Connecting the Dots: Tracing Data Endpoints in IoT Devices

Md Jakaria
North Carolina State University
mjakari@ncsu.edu

Danny Yuxing Huang
New York University
dhuang@nyu.edu

Anupam Das
North Carolina State University
anupam.das@ncsu.edu

ABSTRACT

Smart home devices are constantly exchanging data with a variety of remote endpoints. This data encompasses diverse information, from device operation and status to sensitive user information like behavioral usage patterns. However, there is a lack of transparency regarding where such data goes and with whom it is potentially shared. This paper investigates the diverse endpoints that smart home Internet-of-Things (IoT) devices contact to better understand and reason about the IoT backend infrastructure, thereby providing insights into potential data privacy risks. We analyze data from 5,413 users and 25,123 IoT devices using the IoT Inspector, an open-source application allowing users to monitor traffic from smart home devices on their networks. First, we develop semi-automated techniques to map remote endpoints to organizations and their business types to shed light on their potential relationships with IoT end products. We discover that IoT devices contact more third or support-party domains than first-party domains. We also see that the distribution of contacted endpoints varies based on the user's location and across vendors manufacturing similar functional devices, where some devices are more exposed to third parties than others. Our analysis also reveals the major organizations providing backend support for IoT smart devices and provides insights into the temporal evolution of cross-border data-sharing practices.

1 INTRODUCTION

Internet of Things (IoT) devices have seen widespread adoption across the globe. The global IoT market is $662 billion today and is anticipated to increase to $3,353 billion in 2030 [36]. People worldwide use a wide range of IoT devices in their daily lives, including security cameras, voice assistants, smart home appliances, smart TVs, etc. These devices typically collect and share user data with different computing infrastructures for operational and analytic purposes [47, 61]. The data varies from users' personal information (e.g., email address, location) to users' activities (e.g., what a user is doing, whether the user is awake or not, what food the user prefers, what the user is watching). This data can readily be employed to create user fingerprints, raising security and privacy concerns for users (e.g., determining whether anyone resides at home or serving targeted advertisements). While for typical devices such as desktops and smartphones, there are controls to opt out of data sharing; such options are very limited for IoT devices other than installing firewalls on the gateway routers of the home network.

In recent years, researchers characterized IoT traffic in terms of the relationship between traffic generators and receivers. However, existing works collected and characterized traffic from a few dozen of devices in controlled lab settings [5, 13, 24, 26, 27, 49, 54, 56, 61, 67, 68, 80] that may not represent organic user traffic in the wild. Whether these existing techniques would still apply to a larger IoT traffic dataset is unclear. Work like Kumar et al. examined millions of IoT devices, although their dataset is proprietary, and the results are difficult to reproduce [39]. Crowdsourced datasets have also been used to characterize IoT traffic [30, 48]; however, such analyses have primarily focused on security aspects such as the use of TLS versions and the prevalence of unencrypted traffic. Little research has been done to identify the thousands of Internet endpoints that millions of IoT devices contact and their role in the overall IoT ecosystem. Previous works have categorized remote endpoints into three categories [47, 61]: first-party, support-party, and third-party. Such categorizations are based on whether a device manufacturer owns a remote endpoint, whether the endpoint provides CDN or cloud-based services, or none of the above. Mandalari et al. suggested that all third-party destinations are non-essential and can potentially be filtered/blocked [47].

Existing works have recognized different endpoint types, but their approach is manual and unscalable for a large number of IoT devices [47, 61]. To resolve this issue, we first develop techniques to augment the identification and categorization of smart home devices using a large language model (LLM). Next, we develop a semi-automated system to categorize endpoints into first, support, or third party with minimal human input. We collect device manufacturer and remote endpoint information from open data sources like WHOIS database, web scraping, SERP scraping, etc. We incorporate spaCy [75], an NLP technique to determine the organization names from unstructured streams of text. We also consider the parent-subsidiary relationship among various device vendors and manufacturers. With the information collected about the device vendors and the remote endpoints, we then map the relationship between IoT devices and remote endpoints. Using our approach, we conduct an analysis of the IoT Inspector [30] dataset with respect to the contacted endpoints. Our proposed methodology can also be applied to other independent datasets (as shown in §4.2). Such generalizability arises from the fact that the LLM model used for device identification is not fine-tuned, and the sources we used for endpoint classification are independent of the IoT Inspector dataset.

