Cloudflare WAF might have error pages with titles such as "Attention Required! | Cloudflare".

- **HTTP Body**: To increase identification accuracy, we also examine the body content of error pages. For example, websites using Akamai WAF may return an error page with the title "Access Denied". However, since this title is similar to those returned by some HTTP servers for 403 errors, we also review the body content. If it contains "errors.edgesuite.net", it is typically indicative of Akamai WAF usage.
- **HTTP Headers**: Headers are also part of the detection mechanism. For example, if the "x-amzn-waf-action" header is present in an HTTP response, it typically indicates the use of AWS WAF.

Table 3.3: WAF&Bot Identifiers - Cloudflare

Title	Body	Headers	Feature
Just a moment...	/	CF-RAY	Cloudflare WAF
Attention Required! | Cloudflare	/	/	Cloudflare WAF
Access denied	Cloudflare Ray ID	CF-RAY	Cloudflare WAF
/	/	one of the following: cf-chl-out cf-mitigated	Cloudflare WAF

Table 3.4: WAF&Bot Identifiers - CloudFront

Title	Body	Headers	Feature
ERROR: The request could not be satisfied	Generated by cloudfront (CloudFront)	/	AWS WAF
Human Verification	awswaf.com	/	AWS WAF
/	/	x-amzn-waf-action	AWS WAF

Table 3.5: WAF&Bot Identifiers - Akamai

Title	Body	Headers	Feature
Access Denied	errors.edgesuite.net	/	Akamai WAF

3.1.4 Cloudflare WAF&Bot Paid User Identification

In contrast to other CDN service providers, where access to WAF and Bot services typically requires paid subscriptions, Cloudflare offers partial functionalities to its free users. To distinguish between free and paid users of Cloudflare's WAF & Bot features, we utilized the following criteria:

- **Custom Pages (Error and Challenge)**: This rule matches domains that return a custom error page, a feature available only to paid users.
- **Cloudflare 1XXX Errors**: This rule matches domains that return 1xxx errors, with restrictions applicable only to paid users.
- **TLS Fingerprint Detection**: This rule matches domains that have enabled the advanced Bot detection mechanism, a feature available only to paid users.

For the TLS fingerprint detection rule, we aim to determine whether a website employs this technique as a bot protection measure. Our methodology is as follows:

Initially, we send a request to the website using typical legitimate browser headers (including User-Agent) via the Ruby net/http library's TLS fingerprint. We then observe whether the request is challenged or blocked. If the request is challenged or blocked, we resend the same request headers but spoof the browser's TLS fingerprint to that of Chrome 116 or Safari 15.5. We then re-evaluate if the request is challenged or blocked. If the request is still challenged or blocked, this suggests that the protection mechanism is not due to TLS fingerprint detection. Conversely, if the request is accepted, this indicates that the website utilizes TLS fingerprint detection as a bot protection technique, implying that the corresponding Cloudflare WAF & Bot user is a paid user.

3.2 Measurement and Correlation

This section outlines the process for collecting and correlating data related to CDN usage, CDN feature usage, and domain categorization. By applying the detection and identification methods described in the previous section, we performed large-scale measurements. We then obtained domain categorization data and correlated it with our findings to gain deeper insights. The following steps detail our approach:

3.2.1 Tranco-ranked websites

To detect the CDN, paid CDN features, and CDN protection configurations (WAF & Bot) used by websites, we implemented automated ruby scripts based on the detection methods designed in our methodology. First, we categorized the required attributes into the following three categories:

1. **DNS records**, including: CNAME, NS, A, AAAA, PTR
2. **AS information**, including: AS-Country, AS-Description, and ASN
3. **HTTP response**, including: HTTP Code, HTTP Headers, HTTP Body, and TLS Certificate

We used Unbound configured as a local DNS Server for DNS queries, and the datebase from iptoasn.com to resolve IP-to-AS information, then sent HTTP requests to websites to retrieve HTTP responses and corresponding TLS certificate information. After obtaining the necessary information, we cross-matched it based on the identification methods outlined in our methodology to determine the CDN and paid CDN features used by the websites.

We developed a Ruby script, measured the top 10,000 websites ranked by Tranco[23]. We utilized multithreading to reduce runtime. It takes approximately 3 minutes on a system running Debian 12 with an AMD EPYC 9534 16-core processor and 32GB RAM. We manually sampled and analyzed the scan results, which were accurate and met our expectations. Consequently, we conducted large-scale measurements on the top 1 million websites ranked by Tranco.

To further improve the script's efficiency, we analyzed the measurement results and found that most websites' CDNs could be identified using CNAME, PTR, and ASN records. Websites identified as using CDNs via HTTP Headers accounted for only 0.41%, as shown in FIGURE 3.2. Therefore, we decided to send HTTP requests only to websites identified as using CDNs to further detect the paid CDN features, thereby reducing runtime. The optimized execution time was approximately 5 hours on the same system. To verify data stability, we also did daily measurement of the top 1 million websites ranked by Tranco. The analysis of the results showed stable measurement data.

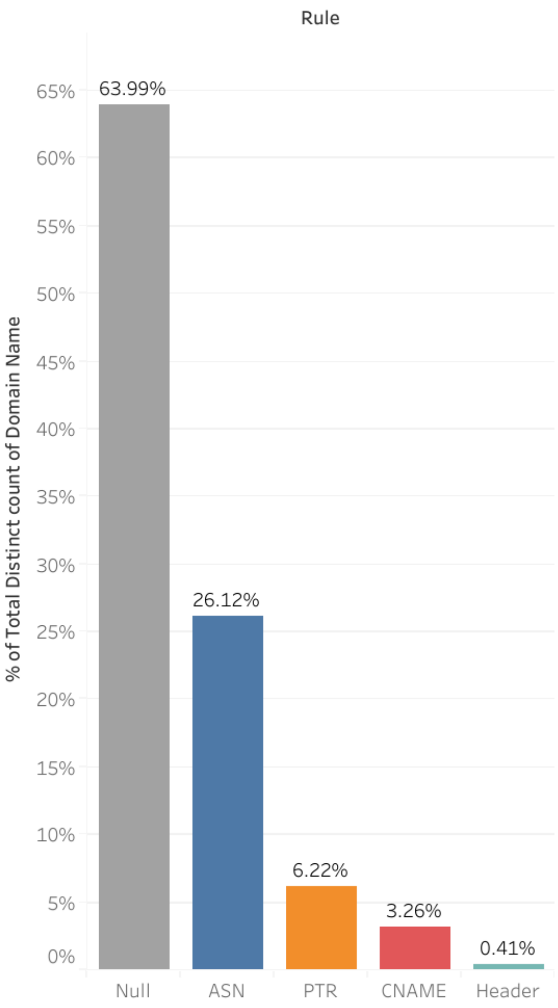


Figure 3.2: Percentage of Domains Identified by Header

Note: We used the Tranco ranking list with subdomains for measurement because different subdomains and top-level domains may use different CDNs or CDN features. Therefore, we used the list with subdomains to improve the accuracy of the measurement results. Additionally, for domains that require sending HTTP requests, some domains may redirect the request to a subdomain or another domain. We only follow redirects within the same FQDN to obtain accurate HTTP response and TLS certificate information, thus enhancing the accuracy of the measurement results.

3.2.2 Domain Categorization and Correlation

To analyze the industry distribution of websites using CDNs and their paid features, We obtained domain classification information from the IBM X-Force Exchange API[19]. This API provides comprehensive domain categorization data, classifying websites into various industry categories such as technology, finance, healthcare, retail, and more. New users can access a 30-day Standard trial with 1 million API calls, and the rate limit is set at

3000 requests per minute. As an alternative, the Cloudflare API is available, but with limitations for free tier users who can retrieve information on up to 100 domains per month, insufficient for our research needs.

We then correlated the classification data from the IBM X-Force Exchange API with our previously collected measurement results. This integration involved aligning each domain's classification with the identified CDN and its associated features, enabling further detailed analysis.

Chapter 4

Results

The Results section provides a comprehensive analysis of the data collected on CDN usage distribution, focusing on different aspects of the market and the characteristics of various CDN providers. This section is structured as follows:

1. **CDN Usage Distribution**:
   - **Overall Usage**: This part shows the market share of various CDN providers from July 1st to July 7th, highlighting Cloudflare's dominance. FIGURE 4.1 and 4.2 visualizes the usage proportions among the top one million websites.
   - **Ranking Distribution**: Using ECDF charts (FIGURE 4.3 and 4.4), this subsection examines how different CDN providers are utilized across various website rankings, revealing distinct market strategies and user preferences.
   - **Categorized Usage**: This analysis correlates CDN usage with domain classification data, illustrating the distribution of CDN providers across different industries in FIGURE 4.5-4.7, and Appendix B.
2. **CDN Features**:
   - **Overall Usage**: This subsection provides insights into the overall usage proportions of paid CDN features. FIGURE 4.8 and 4.9 shows the market share of paid features among different CDN providers.
   - **Ranking Distribution**: FIGURE 4.10 and 4.11 depict the distribution of paid feature usage across different website rankings, highlighting the preferences of high-traffic websites.
   - **Detailed Usage**: This part provides a deeper analysis of CDN feature usage for each CDN provider, as shown in Appendix C.
   - **Feature Combinations**: FIGURE 4.12 presents a heatmap illustrating the popularity of various CDN feature combinations among different CDN providers.
   - **Categorized Usage**: This part analyzes the usage of CDN features across various website categories, as shown in FIGURE 4.13-4.15.
3. **WAF & Bot**:
   - **Overall Usage**: This subsection examines the distribution of Web Application Firewall (WAF) features among different CDN providers. FIGURE 4.16 provides a detailed view of WAF usage.

- **Categorized Usage**: FIGURE 4.17-4.19 show the distribution of WAF features across different website categories.
- **Ranking Distribution**: The ECDF charts (FIGURE 4.20-4.22) illustrate the usage of WAF features among websites of various rankings.

Throughout this section, each figure provides visual representation and detailed analysis to support the findings, offering a comprehensive understanding of the CDN market.

4.1 CDN Usage Distribution

4.1.1 Overall Distribution

Based on our measurement data of the Tranco top 1 million websites, we conducted a CDN Usage Distribution analysis to gain insights of the CDN market. As shown in FIGURE 4.1, 34.38% of the measured websites use detectable CDNs. FIGURE 4.2 shows the CDN usage distribution. We reviewed the measurement data for the period from July 1st to July 7th. The data analysis reveals that, aside from minor fluctuations observed in Cloudflare and Google Cloud CDN, the proportions of other CDN service providers remained consistent and without any visually perceptible variations. This consistency across different providers suggests that our measurement methodology is robust and reliable.

As shown in FIGURE 4.2, Cloudflare overwhelmingly dominates the market with a usage proportion of approximately 70%, significantly surpassing other CDN providers. Amazon CloudFront and Akamai follow closely behind, but their usage proportions are noticeably lower than that of Cloudflare. Google Cloud CDN and Fastly rank fourth and fifth, respectively, demonstrating their significant positions in the market. However, the market shares of these providers remain substantially lower compared to Cloudflare. Microsoft Azure CDN/Azure Front Door holds a relatively low market share, reflecting its limited influence in the CDN market.

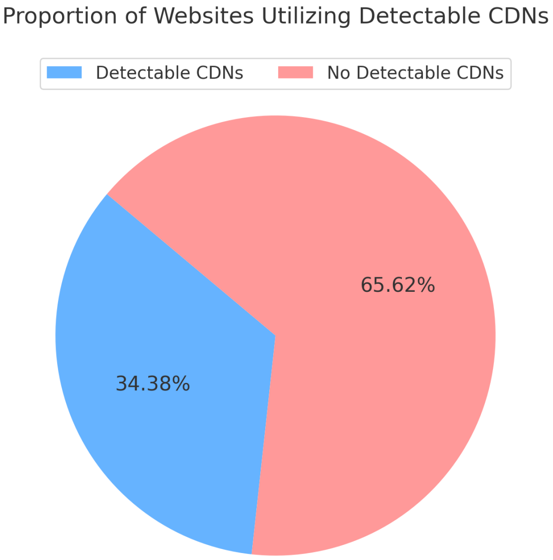


FIGURE 4.1: Proportion of Websites Utilizing Detectable CDNs

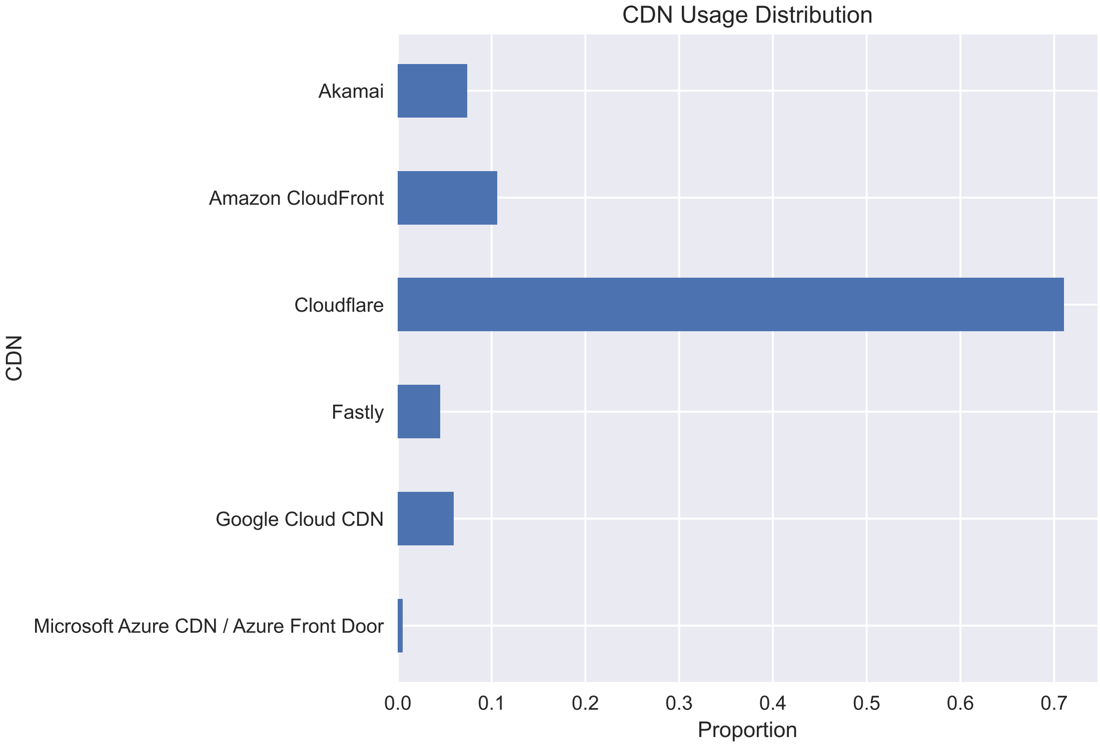


FIGURE 4.2: CDN Usage Distribution of Tranco Top 1 Million Sites

4.1.2 Ranking Distribution

We further analyzed the distribution characteristics of each CDN provider across different ranking intervals. The empirical cumulative distribution function (ECDF) chart reveals the usage and ranking distribution characteristics of different CDN providers among the top one million websites, as shown in FIGURE 4.3 and 4.4. Amazon CloudFront, Akamai, and Fastly exhibit similar distribution patterns, with higher usage proportions among highly-ranked websites. This distribution characteristic reflects the specific advantages and market positioning of these CDN service providers in high-traffic websites. They likely focus more on meeting the service demands of high-traffic, high-value websites by providing optimized performance and reliability to meet their stringent requirements. In contrast, Cloudflare and Google Cloud CDN display a more uniform ranking distribution, with widespread usage across all ranking intervals. This uniform distribution indicates that Google Cloud CDN and Cloudflare have adopted more universal strategies in the market, serving not only high-ranking websites but also a substantial number of mid- to low-ranking websites. This reflects their attractiveness and competitiveness across different market tiers. Microsoft Azure CDN/Azure Front Door, on the other hand, shows a distinct distribution characteristic. Its usage proportion is relatively low in the overall market but highly concentrated among websites ranked after the 75th percentile. This may indicate that Azure primarily attracts small to medium-sized websites, possibly due to its pricing strategy, service characteristics, or market promotion strategies that better align with the needs of these websites.