There is a lack of transparency regarding where sensitive user data goes and with whom it is potentially shared. Furthermore, little is known about how cross-border data-sharing practices in the smart-home IoT ecosystem vary over time. To address these gaps, we answer the following four research questions using our proposed endpoint detection system. **RQ1:** *How does the distribution of different types of endpoints vary across different categories of IoT devices?* We want to determine which devices contact more third parties than compatible devices. With such insights, consumers might be able to determine which devices are more dependent on

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–28

https://doi.org/XXXXXXX.XXXXXXX

support parties and more likely to share or sell user data to third parties. **RQ2:** *Do the contacted endpoints significantly change based on the user's location?* We want to see if the distribution of first, support, and third parties changes based on the user's location. More specifically, we want to determine which regions are more exposed to third parties. **RQ3:** *How frequently does user data cross borders?* We want to investigate cross-border data exchanges that take place. Such analysis can help shed light on potential violations of privacy regulations. **RQ4:** *Does the distribution of different types of contacted endpoints vary over time?* We want to investigate if the contacted first, support, and third parties for a given device change over time. We focus on smart home devices in our analysis, and throughout the paper, we use "IoT device" and "smart home device" interchangeably. In summary, we make the following contributions:

- First, we identify and categorize smart home devices. We examine 54,950 devices, exclude devices that are not typical IoT devices (such as PCs, phones, routes, etc.), and end up with 25,123 IoT devices across eight categories.
- We develop a semi-automated method to classify endpoints into first, support, or third party with minimal user input. Through manual validation, we show that our approach can achieve an average of 97% accuracy in identifying different endpoints.
- We apply our semi-automated endpoint categorization technique to traffic generated by 25,123 IoT devices used by real-world users worldwide and determine the distribution of first, support, and third parties for each device.
- Our analysis identifies the dominant back-end infrastructure for the IoT ecosystem, which can foster data transparency research for IoT devices, something existing literature needs to include.
- Lastly, we showcase how the distribution of contacted support and third parties varies based on the user's location and to what extent user data cross borders. We also perform the *first* temporal large-scale analysis of smart home IoT traffic.

Interested readers can find all the relevant code and anonymized data in the GitHub[1] repository.

2 RELATED WORK

Characterization of Real-world IoT Traffic. Recent works in IoT measurement have leveraged active and passive measurement data to characterize IoT device traffic. In particular, Kumar et al. leveraged a corporate dataset of 83 M devices from around the world and characterized their device properties as well as their security profiles [39]. This data was primarily collected through active probing of devices on the network. In contrast, Mazhar et al. instrumented the gateway software in 220 homes to collect passive data and investigate similar device properties [48]. Most recently, IoT Inspector [30] crowdsourced IoT network traffic from more than 5,500 real IoT users around the world to allow individual participants to identify potential security risks, such as the use of TLS versions and the prevalence of unencrypted traffic.

Fingerprinting IoT Traffic. In recent years, researchers have started analyzing the network traffic generated by IoT devices to uniquely identify IoT devices on the network. Researchers have shown that it is possible to infer not only IoT devices [6–9, 12, 21, 50, 63, 65, 66], but also certain device-level activities [1, 15, 55, 61, 72] from encrypted traffic. However, most of these works focus on building and evaluating models that work well on a relatively small dataset (typically less than 50 devices) and lack any analysis of how such models generalize to other datasets, often collected under different settings. Moreover, they lack any comprehensive open-world analysis — something that an adversary is bound to face in any real-world setting. Dilawer et al. have looked at the generalizability of such fingerprinting techniques across independent datasets collected across different settings [2] and have analyzed how device fingerprints evolve over time and generalize across different geolocations.