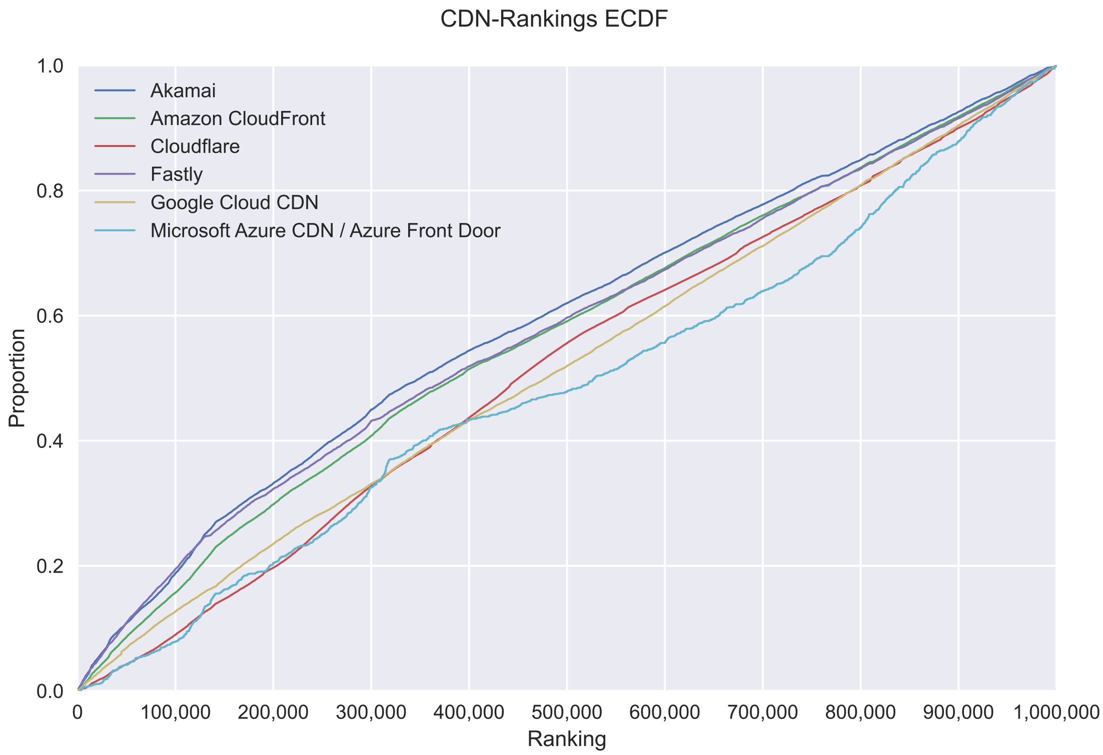


FIGURE 4.3: CDN-Rankings ECDF

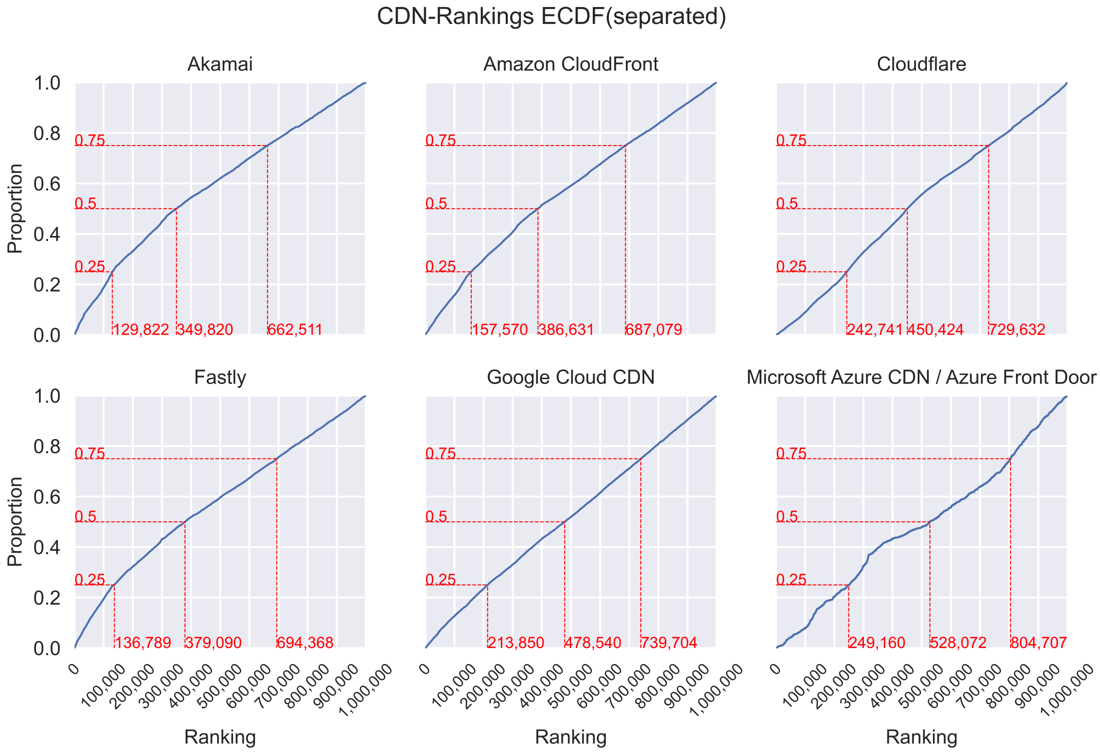


FIGURE 4.4: CDN-Rankings ECDF(separated)

4.1.3 Categorized Distribution

To further analyze CDN usage across different categories of websites, we correlated the measurement results with domain classification data. This enabled us to better understand the market penetration and usage patterns of different CDN providers in various industries. FIGURE 4.5-4.7 illustrate the usage distribution of each CDN provider across different website categories, revealing their market strategies and user preferences.

As shown in FIGURE 4.5, Cloudflare dominates across all categories, particularly in the "General Business" category, where its usage far surpasses that of other CDN providers. This further demonstrates Cloudflare's broad applicability and widespread recognition, indicating its strong appeal among a wide range of commercial websites. This might be attributed to the attractiveness of its free plan. Notably, Akamai has the highest usage in the "Cloud" category, suggesting its greater popularity among technology-intensive and cloud computing-related websites. Additionally, Cloudflare's usage in the "Pornography" category is significantly higher than in other categories.

To more accurately analyze CDN usage among other categories, we excluded Cloudflare and the "General Business" category, as shown in FIGURE 4.6. It reveals some significant differences in CDN usage across various categories. For example, Amazon CloudFront has a higher usage in the "Education" category, likely due to the adoption of AWS by various third-party learning management system solutions, such as Instructure Canvas and Blackboard; Akamai holds a larger market share in the "Cloud" and "Shopping" categories. However, no single CDN provider dominates all categories. It is also noteworthy that, although Microsoft Azure CDN/Azure Front Door has a relatively low market share across categories, it maintains a presence in the "Banner Advertisements" category.

FIGURE 4.7 further clarifies the proportions of different website categories across various CDN providers. The separated CDN Usage Distribution of different website categories are shown in Appendix B.

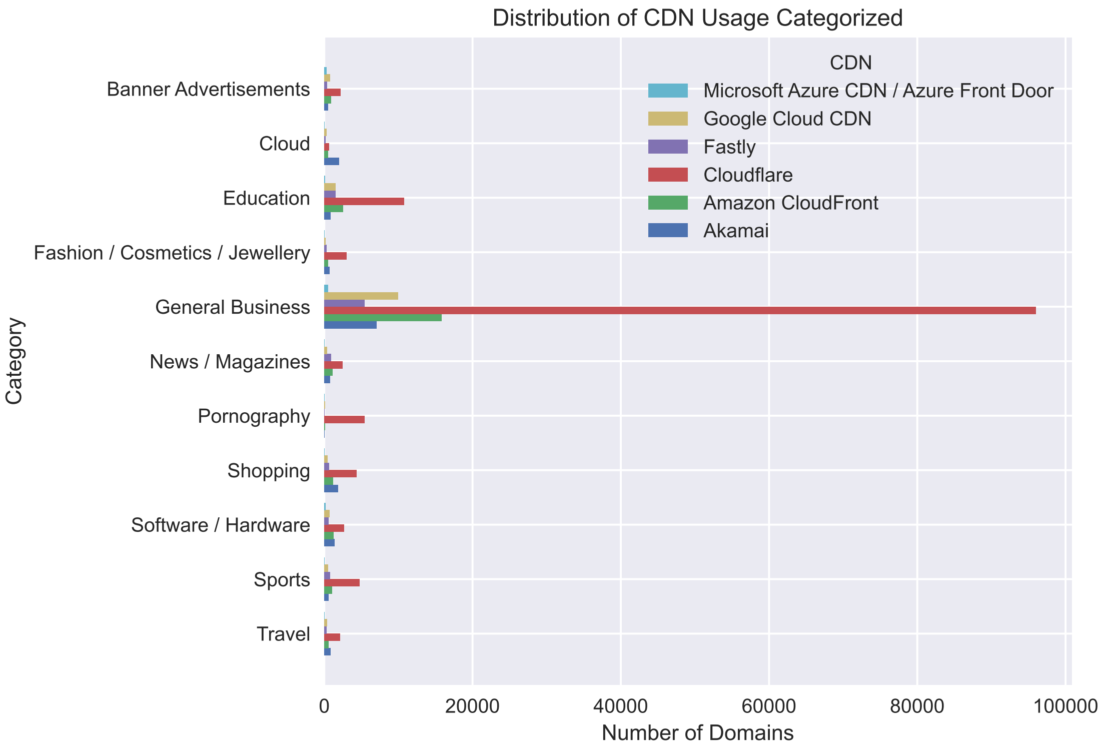


FIGURE 4.5: Distribution of CDN Usage Categorized

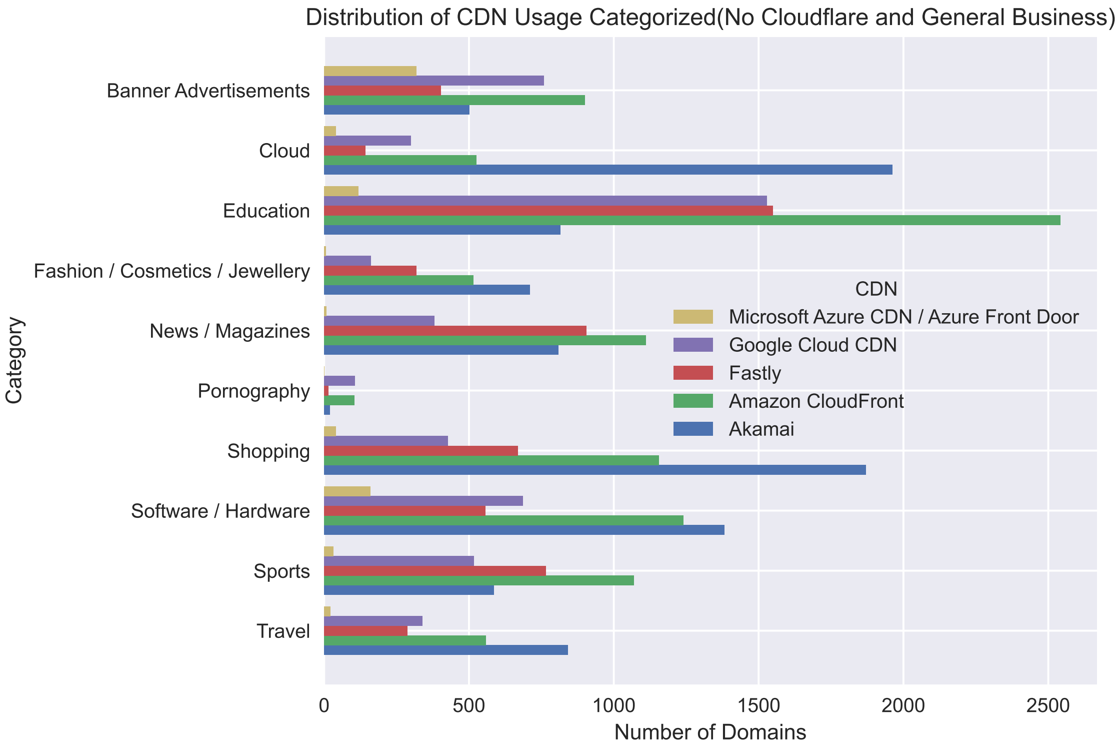


FIGURE 4.6: Distribution of CDN Usage Categorized(No Cloudflare and General Business)

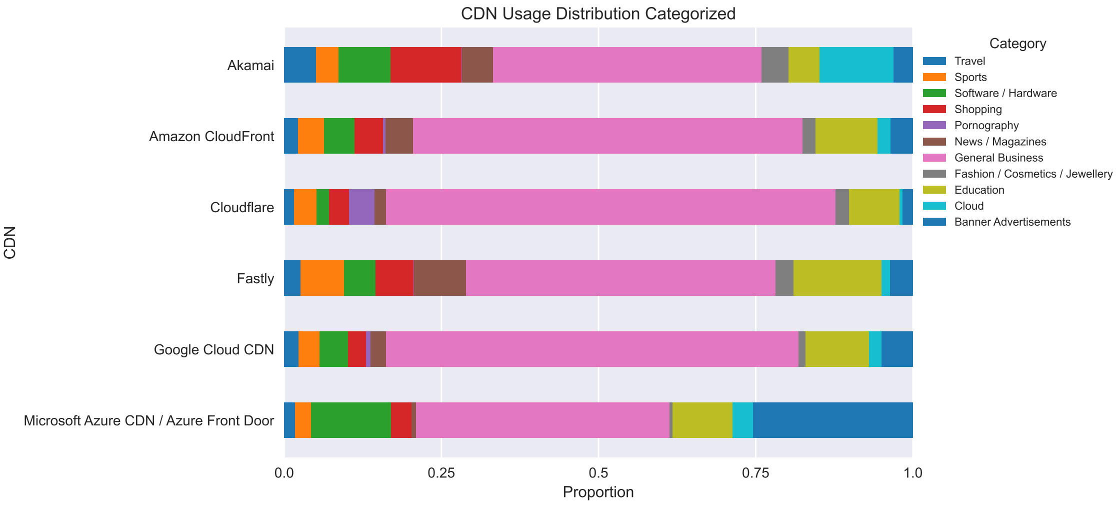


FIGURE 4.7: CDN Usage Distribution Categorized

4.2 CDN Features

4.2.1 Overall Usage

In the analysis of the usage of CDN paid features, we can observe the market strategies of various CDN providers in offering value-added services and advanced functionalities, as well as the distribution characteristics of their user bases. As shown in FIGURE 4.8, 22.44% of the measured websites utilize at least one detectable paid feature offered by CDNs. FIGURE 4.9 shows the distribution of paid CDN users, We reviewed the measurement data for the period from July 1st to July 7th. The data analysis reveals that, aside from minor fluctuations observed in Google Cloud CDN, the proportions of other CDN users using paid features remained consistent and without any visually perceptible variations. The stability of the data further confirms the reliability of our approach in capturing accurate and representative measurements of CDN paid features.