Security and Privacy Awareness for IoT Devices. Numerous previous works have demonstrated that the rapidly evolving IoT ecosystem is a significant source of privacy and security risks [4, 11, 71, 73]. Alasdair revealed that when plug-and-play IoT devices evolve, two groups of users emerge: aware and unaware. Many consumers blindly follow gadget prompts without reading terms [23]. Koohang et al. showed that IoT awareness could positively influence users' IoT privacy and security knowledge, as well as their trust and continued intention to use IoT devices [37]. Emami-Naeini et al. [18, 19] have showcased how security and privacy labels can influence users' purchase behavior for IoT products. Babun et al. demonstrated that IoT platforms are not generic, so most IoT users struggle to configure IoT devices correctly [10]. Also, IoT devices are vulnerable to information leakage, which exposes a user's private data, and most IoT platforms still lack effective privacy protections [10]. Saidi et al. analyzed backend infrastructures and revealed that 35% of IoT traffic is exchanged with IoT backend servers in other continents [62].

Distinction with Existing Work. We analyze traffic from IoT devices under real-world settings from smart homes worldwide, in contrast to lab studies such as Ren et al. [61]. Therefore, our work is closely related to the existing works listed in Table 1. Compared to Mazhar et al. [48], our work covers both a larger diversity of devices (1,103 distinct IoT products compared to 66) and geographic regions (globally compared to one US city). Furthermore, we cover a longer duration of time (3 years compared to 19 days). When comparing our work with Kumar et al. [39], we see that while Kumar et al. cover more devices globally, the data was collected only for one month, based on active scans, and did not include passive network traffic, which would be useful in understanding activities of devices. Moreover, their dataset is proprietary, as Avast collected it. Lastly, although we use a similar dataset as Huang et al. [30], we are different from the original paper in the following aspects: 1) longer duration (3-year period compared to 9 months), 2) more diverse devices (original paper analyzed only 25.8% of all devices in our current dataset), and 3) performs a more holistic analysis of contacted endpoints in terms of first, support, and third parties where existing work relies on blocklists primarily for the web. Moreover, none of the existing works have performed temporal analysis of IoT traffic. About the classification of endpoints, Varmarken et al. [74] and Razaghpanah et al. [60] have employed semi-automated methods as well. Their approach relied on proprietary databases (Crunchbase or D&B) for endpoint-to-organization

[1] https://github.com/jakariamd/IoT-Measurement.git

Table 1: Comparison with existing work on IoT traffic from *real-world users* and not from a lab setup.

Work	No. of IoT devices	Unique IoT products	Collection duration	Geographic regions	Temporal analysis	Analyzes passive network traffic	Data proprietary
Kumar et al. [39]	83 M	14K	1 month	global	✗	✗	✓
Mazhar et al. [48]	240	66	19 days	one US city	✗	✓	✗
Huang et al. [30]	6,776	81	9 months	global	✗	✓	✗ †
This paper	25,123	1,103	3 years	global	✓	✓	✗ †

† subject to IRB approval

mapping, which becomes problematic when organizational information is absent or the organization is newly established. Instead, our approach leverages multiple open-source data sources to map even more domains to organizations. Also, their datasets lack the diversity present in our dataset. Lastly, Saidi et al. [62] examined a limited set of backend providers over a short period (2 weeks). In contrast, our analysis identifies backend providers through organic communications spanning a longer duration (3 years).

3 DATASET

Background on IoT Inspector. IoT Inspector is an open-source tool that anyone in public can download on their computers to collect, analyze, and visualize the network traffic of their IoT devices. The tool also sends anonymous data to the IoT Inspector server. This data includes (i) information that suggests possible identities of devices, such as the Organizationally Unique Identifier or OUI (i.e., the first 3 bytes of a MAC address), the HTTP User Agent, mDNS [14] responses, and UPnP [51] announcements, along with any annotations that users voluntarily provide, including the names and manufacturers of devices; and (ii) statistics of network traffic, aggregated every five seconds, including the remote hostnames, remote IP addresses, remote ports, and the number of bytes sent and received. It collects statistics on the network traffic with ARP [77] spoofing, effectively man-in-the-middling all connections between the smart home devices and the gateway. IoT Inspector collects these data from the home networks of organic users.

Overview of Dataset. IoT Inspector was launched in April 2019 and has been in operation since. For this paper, we have requested a subset of the data from the IoT Inspector team for the time period between April 08, 2019, and July 20, 2022. The entire IoT Inspector dataset includes 216,671 devices across 11,787 global users.