As shown in FIGURE 4.9, significant differences in market share for paid features among CDN providers are evident. Fastly and Amazon CloudFront exhibit higher proportions of paid feature usage, particularly Fastly, where over 75% of its users utilize paid features. This indicates that Fastly and Amazon CloudFront have a significant market advantage in offering certain advanced features and services, likely due to their ability to meet the high traffic and performance demands of websites. Additionally, Google Cloud CDN also shows a notable market share in paid feature usage, highlighting its attractiveness for advanced feature needs. Cloudflare and Akamai have relatively lower proportions of paid feature usage. Despite Cloudflare's market dominance, a large portion of its users may prefer its free basic features, which aligns with its market strategy and broad user base. For Akamai, the lower usage of paid features might be due to its comprehensive basic offerings, reducing the need for additional paid features. Microsoft Azure CDN/Azure Front Door has the lowest proportion of paid feature usage, consistent with its lower overall market share.

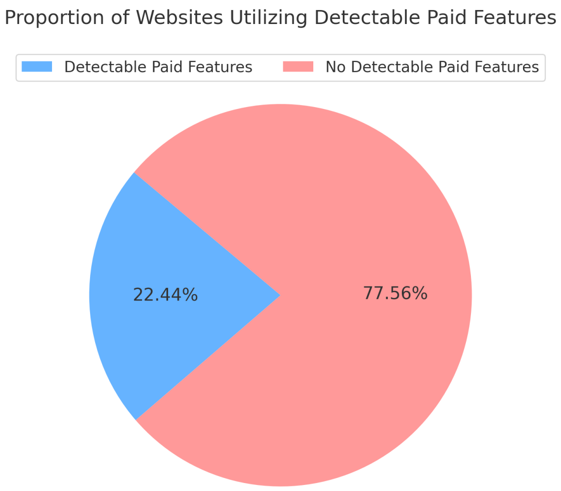


Figure 4.8: Proportion of Websites Utilizing Detectable Paid Features

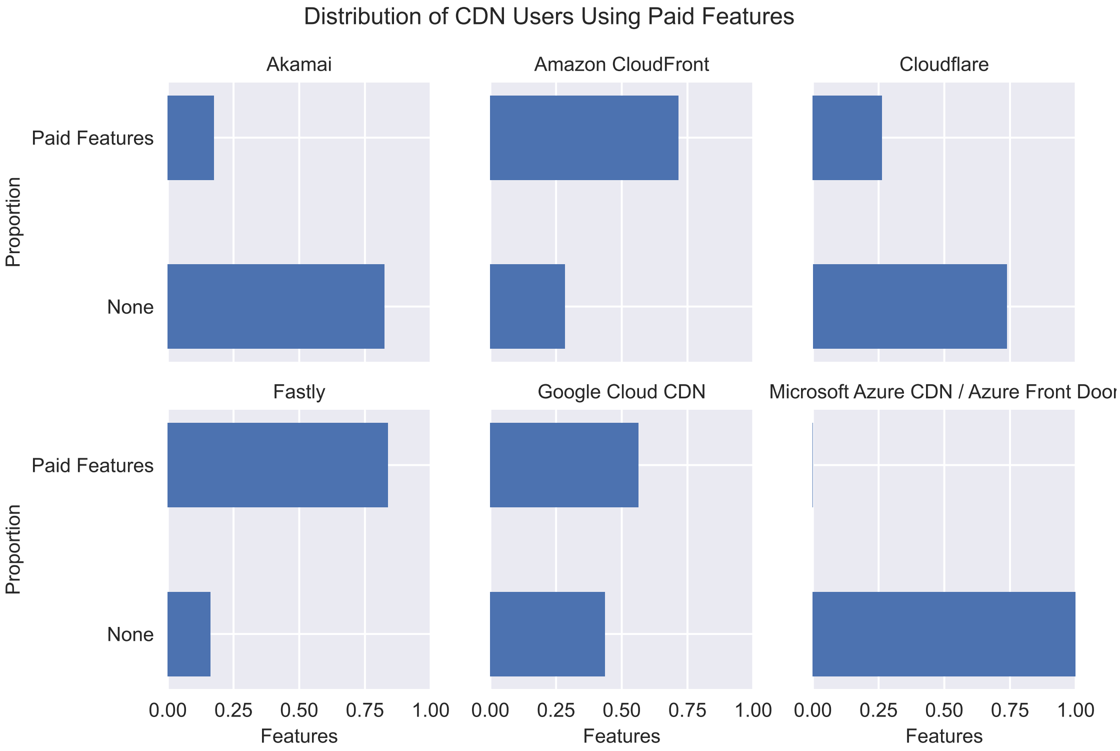


Figure 4.9: Distribution of CDN Users Using Paid Features

4.2.2 Ranking Distribution

FIGURE 4.10 provides an overall comparison of all CDN providers. Akamai distinctly shows a higher proportion of paid feature usage among high-ranking websites, with the steepest curve indicating that its paid features are primarily adopted by high-ranking, high-traffic websites. Fastly, Amazon CloudFront, and Google Cloud CDN follow, with notable usage of paid features in mid-to-high-ranking websites. Cloudflare's paid feature usage curve fluctuates more, with coverage across websites of various rankings, showing a more pronounced presence in mid-to-high-ranking websites but with an overall lower proportion compared to the aforementioned CDN providers. Due to the minimal number of Microsoft Azure CDN/Azure Front Door users using paid features, it is omitted from the chart.

Finally, FIGURE 4.11 further illustrates the usage of paid features by different CDN providers across various ranking intervals. Akamai's paid features show the highest concentration among high-ranking websites compared to other CDN providers. Its Edge DNS and Global Traffic Management features are significantly used in high-ranking websites. Amazon CloudFront, Fastly, and Google Cloud CDN also exhibit high usage proportions of paid features in mid-to-high-ranking websites. Cloudflare offers a range of paid features, such as Advanced/Custom Certificates, Partial (CNAME) setup, and Secondary DNS, with usage distributed across different ranking intervals but showing some concentration in mid-to-high-ranking websites. Notably, Cloudflare's Spectrum feature is predominantly used by mid-to-low-ranking websites, possibly due to our ASN-based identification method, which may overlook high-ranking websites that use their own IP addresses, which may not be announced in the same ASN, as explained in Cloudflare Spectrum – BYOIP[8]. Due to the minimal usage of paid features by Microsoft Azure CDN/Azure Front Door users, its result is omitted from the chart.

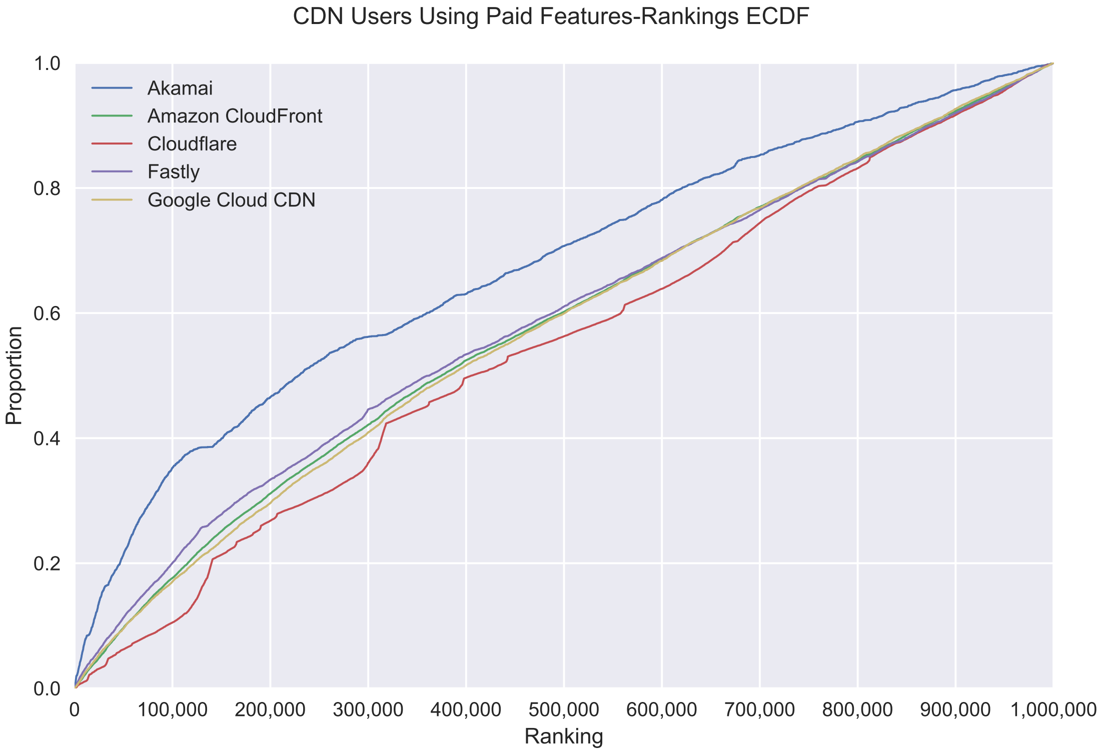


FIGURE 4.10: CDN Users Using Paid Features-Rankings ECDF

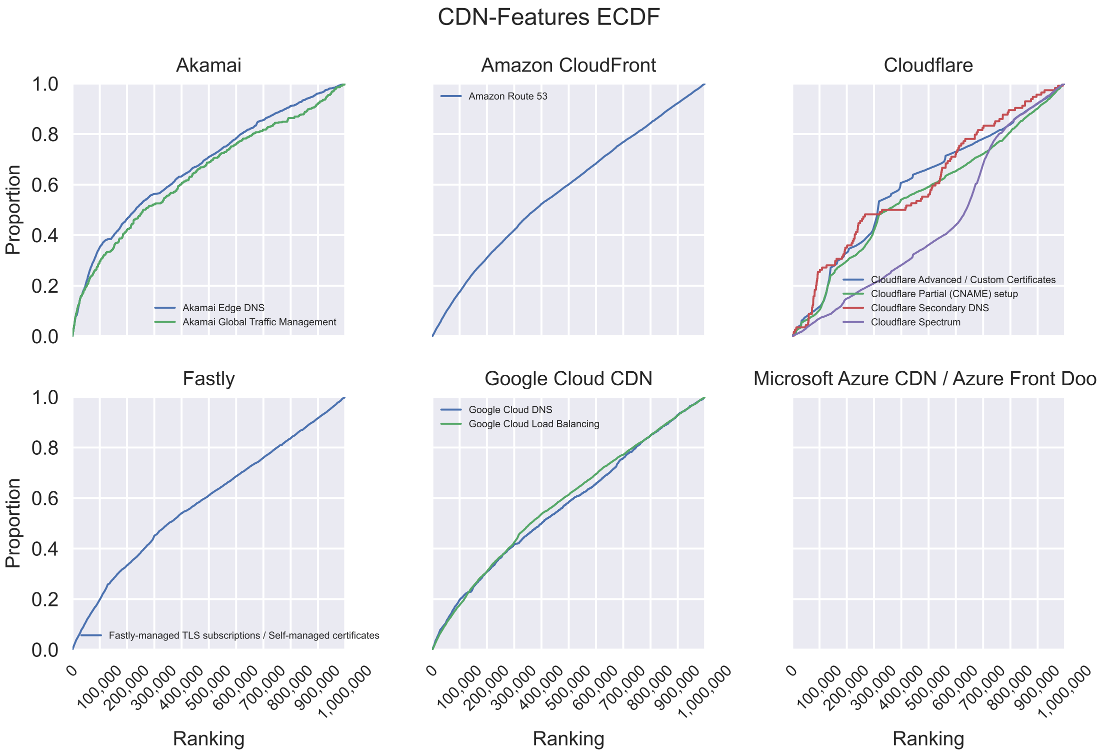


FIGURE 4.11: CDN-Features ECDF

4.2.3 Detailed Usage

Next, we detailed the usage proportions of paid features for users of each CDN service provider. Based on the six charts in Appendix C illustrating the usage of paid features by different CDN users, we can observe that Akamai's paid features usage primarily concentrates on Edge DNS and Global Traffic Management. These two features are evidently the most popular advanced services among Akamai users. However, most users still opt not to use any additional paid features that we can detect, indicating that its core user base may rely more on its robust basic services. For Amazon CloudFront, Route 53 is the most commonly used paid feature, occupying a large proportion. In contrast, the usage rates of other paid features are almost negligible. This suggests that Route 53 has a very high acceptance and usage rate among Amazon CloudFront users, likely due to its advantages in DNS management and traffic routing. Cloudflare's chart shows frequently used paid features, including Advanced/Custom Certificates, Partial (CNAME) Setup, and Spectrum. However, the vast majority of users still do not use any detectable paid features, consistent with Cloudflare's strategy of offering a wide range of free services in the market. A significant proportion of Fastly users choose to use paid certificate features, with Amazon Route 53 also having a notable usage proportion. For both Cloudflare and Fastly, paid certificates are the most commonly used paid feature, occupying the majority proportion. This may be due to the need for managing multiple domains/subdomains, with paid certificates allowing them to choose more certificate authorities, use dedicated IP certificates, and benefit from automatic renewals. Google Cloud CDN users primarily utilize paid features such as Google Cloud Load Balancing and Google Cloud DNS. These two features are evidently the most popular advanced services among Google Cloud CDN users, likely due to their advantages in load balancing and DNS management. It is noteworthy that Amazon Route

53 still has a certain proportion among its user base, slightly lower than Google Cloud DNS. Other paid features have lower usage rates but still show some diversity. Microsoft Azure CDN/Azure Front Door's chart shows that its users almost do not use any paid features. This might reflect Azure's weaker competitiveness in the CDN market or that its users rely more on basic services.

4.2.4 Feature Combinations

To further analyze websites using the combination of CDN paid features from multiple different CDN service providers, we created a corresponding heatmap. This heatmap allows us to visually observe the popularity of various CDN feature combinations. As shown in FIGURE 4.12, Amazon Route 53 is one of the most popular DNS services, especially for users of Fastly's paid TLS certificates. The number of users utilizing Amazon Route 53 far exceeds those using other DNS services. Upon reviewing Fastly's documentation, we found that Fastly uses Amazon Route 53 as the default DNS service provider for CDN configuration in its official documentation[13]. This likely explains the high popularity of this combination in the heatmap. Additionally, Fastly does not offer paid DNS services, which may indirectly lead users with advanced DNS service needs to choose other DNS providers. The combination of Amazon Route 53 and Cloudflare Spectrum is also quite popular, surpassing other combinations of Cloudflare's paid features. However, we could not find a reasonable explanation for this result. The combination of Google Cloud Load Balancing and Google Cloud DNS is more popular compared to its combination with Amazon Route 53, indicating Google Cloud CDN's competitiveness in providing an integrated solution and its users' reliance on load balancing and DNS services.

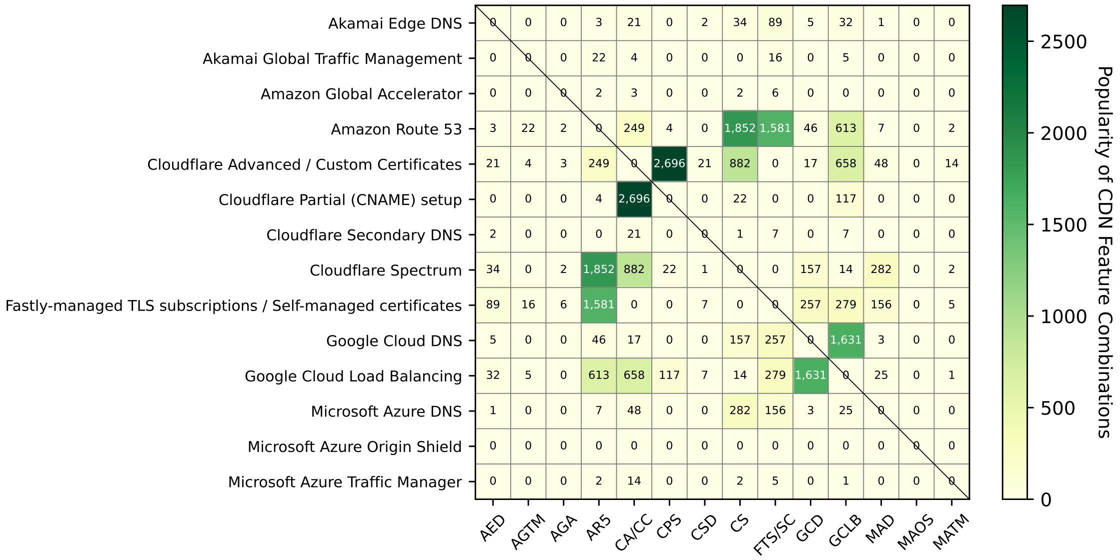


FIGURE 4.12: Popularity Heatmap of CDN Feature Combinations

4.2.5 Categorized Usage

FIGURE 4.13-4.15 illustrate the proportions of CDN features by corresponding website categories, we can gain deeper insights into the usage of each CDN provider's features across different website categories. As shown in FIGURE 4.13, Amazon Route 53 is the most popular among all identified features. Websites in the General Business category make up the majority of users utilizing CDN paid features.

As shown in FIGURE 4.14, after removing the General Business category, we can more clearly see the specific usage in other categories. Amazon Route 53 shows a high usage rate across multiple categories, particularly in the Education category, where its usage far exceeds other categories. This indicates a high acceptance of Amazon Route 53 in technology-intensive fields. Additionally, it is evident that educational websites have a significantly higher number of paid feature usages compared to other categories. This may be due to the complex domain setups, intricate network structures, and numerous infrastructures of educational websites, leading to a higher demand for paid CDN features.

FIGURE 4.15 further clarifies the proportions of different website categories across various CDN feature users.

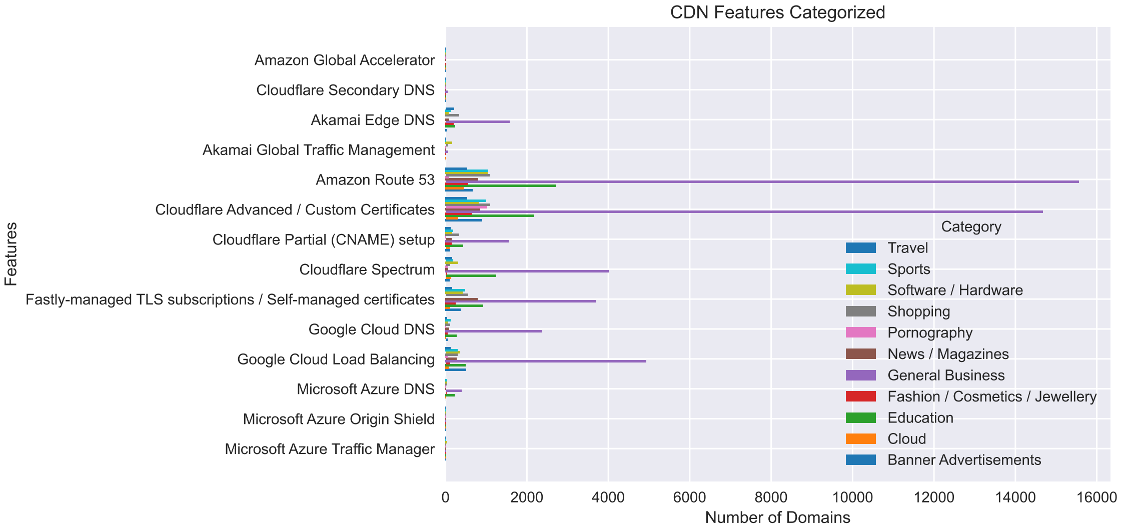


FIGURE 4.13: CDN Features Categorized

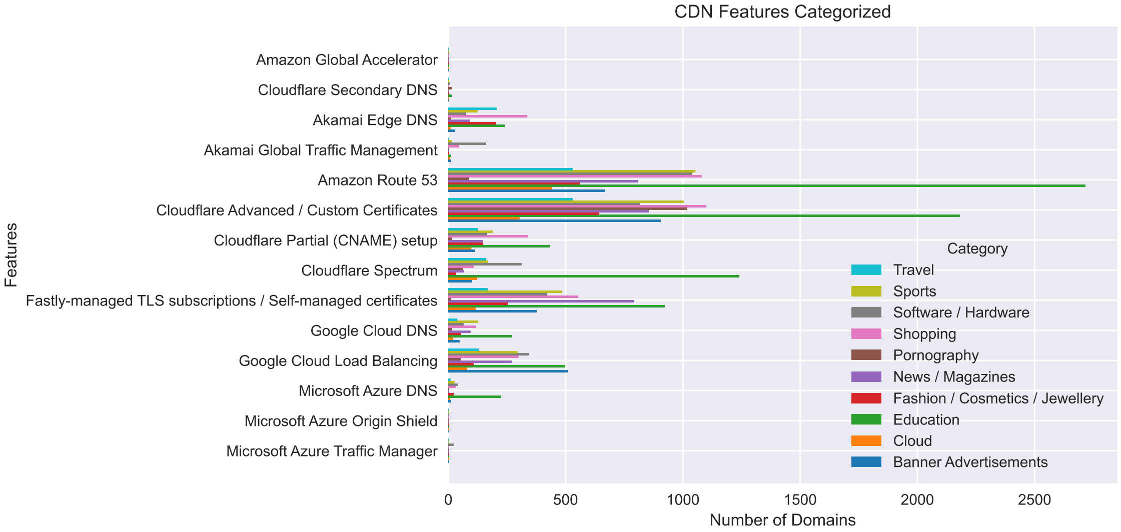


Figure 4.14: CDN Features Categorized(no general business)

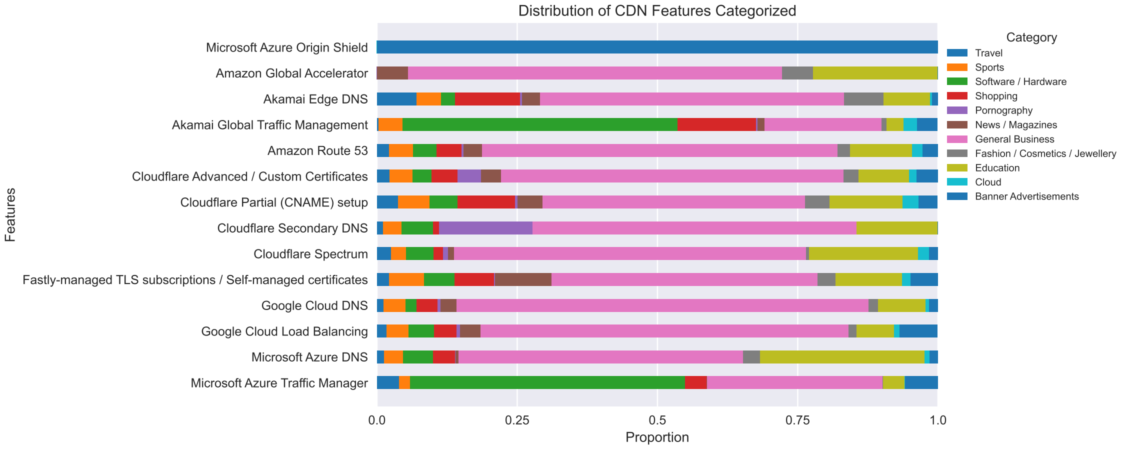


Figure 4.15: Distribution of CDN Features Categorized

4.3 WAF & Bot

4.3.1 Overall Usage

We also did separate analysis on CDN WAF and Bot Features. FIGURE 4.16 show the usage distribution of CDN WAF features, from which we can draw the following key conclusions: Cloudflare's WAF (including paid WAF) dominates overall usage, accounting for over 90% of the detectable WAF products. The usage rates of AWS WAF and Akamai WAF are similar, with Akamai WAF having a slightly higher user proportion than AWS WAF, but both fall far behind Cloudflare.

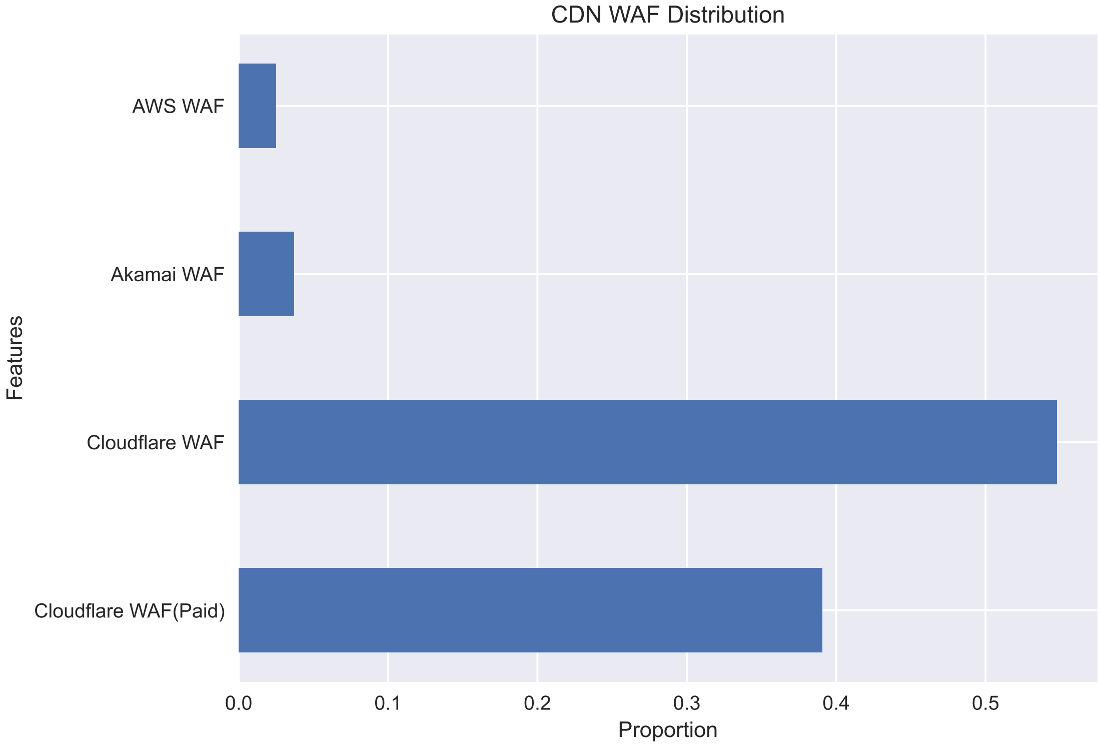


Figure 4.16: CDN WAF Distribution

4.3.2 Categorized Usage

FIGURE 4.17 provides a more intuitive view of the usage of various WAF features across different categories. Including the General Business category, Cloudflare WAF and Cloudflare WAF (Paid) have significantly higher user numbers in the General Business category compared to other categories. Moreover, the General Business category websites also dominate the overall market, particularly evident in the usage of Cloudflare's WAF features.

After removing the General Business category, as shown in FIGURE 4.18, we can more clearly see the specific usage in other categories. Cloudflare WAF and Cloudflare WAF (Paid) continue to dominate most categories, particularly in Education, Shopping, and Sports. Akamai WAF has higher user numbers across categories than AWS WAF, especially in the Shopping category. Additionally, Cloudflare WAF (Paid) has higher usage numbers in all categories compared to the free version. Notably, Education and Shopping websites show significantly higher usage of WAF products than other categories, indicating a higher demand for advanced security features in these categories. For instance, Shopping websites often require protection against bots and complex web attacks, making them more likely to pay for advanced WAF features.

FIGURE 4.19 further clarifies the proportions of different website categories across various CDN WAF products.