4 DEVICE IDENTIFICATION

Device identification *(product name and manufacturer names)* is a crucial challenge for our dataset. Several potential metadata for identifying the devices are included in the dataset. Below, we detail the strengths and weaknesses of each metadata:

User Annotation. The users of IoT Inspector have voluntarily annotated 19,210 out of the 216,671 devices (8.87%) with product information (device and manufacturer names). User labels, while useful for device identification, can be inaccurate and unreliable, raising concerns about their authenticity.

OUI. The OUI represents the device manufacturers. For example, given an OUI `c8:3a:6b`, we can infer the manufacturer as `Roku`. One limitation with OUI is that many IoT manufacturers outsource the microprocessor unit/WiFi chips from third parties (e.g., Texas Instruments, AzureWave Technology, Espressif Systems). Devices with outsourced microprocessor units/WiFi chips broadcast their MAC addresses with third-party OUIs. Again, some manufacturers share the same OUI with others, and for those cases, `IEEE REGISTRATION AUTHORITY` is presented as the manufacturer name.

HTTP User Agent. HTTP User Agent can also be used to identify devices. For example, user agent `HbbTV/1.4.1 (+DRM+MEDIA360; Samsung; SmartTV2017; T-KTMAKUC-1262.0)` tells that the given device is a 'Samsung Smart TV'. Nonetheless, the shortcoming is that User Agent is not available for all devices.

mDNS & UPnP. The most substantial metadata for device identification is the outputs of Netdisco API, which resolves device identification from mDNS responses and UPnP announcements [38]. Several attributes of Netdisco result such as `name, device_type, upnp_device_type, properties, model_number, manufacturer` are used to confirm the device identity. For example, the following "netdisco" information for a device confirms that the device is a smart TV manufactured by `LG`.

```
{ name: 'LG webOS TV [REDACTED]', upnp_device_type:
'urn:schemas-upnp-org:device:Basic:1',
device_type: 'webos_tv', model_number: 'OLED65B6P-U',
model_name:'LG Smart TV',manufacturer:'LG Electronics'}
```

The problem with this source is that mDNS or UPnP information is very sparse and insufficient for identifying all devices.

FingerBank. To supplement the crowd-sourced annotations, IoT Inspector uses an external proprietary API, FingerBank [22], to infer the product names for every device. This proprietary API takes the OUI, User Agent, and remote hostnames as the input; it outputs the possible name of the product. For example, given the OUI `c8:3a:6b` of a device, FingerBank infers the product name as `Audio, Imaging or Video Equipment/Television/Roku TV`.

We employ two techniques to identify devices: manual identification and identification using Large Language Model (LLM). These techniques complement each other and help to identify a mutually exclusive subset of devices (detailed in §4.3). These techniques are explained in the following subsections.

4.1 Inferring Device Identities Manually

Since each metadata for a device mentioned above has a different level of strength and weakness, we use a combination of them and follow a vetting process to identify the device manually. Identifying a device means inferring the device vendor or manufacturer and the device type. For example, `Amazon-Fire-TV` is identified as `Device-vendor: Amazon`, `Device-type: TV`. The step-by-step device identification process is described below. For a given device:

(1) If HTTP User Agents or the Netdisco information contains the product name and manufacturer name, we record it as the ground truth label of the device. *OR*
(2) If the `model_number` is present in Netdisco information, we search on the Internet with the model number and record the

device identity from the search result (example: the first search result for `'OLED65B6P-U'` says the device is a "LG Smart TV").
OR

(3) If the product name and the manufacturer name from the user annotation are consistent with FingerBank and/or OUI, we also record those as the device identity.

One of our researchers manually inspected the device identification metadata of a subset of the dataset (54,950 devices) and identified 21,653 devices. We also inferred vendor names and device types of 13,954 other devices. We say "inferred" because we only have user annotation or Fingerbank labels, but we do not have any ground truth metadata for those devices. Thus, we have 35,607 (21,653+13,954) devices with manual device identification.

4.2 Inferring Device Identities with LLM

As explained in the original paper [30], IoT Inspector collects metadata of network traffic (e.g., source/destination IP addresses and ports), some payloads (e.g., DNS), and crowdsourced user labels. The dataset does not explicitly identify vendors, models, products of devices. This section discusses how we *infer* identities of devices, including the vendor and category information, based on this dataset. We say "infer" because we do not have the ground truth due to the crowdsourced nature; we can only validate our findings across different internal data sources and/or through manual inspection.