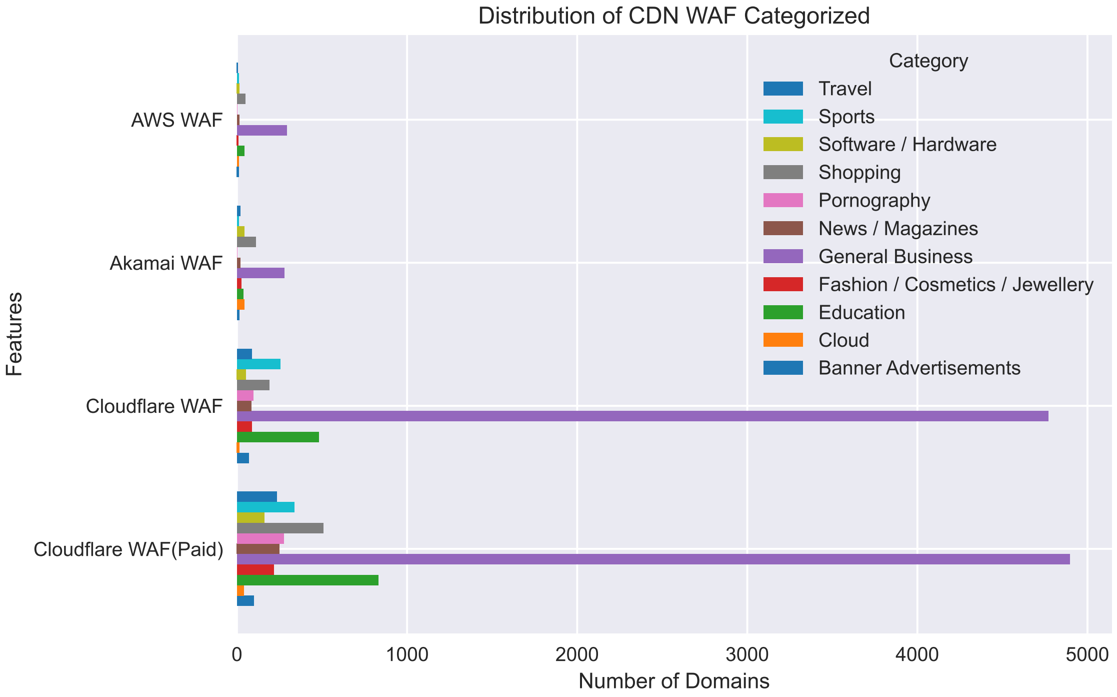


FIGURE 4.17: Distribution of CDN WAF Categorized

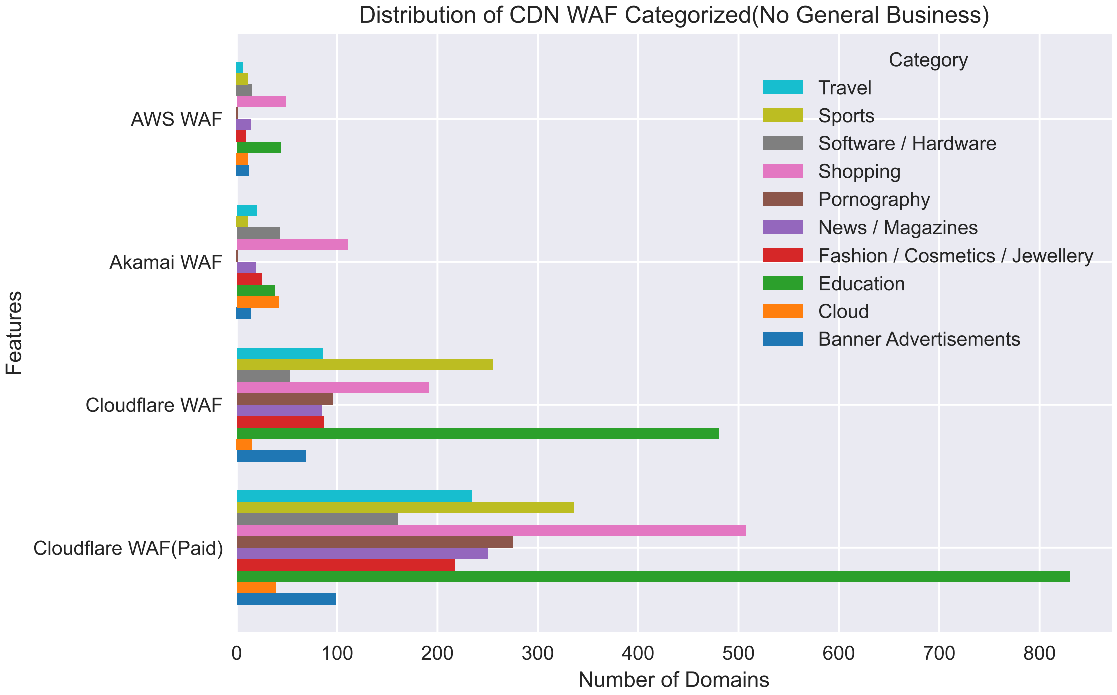


FIGURE 4.18: Distribution of CDN WAF Categorized(no general business)

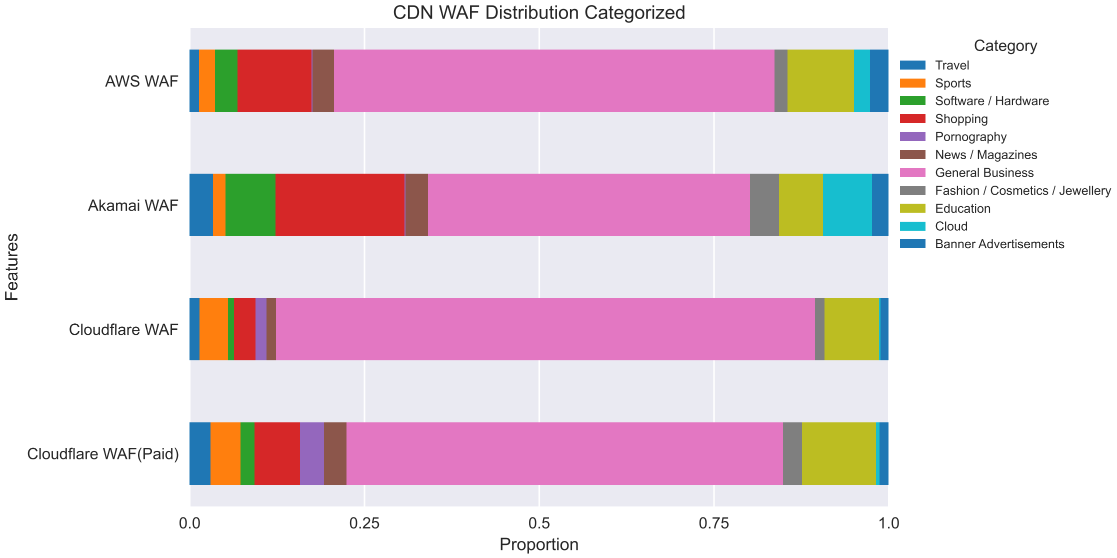


Figure 4.19: CDN WAF Distribution Categorized

4.3.3 Ranking Distribution

To gain a deeper understanding of the usage of WAF products by different CDN providers across various website rankings, we created an ECDF chart for CDN WAF products. As FIGURE 4.20-4.22 illustrate, the higher the website ranking, the higher the proportion of paid WAF usage. These high-ranking websites are likely more concerned with security and opt for paid WAF products. This is particularly evident in Cloudflare's chart, where the curve for Cloudflare's paid WAF shows a significantly higher user proportion among the top 300,000 websites compared to the free Cloudflare WAF. Paid WAF usage holds a larger market share among high-ranking websites, while the free WAF is more widely used among mid- to low-ranking websites. This also indicates that high-ranking websites tend to choose paid WAF services to enhance security, whereas mid- to low-ranking websites more often opt for free or basic WAF services.

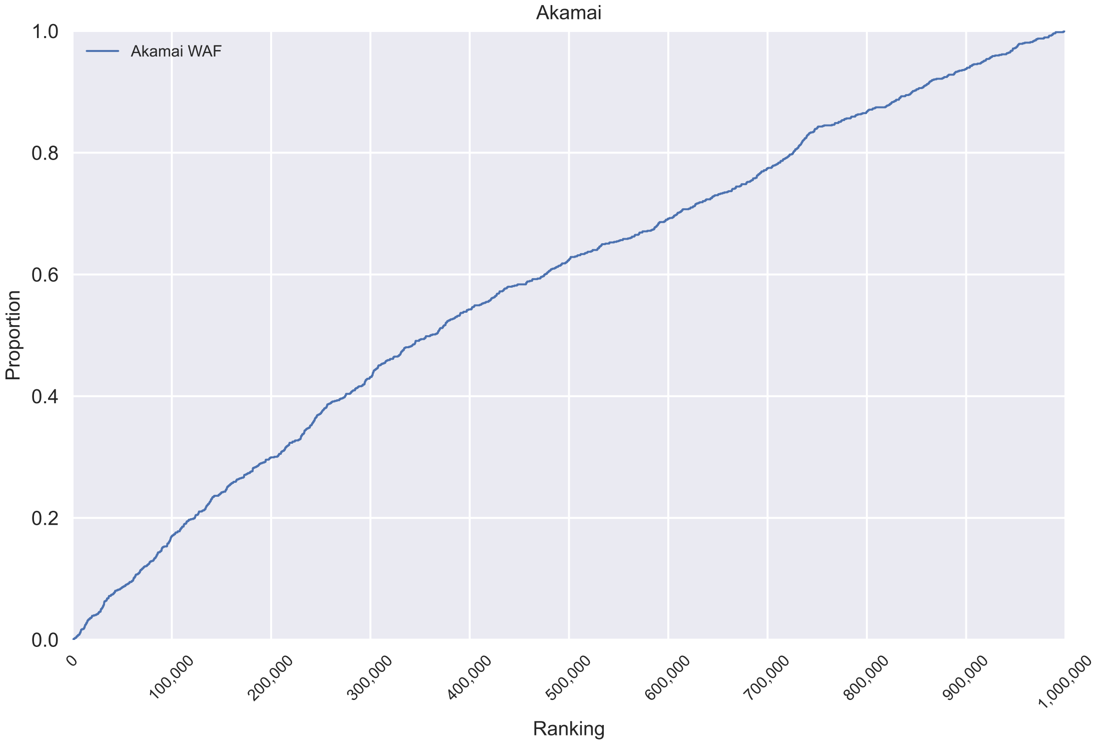


FIGURE 4.20: CDN WAF ECDF-Akamai

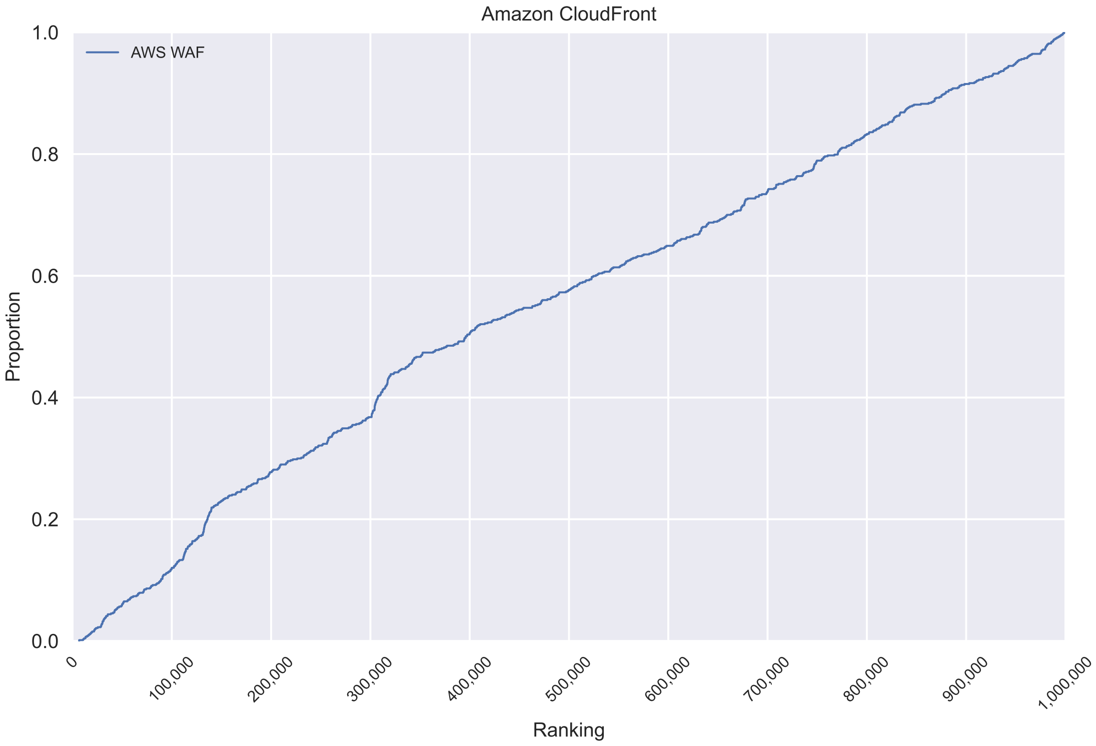


FIGURE 4.21: CDN WAF ECDF-AWS

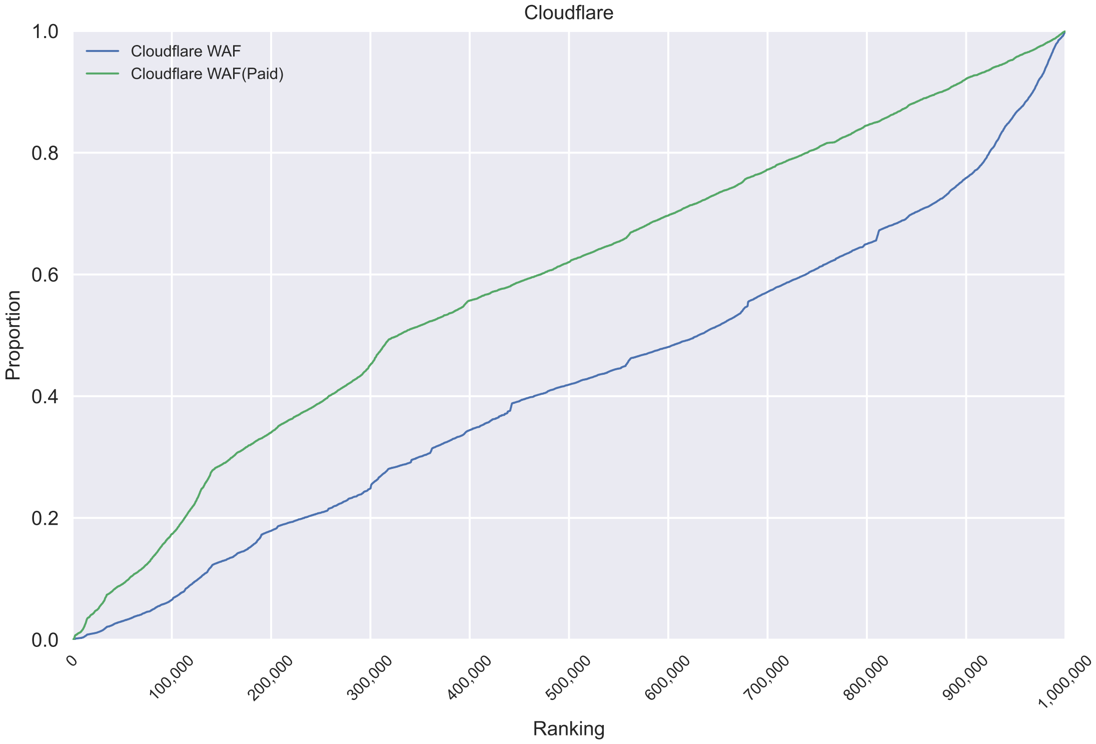


Figure 4.22: CDN WAF ECDF-Cloudflare

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we have conducted a detailed study of Content Delivery Networks (CDNs), focusing on their usage distribution, features, and security configurations. Through our methodology, we identified and analyzed the major CDN service providers and their market shares, investigated the adoption rates of paid CDN features, and evaluated the protection configurations such as Web Application Firewalls (WAF) and bot management mechanisms.

Our findings provide valuable insights into the CDN ecosystem. Cloudflare emerged as a dominant player in the CDN market, with a substantial share among both high-ranking and lower-ranking websites. Other major providers like Amazon CloudFront, Akamai, Google Cloud CDN, Fastly, and Microsoft Azure CDN also demonstrated significant presence, each with unique distribution patterns and feature adoption rates.

The analysis of paid CDN features highlighted the varying strategies and offerings of different providers. Fastly and Amazon CloudFront showed higher proportions of paid feature usage, particularly among high-traffic websites. Cloudflare, despite its widespread free service adoption, also had notable usage of advanced features among mid-to-high-ranking websites. The popularity of certain features, such as Amazon Route 53 and Fastly's TLS certificates, underscored the importance of DNS and certificate management services in the CDN landscape.

Our examination of WAF and bot protection configurations revealed that Cloudflare's WAF solutions are widely used, with a significant proportion of websites employing these features to enhance security. The use of paid WAF features was particularly prevalent among high-ranking websites, indicating a greater emphasis on security and performance optimization in these domains.

Overall, this research contributes to the understanding of the CDN market, highlighting key trends, usage patterns, and the significance of advanced CDN features. These insights can help website operators, security professionals, and businesses make informed decisions regarding the selection and deployment of CDN services to optimize web performance, security, and user experience.

5.2 Future Work

While this thesis has contributed valuable insights into the understanding of CDN usage and security, there are several areas for future research that can enhance our knowledge and address some limitations encountered in this study.

- **Improvement of Detection Methods**: Future work should aim to refine and enhance the detection and identification methods used in this study. This could involve developing more sophisticated algorithms and techniques to improve accuracy and reliability in identifying CDN usage, paid features, and security configurations. Enhanced detection methods would provide a clearer and more detailed understanding of the CDN landscape.

- **Extended Measurement Period**: Future studies could benefit from conducting measurements over an extended period to provide more robust data and allow for the observation of temporal trends and fluctuations in CDN usage and feature adoption. This would help in understanding the dynamic nature of the CDN market more comprehensively.

- **Large Scale Measurement**: Conducting measurements on a larger scale would provide a more comprehensive picture of CDN usage patterns. Specifically, analyzing CDN usage across different top-level domains (TLDs), such as .com, .net, .org, and country-specific TLDs, could reveal variations in CDN adoption and feature utilization across different types of websites.

- **Security Threat Analysis**: Conducting an in-depth analysis of emerging security threats and attack vectors targeting CDNs. Future research could evaluate the effectiveness of current WAF and bot protection mechanisms and explore new approaches to enhance CDN security. This would help in developing more robust defenses against evolving cyber threats.

By addressing these areas, future research can build on the foundation laid by this thesis, providing deeper insights and practical recommendations for the effective use and management of CDN services in today's digital landscape.

Bibliography

[1] Babak Amin Azad, Oleksii Starov, Pierre Laperdrix, and Nick Nikiforakis. Web runner 2049: Evaluating third-party anti-bot services. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA 2020, Lisbon, Portugal, June 24–26, 2020, Proceedings 17*, pages 135–159. Springer, 2020.

[2] AWS. What are aws waf, aws shield advanced;, and aws firewall manager? URL: https://docs.aws.amazon.com/waf/latest/developerguide/what-is-aws-waf.html.

[3] Thomas Barnett, Shruti Jain, Usha Andra, and Taru Khurana. Cisco visual networking index (vni) complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, pages 1–30, 2018.

[4] BuiltWith. Verified cdn usage distribution in the top 1 million sites. URL: https://trends.builtwith.com/cdns.

[5] Elie Bursztein, Artem Malyshev, Tadek Pietraszek, and Kurt Thomas. Picasso: Lightweight device class fingerprinting for web clients. In *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 93–102, 2016.

[6] Elisa Chiapponi, Marc Dacier, Olivier Thonnard, Mohamed Fangar, Mattias Mattsson, and Vincent Rigal. An industrial perspective on web scraping characteristics and open issues. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)*, pages 5–8. IEEE, 2022.

[7] David Choffnes, Jilong Wang, et al. Cdns meet cn an empirical study of cdn deployments in china. *IEEE Access*, 5:5292–5305, 2017.

[8] Cloudflare. Byoip · cloudflare spectrum docs. URL: https://developers.cloudflare.com/spectrum/about/byoip/.

[9] Cloudflare. Cloudflare bot solutions. URL: https://developers.cloudflare.com/bots/.

[10] Cloudflare. Cloudflare web application firewall. URL: https://developers.cloudflare.com/waf/.

[11] Cloudflare. What is a content delivery network (cdn)? | how do cdns work? URL: https://www.cloudflare.com/learning/cdn/what-is-a-cdn/.

[12] Cloudfront. What is a cdn (content delivery network)? URL: https://aws.amazon.com/what-is/cdn/.

[13] Fastly. 5. go live | fastly documentation. URL: https://www.fastly.com/documentation/solutions/tutorials/introduction-to-cdn/5-go-live/.

[14] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. The lockdown effect: Implications of the covid-19 pandemic on internet traffic. In *Proceedings of the ACM internet measurement conference*, pages 1–18, 2020.

[15] Gartner. Global cdn reviews and ratings. URL: https://www.gartner.com/reviews/market/global-cdn.

[16] Milad Ghaznavi, Elaheh Jalalpour, Mohammad A Salahuddin, Raouf Boutaba, Daniel Migault, and Stere Preda. Content delivery network security: A survey. *IEEE Communications Surveys & Tutorials*, 23(4):2166–2190, 2021.

[17] Run Guo, Jianjun Chen, Baojun Liu, Jia Zhang, Chao Zhang, Haixin Duan, Tao Wan, Jian Jiang, Shuang Hao, and Yaoqi Jia. Abusing cdns for fun and profit: Security issues in cdns' origin validation. In *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*, pages 1–10. IEEE, 2018.

[18] Run Guo, Weizhong Li, Baojun Liu, Shuang Hao, Jia Zhang, Haixin Duan, Kaiwen Sheng, Jianjun Chen, and Ying Liu. Cdn judo: Breaking the cdn dos protection with itself. In *NDSS*, 2020.

[19] IBM. Ibm x-force threat intelligence api documentation. URL: https://api.xforce.ibmcloud.com/doc/.

[20] IDC. Idc marketscape: Worldwide commercial content delivery network services 2022 vendor assessment. URL: https://pages.awscloud.com/rs/112-TZM-766/images/IDC_MarketScape_CDN_2022_licensed.pdf.

[21] Fortune Business Insights. Content delivery network market share & size report, 2030. URL: https://www.fortunebusinessinsights.com/content-delivery-network-market-105949.

[22] Gregoire Jacob, Engin Kirda, Christopher Kruegel, and Giovanni Vigna. {PUBCRAWL}: Protecting users and businesses from {CRAWLers}. In *21st USENIX Security Symposium (USENIX Security 12)*, pages 507–522, 2012.

[23] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019. doi:10.14722/ndss.2019.23386.

[24] Xigao Li, Babak Amin Azad, Amir Rahmati, and Nick Nikiforakis. Good bot, bad bot: Characterizing automated browsing activity. In *2021 IEEE symposium on security and privacy (sp)*, pages 1589–1605. IEEE, 2021.

[25] Anália G Lourenço and Orlando O Belo. Catching web crawlers in the act. In *Proceedings of the 6th international Conference on Web Engineering*, pages 265–272, 2006.

[26] K Nagendran, S Balaji, B Akshay Raj, P Chanthrika, and RG Amirthaa. Web application firewall evasion techniques. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 194–199. IEEE, 2020.

[27] Evi Nemeth, Garth Snyder, Trent R Hein, Ben Whaley, and Dan Mackin. Unix and linux system administration handbook. *USENIX Open Access Policy*, 59, 2018.

[28] KyoungSoo Park, Vivek S Pai, Kang-Won Lee, and Seraphin B Calo. Securing web service by automatic robot detection. In *USENIX Annual Technical Conference, General Track*, pages 255–260, 2006.

[29] Grand View Research. Content delivery network market size & share report, 2030. URL: https://www.grandviewresearch.com/industry-analysis/content-delivery-networks-cnd-market.

[30] Scrapfly. How to bypass cloudflare when web scraping in 2024. URL: https://scrapfly.io/blog/how-to-bypass-cloudflare-anti-scraping/.

[31] Behnam Shobiri, Mohammad Mannan, and Amr Youssef. Cdns' dark side: Security problems in cdn-to-origin connections. *Digital Threats: Research and Practice*, 4(1):1–22, 2023.

[32] Rachee Singh, Arun Dunna, and Phillipa Gill. Characterizing the deployment and performance of multi-cdns. In *Proceedings of the Internet Measurement Conference 2018*, pages 168–174, 2018.

[33] Markus Sosnowski, Johannes Zirngibl, Patrick Sattler, Georg Carle, Claas Grohnfeldt, Michele Russo, and Daniele Sgandurra. Active tls stack fingerprinting: Characterizing tls server deployments at scale. *arXiv preprint arXiv:2206.13230*, 2022.

[34] Qi Wang, Jianjun Chen, Zheyu Jiang, Run Guo, Ximeng Liu, Chao Zhang, and Haixin Duan. Break the wall from bottom: Automated discovery of protocol-level evasion vulnerabilities in web application firewalls. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 129–129. IEEE Computer Society, 2024.

[35] Guowu Xie, Huy Hang, and Michalis Faloutsos. Scanner hunter: Understanding http scanning traffic. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 27–38, 2014.

[36] ZenRows. How to bypass cloudflare in 2024: The 9 best methods. URL: https://www.zenrows.com/blog/bypass-cloudflare#passive-detection-techniques.

[37] Minyuan Zhou, Jiaqi Zheng, Guihai Chen, and Wanchun Dou. An efficient approach for (multi-) cdn identification. In *Proceedings of the ACM Turing Award Celebration Conference-China 2023*, pages 156–157, 2023.

[38] Behrouz Zolfaghari, Gautam Srivastava, Swapnoneel Roy, Hamid R Nemati, Fatemeh Afghah, Takeshi Koshiba, Abolfazl Razi, Khodakhast Bibak, Pinaki Mitra, and Brijesh Kumar Rai. Content delivery networks: State of the art, trends, and future roadmap. *ACM Computing Surveys (CSUR)*, 53(2):1–34, 2020.

Appendix A

CDN Features Identification Methods

A.1 Akamai

Table A.1: CDN Features Identification - Akamai

Attribute	Value	Feature
CNAME	akadns.net	Akamai Global Traffic Management
NS	akam.net	Akamai Edge DNS

A.2 Amazon CloudFront

Table A.2: CDN Features Identification - Amazon CloudFront

Attribute	Value	Feature
CNAME	elb.amazonaws.com	Amazon Classic Load Balancer
PTR	awsglobalaccelerator.com	Amazon Global Accelerator
NS	one of the following: awsdns-[0-9]*.com awsdns-[0-9]*.net awsdns-[0-9]*.org awsdns-[0-9]*.co.uk	Amazon Route 53
CERT-san	IP address	Amazon CloudFront Dedicated IP Custom SSL

A.3 Cloudflare

Table A.3: CDN Features Identification - Cloudflare

Attribute	Value	Feature
CNAME	cdn.cloudflare.net	Cloudflare Partial (CNAME) setup (Business or Enterprise only)
ASD / ASN	CLOUDFLARESPECTRUM Cloudflare, Inc. / 209242	Cloudflare Spectrum
NS	secondary.cloudflare.com	Cloudflare Secondary DNS
CERT-issuer	not one of the following: Let's Encrypt Google Trust Services Sectigo DigiCert(Cloudflare)	Cloudflare Advanced / Custom Certificates
CERT-san	one of the following: multiple domain apexes multi-level subdomains IP address	Cloudflare Advanced / Custom Certificates

A.4 Fastly

Table A.4: CDN Features Identification - Fastly

Attribute	Value	Feature
CERT-issuer	not one of the following: Let's Encrypt Certainly	Fastly-managed TLS subscriptions / Self-managed certificates
CERT-san	more than two domains	Fastly-managed TLS subscriptions / Self-managed certificates
CERT-san	IP address	Fastly TLS Dedicated IP addresses

A.5 Google Cloud CDN

Table A.5: CDN Features Identification - Google Cloud CDN

Attribute	Value	Feature
Headers	Via: 1.1 google	Google Cloud Load Balancing
NS	googledomains.com	Google Cloud DNS

A.6 Microsoft Azure CDN

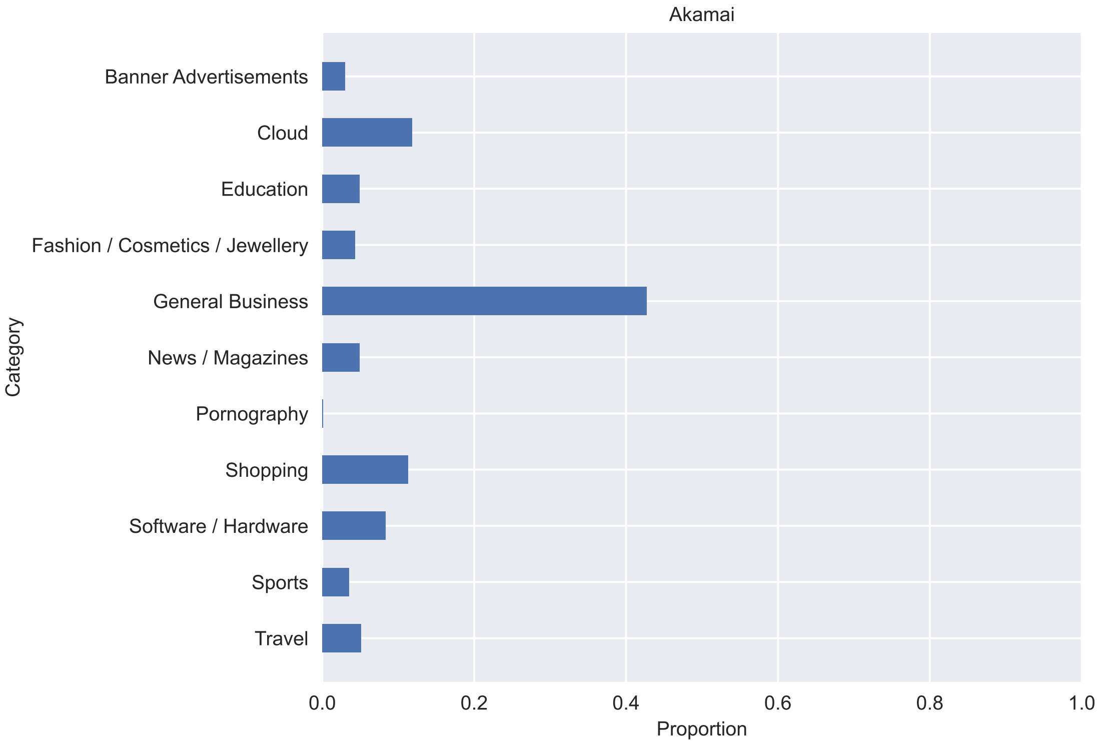
Table A.6: CDN Features Identification - Microsoft Azure CDN

Attribute	Value	Feature
CNAME	trafficmanager.net	Azure Traffic Manager
Headers	X-Azure-Ref-OriginShield	Azure Origin Shield
NS	one of the following: azure-dns.com azure-dns.net azure-dns.org azure-dns.info	Azure DNS