Overview. We obtained a subset of IoT network traffic from IoT Inspector's authors. For each device, we make sure that at least two pieces of the following metadata are available: OUI, DHCP hostname, mDNS/SSDP responses, hostnames contacted, and the user labels. The entire IoT Inspector dataset includes 216,671 devices, of which 25,033 have at least two pieces of the metadata. In the next few steps, we will infer the identities, including the vendor and categories, for these 25,033 devices.

Inferences with ChatGP.T Using OpenAI's Text Completion API [34], we develop prompts to infer device vendors and categories based on a device's DHCP hostname, mDNS/SSDP responses, and user labels. We use this API (base model, GPT 3.5 DaVinci, no fine-tuning) because it is trained on Internet-scale data, which likely includes public knowledge on various IoT devices. Also, we pick these three pieces of metadata because, based on our manual sampling, they are likely to contain identifying information (albeit imperfect), explicitly (i.e., substring) or implicitly. For example, user labels are crowdsourced and sometimes include incorrect spellings [30]; mDNS/SSDP responses often include the vendor and product information, although the exact formats could differ across vendors; and DHCP hostnames may be indicative of the product identity (e.g., the string "cast" often appears in the DHCP hostnames of Google Chromecast devices). We treat all these metadata as unstructured natural languages — especially given the diversity of IoT devices — and feed them into the Text Completion API. We develop the following prompts by iteratively testing different prompts on a small subset of known devices: (i) To infer the vendor names, we ask: "I have an IoT device named '`[metadata]`'. What is the company that makes this IoT device? Output the company's name only." (ii) For device type, we use this prompt instead: "I have an IoT device named '`[metadata]`'. What type of IoT device is this? Output the name of the device type only." We replace `[metadata]` with user labels, DHCP hostnames, or mDNS/SSDP responses, separately, extracted from the IoT Inspector dataset.

We apply these prompts to the 25,033 devices with the Text Completion API. After removing empty or unknown responses, we have the API responses for 24,998 devices. At the time of writing, the API cost was approximately $70 USD in total.

Validating Vendor Inferences To evaluate the API's vendor inferences without ground truth knowledge, we check the *consistency* across the API's outputs based on different metadata inputs. Two outputs are considered consistent if they share a common substring or word that has at least length 3 and is not a stop word (e.g., "the" or "smart"), case insensitive and ignoring punctuations. For each device, we have at least two (out of three) independent metadata: user labels, DHCP hostnames, and mDNS/SSDP responses. We feed the prompt into the API separately for every metadata. If the outputs for at least two pieces of metadata are consistent, then we assume that the API output is correct. Furthermore, it is also possible that we cannot find consistent API outputs using the method above. To supplement the above, we also check the Text Completion API output against the IEEE OUI vendor names (based on the MAC addresses) and the domain names to see if the Text Completion inferred vendor is consistent with the OUI and/or domain names.

To illustrate our method, let us examine one actual device from the dataset. This device has the DHCP hostname `Google-Home` and the following mDNS/SSDP information:

```
{"host": "[REDACTED]", "hostname": "[REDACTED].local.",
 "port" : 8009, "device_type" : "google_cast",
 "properties": {"md": "Google Home", ... }}
```

From both the DHCP hostname and mDNS/SSDP information, the Text Completion API consistently returns "Google"; we thus label this device's vendor as "Google."

There are cases where the vendor's name is not explicitly in the user label, DHCP hostname, or mDNS/SSDP response. For example, one of the devices in the dataset has `Onelink` as the DHCP hostname; the Text Completion API, given this information and our prompt, outputs "First Alert" as the vendor. The device also includes the string "Onelink Safe Sound E380" in the mDNS/SSDP response; the Text Completion API also outputs "First Alert" as the vendor. Because in both cases the API returns consistent responses, we infer the vendor as "First Alert" (even though "First Alert" never appears anywhere in our dataset for this device).

We conduct this consistency check across all possible combinations of user labels, DHCP hostnames, and mDNS/SSDP information for all the 24,998 devices. We find consistent vendor information across 19,096 (76.4%) of these devices. We then sampled 50 random devices from the 19,096 devices to manually check if the inferred vendors and categories are consistent with the device metadata, referring to Google search results when necessary. In all 50 cases, we find that the Text Completion output vendors and categories are consistent with the device metadata.