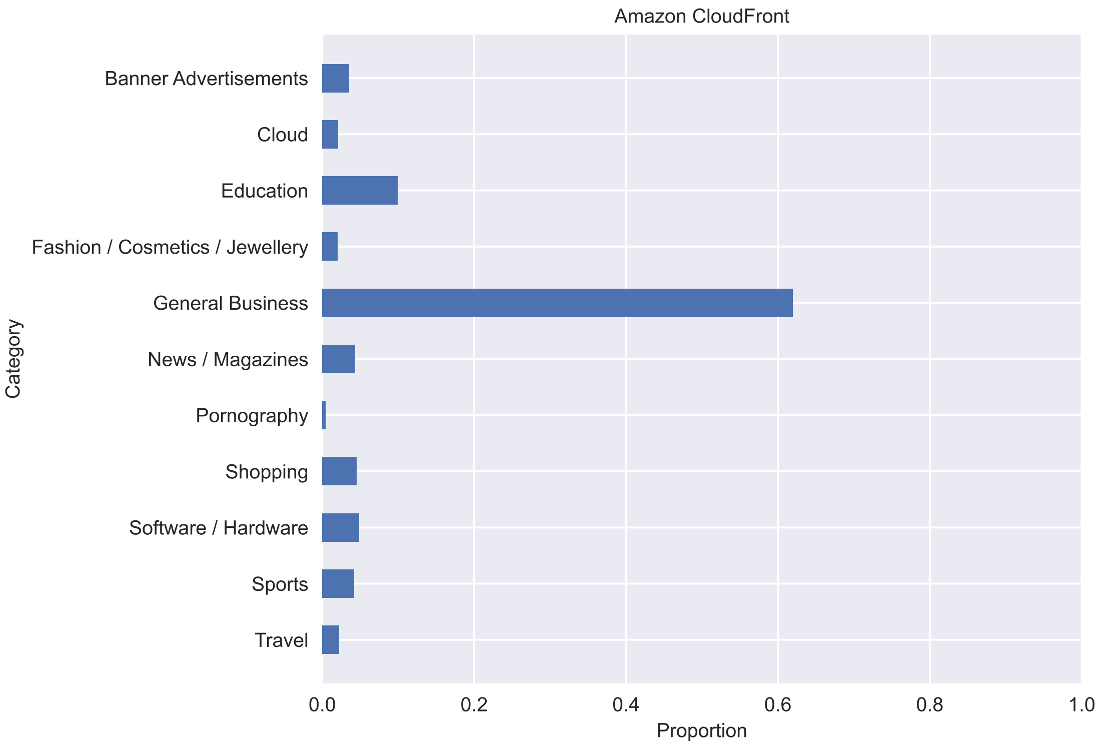
Appendix B

CDN Usage Distribution Categorized

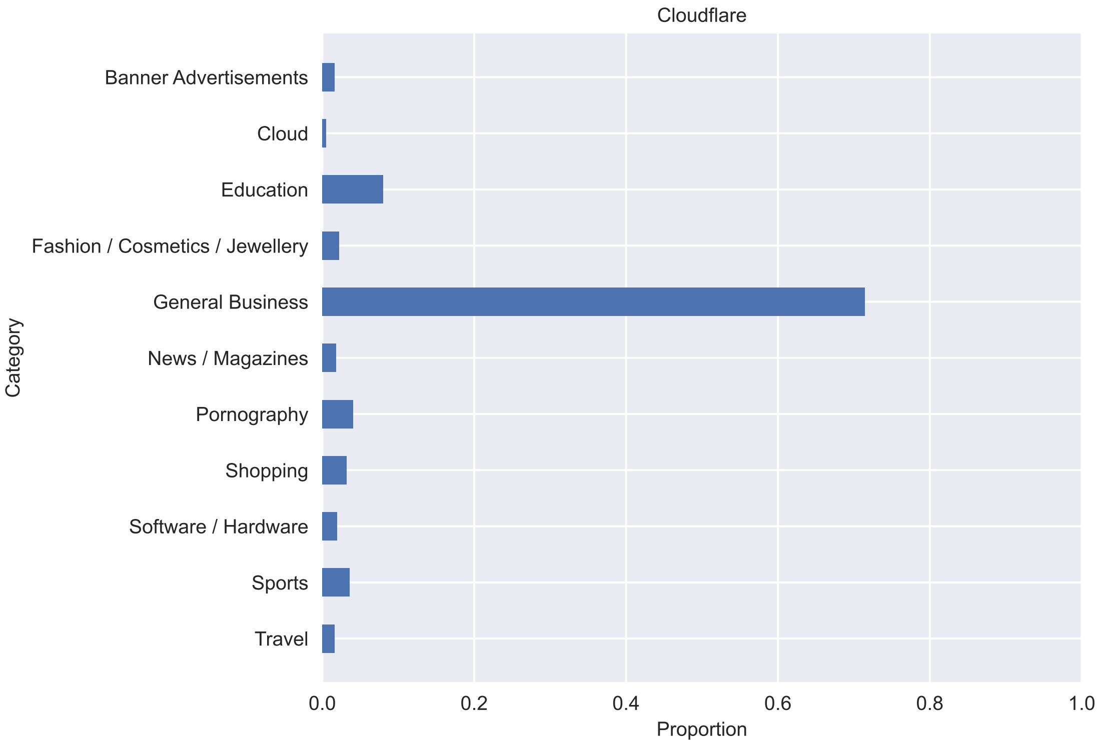
B.1 Akamai



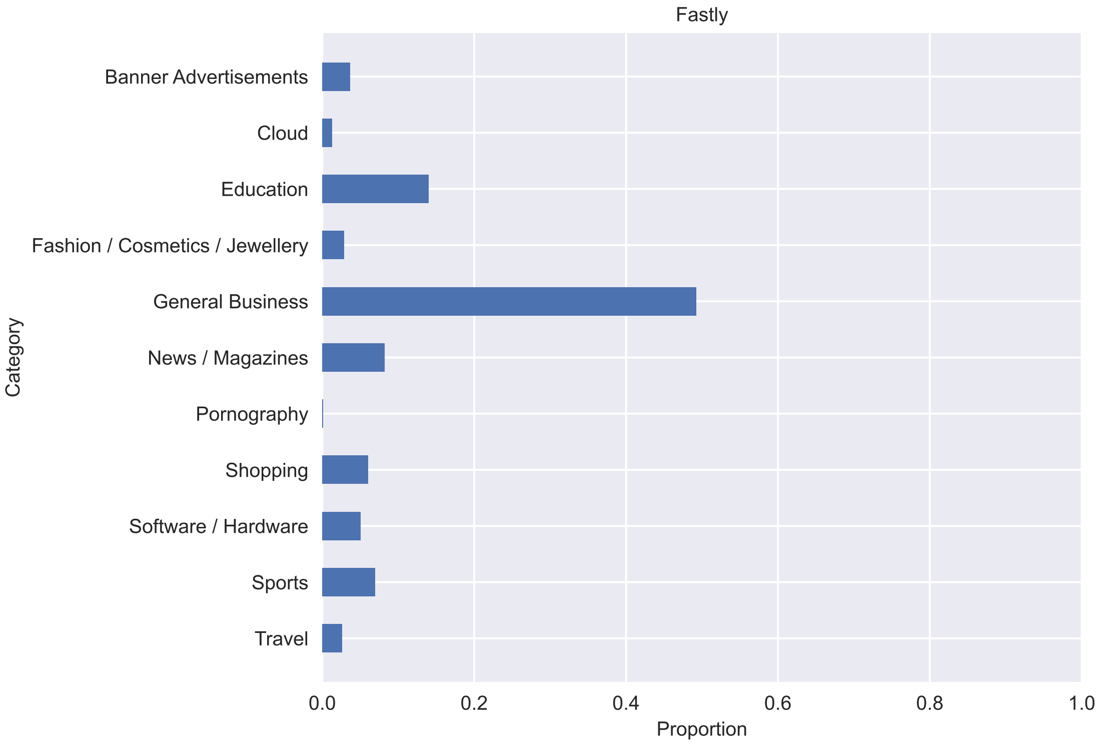
B.2 Amazon CloudFront



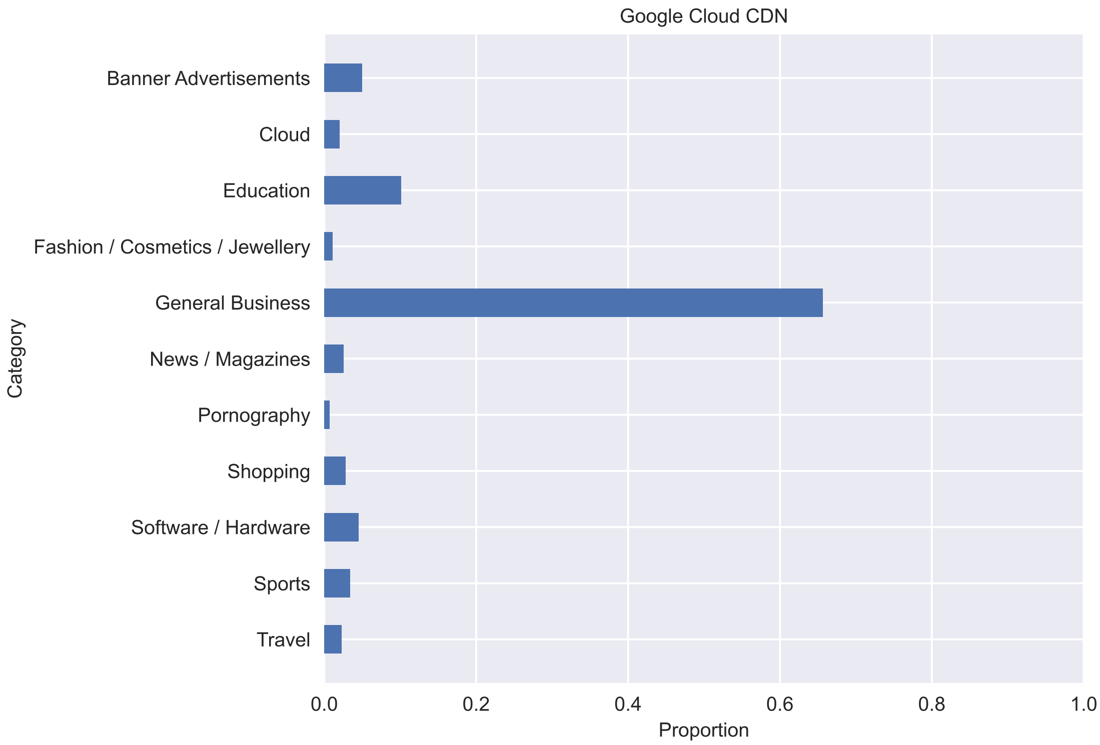
B.3 Cloudflare



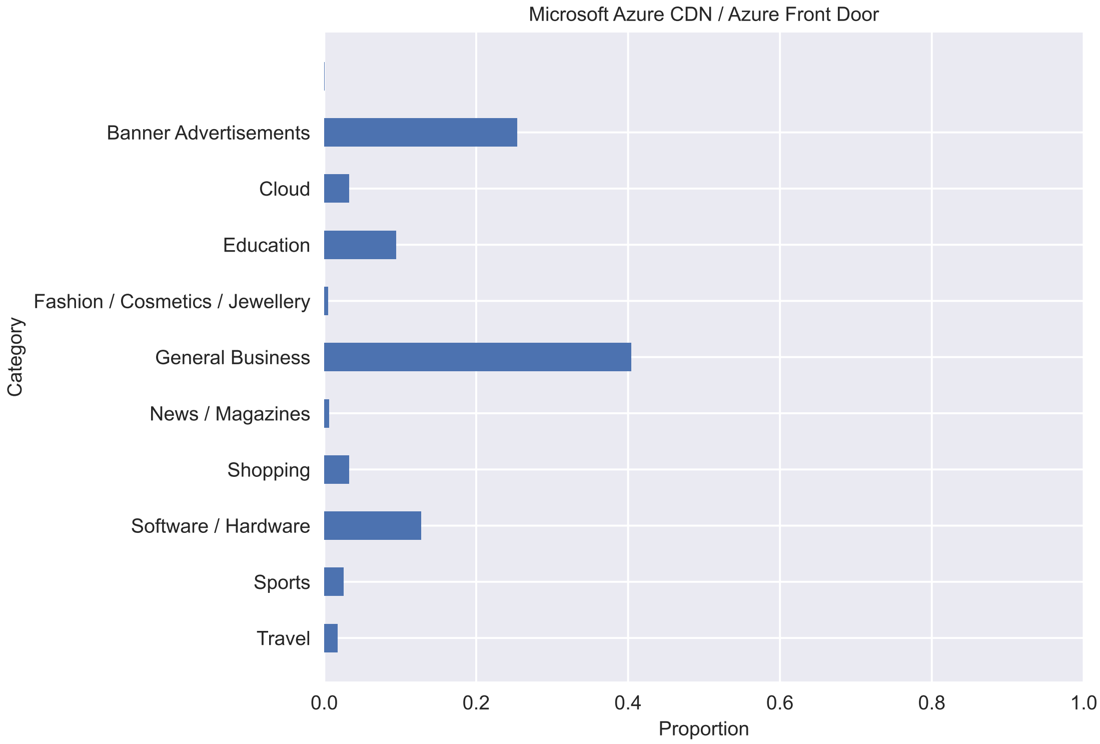
B.4 Fastly


Fastly
Banner Advertisements
Cloud
Education
Fashion / Cosmetics / Jewellery
General Business
News / Magazines
Pornography
Shopping
Software / Hardware
Sports
Travel
Category
0.0
0.2
0.4
0.6
0.8
1.0
Proportion

B.5 Google Cloud CDN



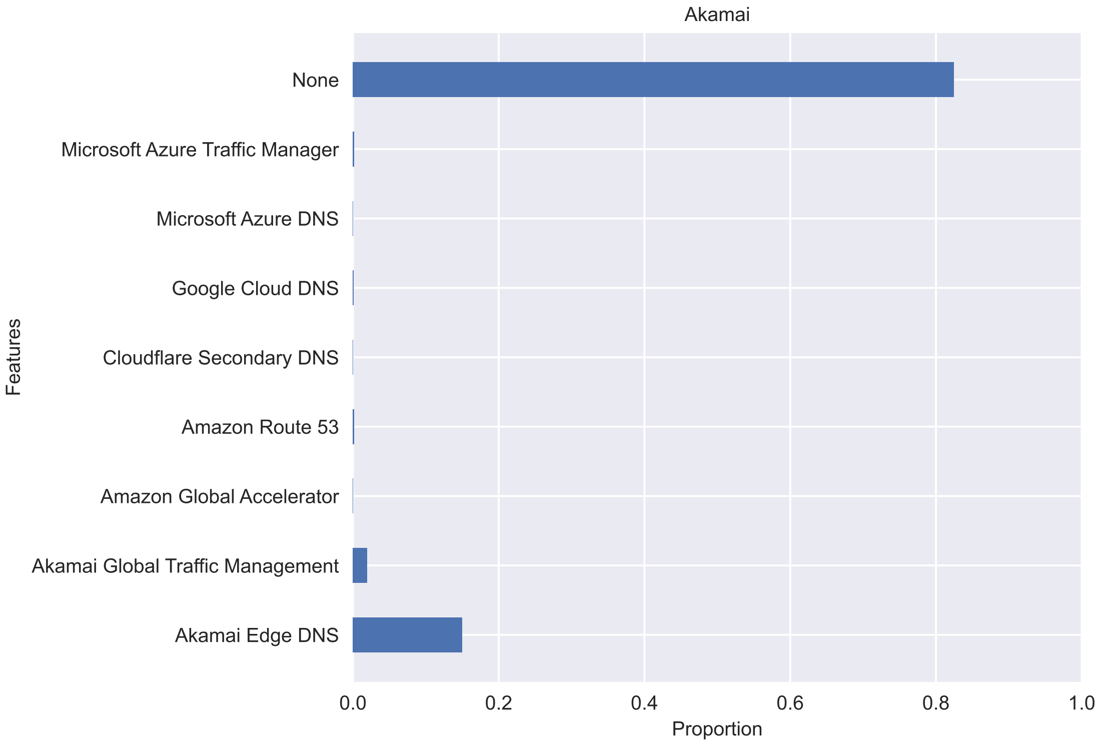
B.6 Microsoft Azure CDN



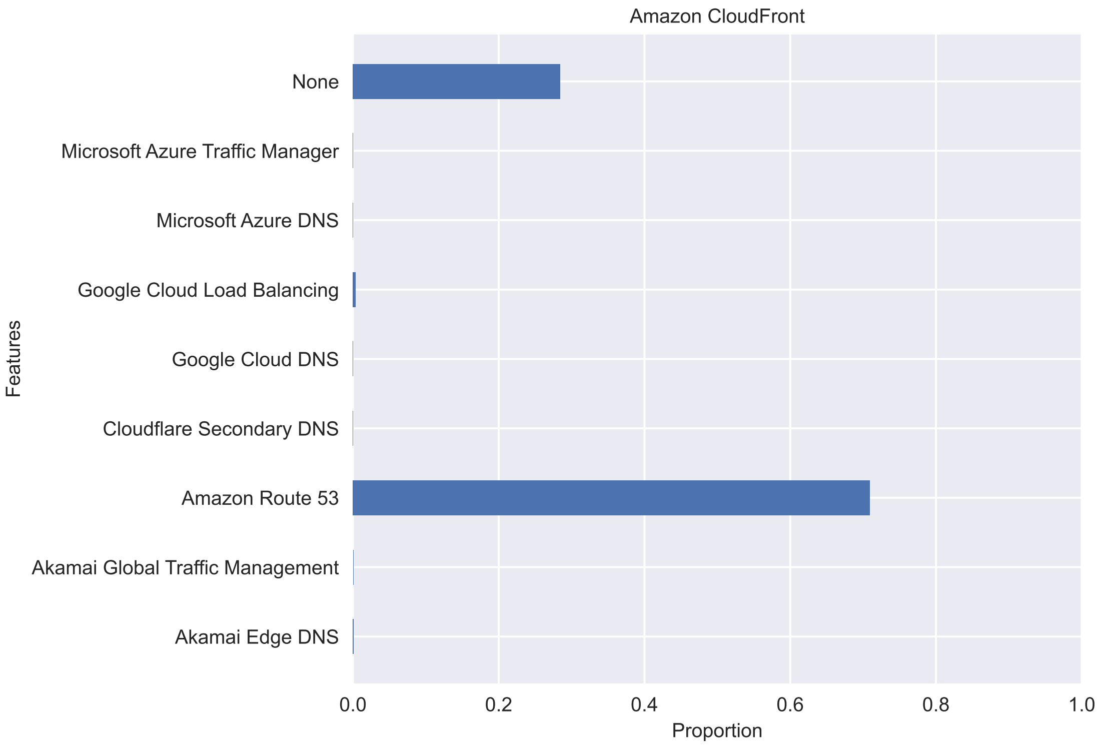
Appendix C

CDN Features Separated

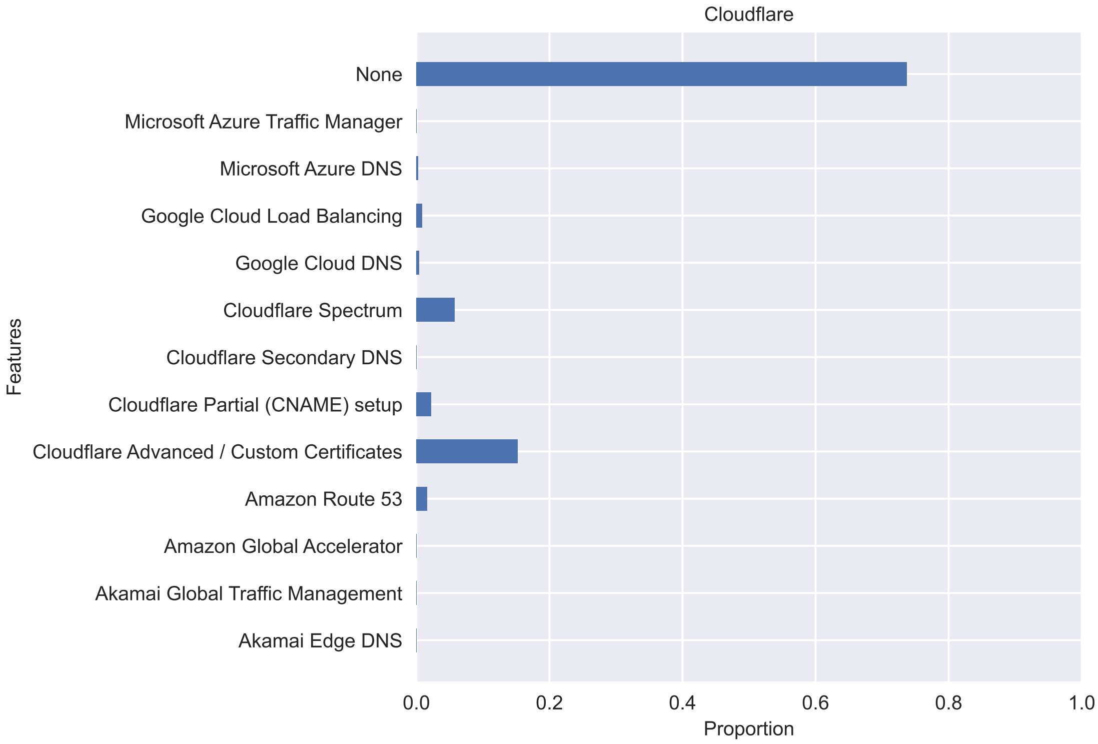
C.1 Akamai



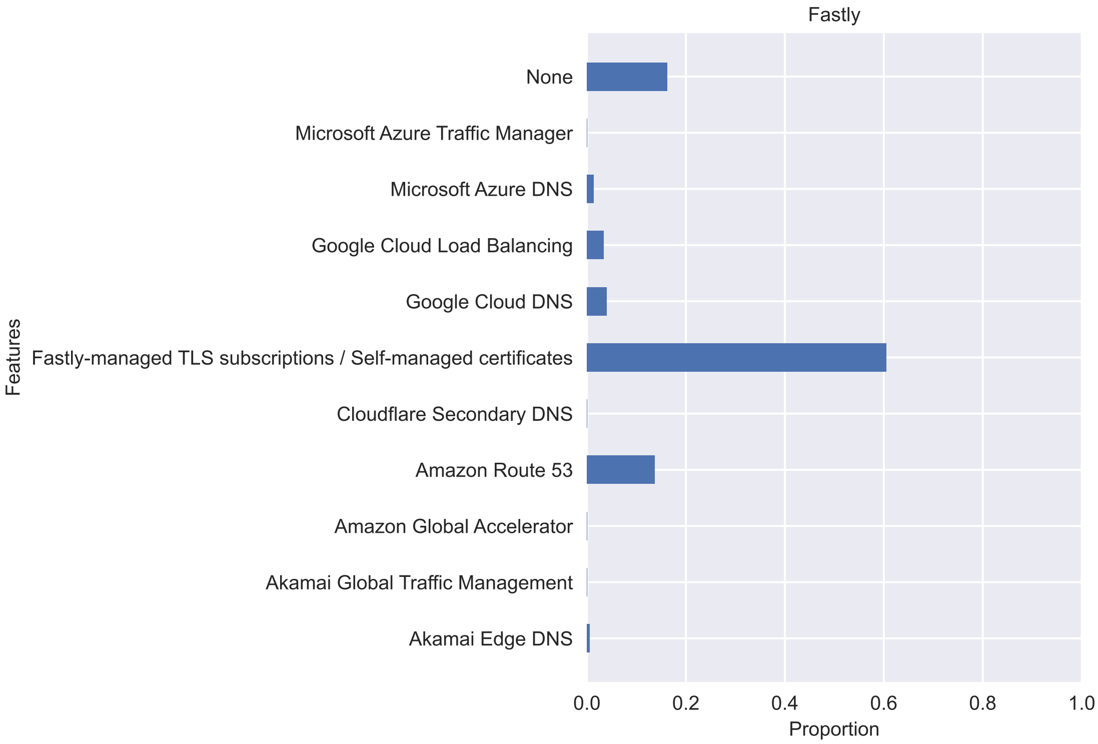
C.2 Amazon CloudFront



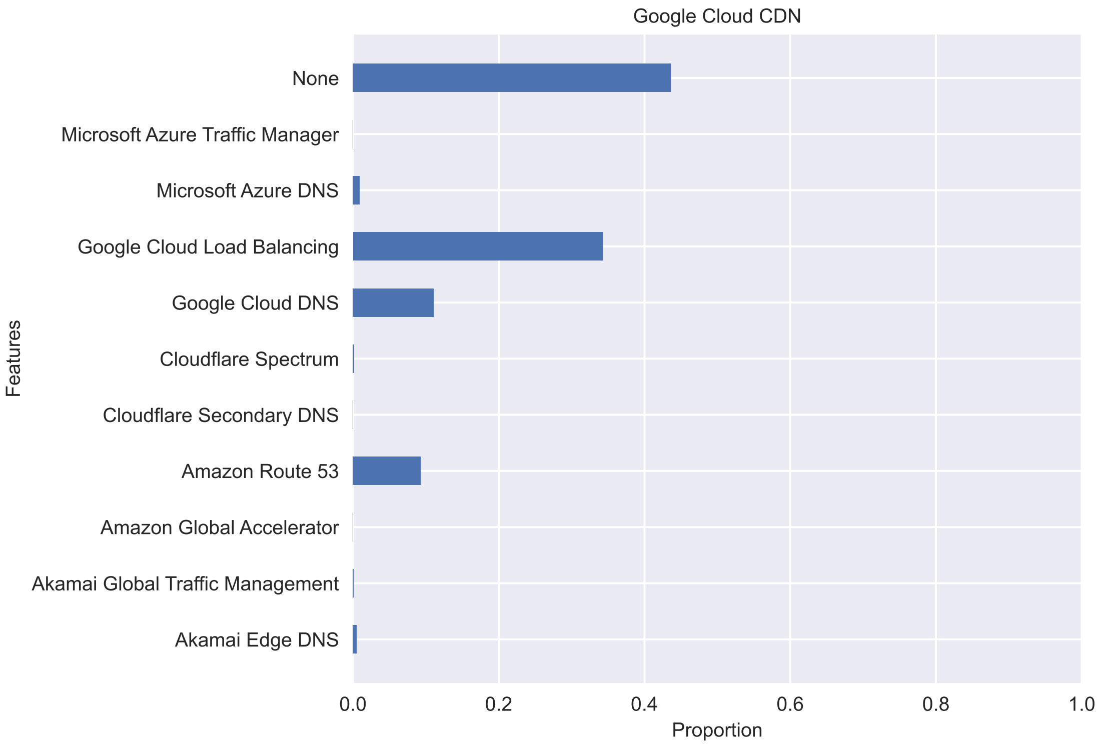
C.3 Cloudflare



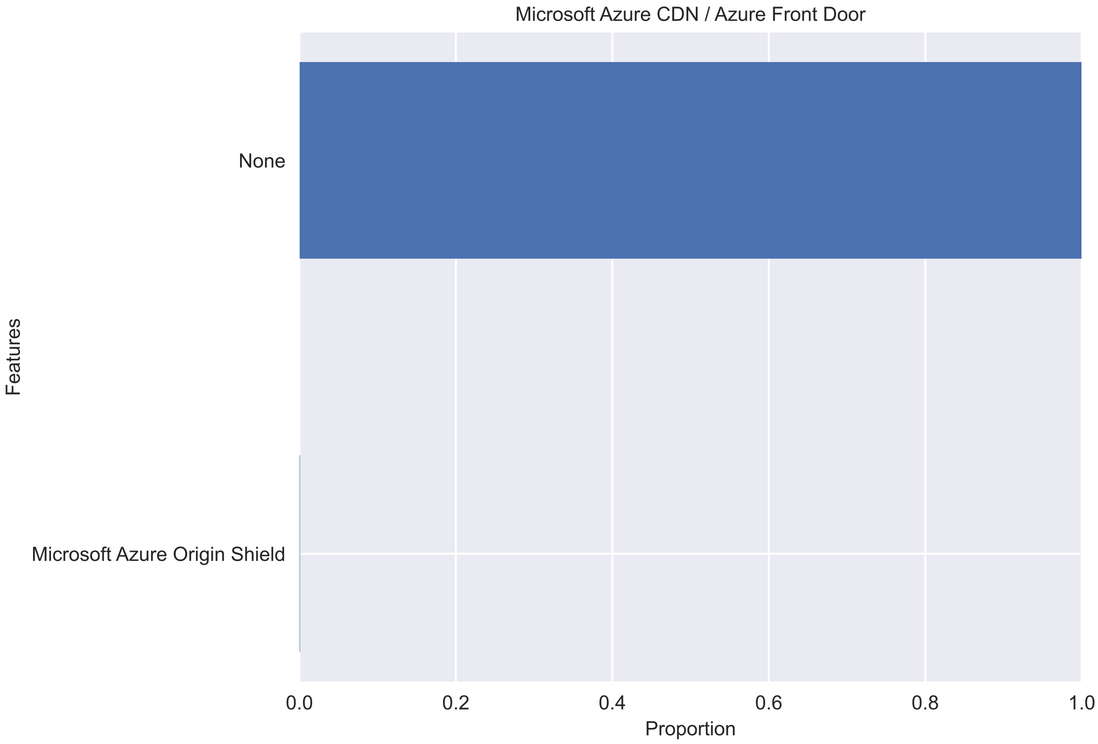
C.4 Fastly


Fastly
None
Microsoft Azure Traffic Manager
Microsoft Azure DNS
Google Cloud Load Balancing
Google Cloud DNS
Fastly-managed TLS subscriptions / Self-managed certificates
Cloudflare Secondary DNS
Amazon Route 53
Amazon Global Accelerator
Akamai Global Traffic Management
Akamai Edge DNS
Features
0.0
0.2
0.4
0.6
0.8
1.0
Proportion

C.5 Google Cloud CDN



C.6 Microsoft Azure CDN


Microsoft Azure CDN / Azure Front Door
None
Microsoft Azure Origin Shield
Features
Proportion