Validation with Auxiliary Datasets. We validate our proposed method of identifying devices using LLM with 2 auxiliary datasets: the UNSW IoT Analytics [66] and the YourThings IoTFinder [58] datasets. We extracted the same device metadata from the PCAP files of these two datasets: OUI, DHCP hostname, mDNS/SSDP

responses, and the hostnames contacted. The sole difference here is that in the IoT-Inspector dataset, we have user annotations for many devices (19k), which we use in device identification using LLM. We refrain from using user labels this time since we use them as ground truths.

A device can be identified if all the following three conditions are met. For a given device, (a) at least 2 out of 5 metadata mentioned above are present; (b) metadata contains identifying information explicitly (i.e., substring) or implicitly (model number or alternative name). Metadata may contain generic information (such as "Mozilla/5.0" as user_agent), which is not useful in identifying a given device. (c) Prompt response for at least two metadata matches. The outcome of our analysis of these two auxiliary datasets is as follows: (i) For the UNSW dataset, among 30 devices, 18 were correctly identified, and 12 were not identified. *"Not identified"* does not necessarily mean incorrectly identified; instead, the above-mentioned device identification conditions are unmet. Table 8 in Appendix A.2 shows the presence of metadata for the 12 unidentified devices. We see that most unidentified devices have less than two metadata, and/or the metadata does not contain identifying information. We experienced a similar aspect for the IoT-Inspect dataset, i.e., our method could not identify many devices because of missing metadata. (ii) For the IoTFInder dataset, among 66 devices, 35 were correctly identified, and 29 devices were not identified (missing metadata for 14 devices ~ see Table 8 in Appendix A.2; and for the remaining 15 devices, no network packet was seen in PCAPs), and 2 devices were incorrectly identified. Regarding the 2 incorrectly identified devices, we argue that our method correctly identified the devices, but there were some mislabeled ground truths (an Apple device labeled as Android Tablet, and a Samsung device labeled as iPhone). We confirm this information by manually inspecting the OUI and user agents. A summary is shown in Table 2.

Our proposed LLM-based device identification method is independent of datasets as it was able to discover different unique devices across the datasets. For example, this method correctly identified a Belkin Crockpot device from the YoutThings dataset, which we did not see in the IoT-Inspect dataset.

4.3 Device Selection and Categorization

We infer 35,607 devices manually and 19,096 with LLM separately, and these two device sets are non-disjoint. We check all manually identified devices against LLM-inferred devices. The resilience of LLM inference over manual vetting is that the former process also considers the hostnames contacted. We retrieve three sets of devices by combining manually vetted and ChatGPT-inferred devices. The number of devices in each set is shown in Table 2. The total number of unique devices we identified in the dataset is **29,212** which is around 47% of devices for which we have the network traffic statistics. We see 216 cases where LLM produces different results (e.g., TCL/Insignia TV labeled as Roku TV). These labels are not entirely incorrect as many devices are being built on firmware provided by other companies. However, manual identification provides more robust labeling in such cases. For the analysis of the IoT Inspector dataset, we combine both approaches to obtain higher device coverage. We present a sample list describing mismatch between LLM and manual device identification in Table 7 of Appendix A.1.

Table 2: Device Identification Statistics

Dataset	Description	# Device
IoT-Inspector	Identified by ChatGPT & manually	13,007
	Identified manually, not by ChatGPT	10,683
	Inferred by ChatGPT, not manually	5,522
	Total Identified	**29,212**
	Non IoT	4,089
	IoT	**25,123**
UNSW IoT Analytics	Identified by ChatGPT	18
	Not Identified by ChatGPT	12
	Total	**30**
YourThings IoTFinder	Identified by ChatGPT	37
	Not Identified by ChatGPT	29
	Total	**66**

Table 3: Device Categorization Statistics

Device Category	# Device
Media/TV	10,695
Home Automation	7,396
Voice Assistant	4,076
Surveillance	1,376
Game Console	720
Work Appliance	637
Home Appliance	162
Generic IoT	61
ToTal	**25,123**

The IoT inspector dataset consists of data of various smart home devices, as well as other devices such as PCs, phones, WiFi extenders/boosters, etc. To understand device characteristics better, we categorize the device types into 15 categories used in previous work [39] showed in Table 5 of Appendix A. A sample device type to category mapping is shown in the Table 6 of Appendix A. For further analysis, we only consider IoT devices (shown in the lower portion of the Table 5) and exclude the non-IoT devices. The number of IoT and non-IoT devices are shown in Table 2. For the rest of the analysis, we will focus on the **25,123** IoT devices. The number of devices in each category is shown in Table 3. Among these **25,123** IoT devices, we find 1,103 (listed in Table 1) unique IoT products. We define an *IoT product* as `device_vendor:device_type`.

5 ENDPOINT CATEGORIZATION

This section will explain our mapping process of contacted endpoints into first, support, and third parties. Our process is semi-automated, requiring minimal human intervention.

5.1 Finding First Parties

Given a device vendor and contacted hostname, we resolve the vendor organization names and server organization names separately. Since a device can be manufactured by a subsidiary organization and then re-branded as parent brand or vice-versa, we also collect the parent brand, all the sub-brands, and subsidiaries of the given brand through Google Search. Therefore, for a given brand, we come up with a list of organizations that represent this brand. We name this as "`device-org`" list. After constructing the "`device-org`" list, we find the organization names for the hostname the given device communicates. We first find the effective domain name (eTLD) from

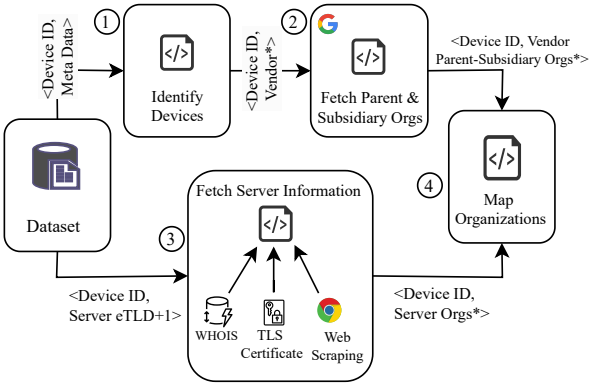


Figure 1: Step-by-step process of first-party detection. ① Identify device vendors and their organization name; ② Fetch parent and subsidiary organizations from vendor organizations in publicly available sources; ③ Leverage a script to automatically mine domain hosting organization information from several sources including WHOIS, TLS certificate, domain based web-scarping, etc.; ④ Map device vendor organization and the remote server organization. If they match, we say the remote server is the first party for the given device. * represents the output from the previous step.

the hostname. Then we collect the organization names that possess this eTLD. We collect organizational information from several complementary sources shown in Figure 1. These sources include:

❶ **WHOIS Databases.** WHOIS [78] is a protocol to retrieve information from databases containing data about registered domain names, IP addresses, autonomous systems, etc. We make WHOIS queries to look for registrant information. The registrant field of a WHOIS response typically represents the organization owning a given domain. The registrant field in the WHOIS response can be missing or be privacy-protected. Many registrars offer privacy or proxy services that hide the registrant's information. We generate a list of privacy masks from the WHOIS query results. Our approach for collecting domain owner names from the WHOIS database is as follows: a) we make a WHOIS query with a domain name and retrieve the registrant organization name from the query result, and b) if the registrant's name is protected with the privacy mask, we filter out that registrant's name.

❷ **TLS Certificate.** Another interesting source of information about the organization of a server is the TLS certificate. The root of a certificate chain contains the organization name. The caveat is that many servers lack HTTPS service, resulting in the absence of a TLS certificate for organization name lookup. Additionally, even if a remote server provides HTTPS service, the organization name field might be absent in the leaf certificate. Our experimental results show that we found organization names in only one-third of server URLs. Using a TLS certificate for organizational information is promising because the results are trustworthy and devoid of privacy masks, unlike WHOIS queries.

❸ **Web Scraping.** We complement the company information data collection through web scraping. First, we crawl the landing web page and see if there is copyright information within that page. We take the organization names from the copyright information of the landing page. We also extend our quest for organizational information in the contact information, privacy policy, and terms and services pages. We incorporate Named Entity Recognition (NER) [41], a cutting-edge NLP technique to extract organization names from unstructured text streams. Specifically, we use *spaCy* [75] language model, which has an F-1 score of 0.86.

We also utilize the Netify[2] network lookup tool to find the domain-level information. This information can supplement the other resources mentioned above. We say supplement because Netify only has information for some of the domains in the dataset. We also noticed that some domains are exclusive to a specific vendor, which means that devices of a particular vendor exchange traffic with the given remote endpoint. We call those domains to be exclusive domains. To identify the exclusive domain, we examine whether over three devices from a specific vendor communicate with a domain, and no devices from other vendors communicate with the same domain. We then check if the exclusive domains are the first party of the particular vendors. Since only several hundred of them exist, we just manually check them. We found some first-party relationships that are not possible via other approaches. One example is `xbcs.net`, the first-party endpoint for Belkin devices.

After obtaining organization names for a device vendor and the related hostname, we map these two sets of names. Direct string matching does not work as organization names for a given IoT vendor and the organization names for a server may not match exactly, although they represent the same company. For example, a device made by "Samsung Group Inc." communicates to the following domain "samsungcloud.com" which is registered by "Samsung Electronics Co., Ltd.". These names do not match directly but their base names represent the same company. We follow a simple but effective process for organization name matching. We first replace the non-ASCII characters, remove punctuation and common legal business suffixes, and finally remove the common suffix in the organization names. In this process, the full organization names "Samsung Electronics Co., Ltd." and "Samsung Group Inc." are reduced to "Samsung". Even after getting the base names, in some cases, the organization names do not match perfectly. To increase the probability of matching, we leverage some fuzzy string matching algorithms, including Discounted Levenshtein [43], String Subsequence Kernel Similarity [45], and Token Set Ratio [64]. If a device manufacturer's name and the domain's organization name match at least a certain threshold (set to 90% empirically), then we say that the connected domain is a first party for the device.

To evaluate the performance of this approach, we sample 100 first-party domains and manually vet the relationships. We check whois.com[3], netify.ai[4], and the target domain's web server, to determine the owner of the domain. We find that 96 first-party relations are correctly predicted. The remaining four domains are third-party; however, the relationships are incorrectly determined since the remote endpoints explicitly mention the corresponding vendor name on their web pages. For example, Stan [46] is a third-party streaming media but is classified as a first-party endpoint for Samsung since the scrapper found "Samsung" on the landing page of Stan.

[2]Example: https://www.netify.ai/resources/domains/nest.com
[3]https://www.whois.com/whois/<domain>
[4]https://www.netify.ai/resources/domains/<domain>

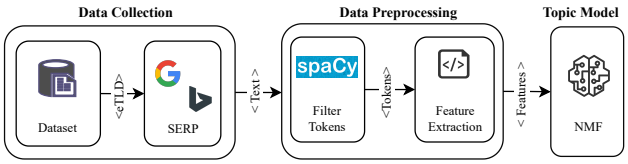


Figure 2: Step-by-step process of support parties detection. Data Collection collects textual information about a server from SERP. Data Pre-processing filters unnecessary words from the text and produces tokens. Then the stream of tokens is fed into the TF-IDF feature extraction module to produce useful features. Topic Model use NMF to find the topic clusters. From these clusters, we extract only those clusters that represent support parties.

5.2 Finding Support Parties

We define *support parties* as non-first-party providers that offer a computing platform as a service. This category includes IoT backend providers like Tuya, SmartThings, HomeKit, etc.; OS vendors like Roku[5] and Fire TV; cloud platforms like Amazon Web Services (AWS), Google Cloud Platform, and IBM Cloud Computing; as well as content delivery network (CDN) services like Akamai, Fastly, Egdecast, CDN77, etc. Knowing the services a remote server offers is necessary to determine support parties. Some tools like WebXray [28] provide this information; however, WebXray targets web servers and only includes a few of the support party information for IoT devices. Therefore, we adopt a different strategy. Search engines are alternative sources for this type of information. We search Google and Bing with the query "`What is + $domain`", which returns a list of results related to the specified domain. We then scrape the Search Engine Results Page (SERP) for organic search results. We only focus on the text under each result, as it best reflects the type of service a remote domain offers. At this point, we require an automated tool to comprehend the outcome and identify the type of service being described in this text. An apparent choice is employing Natural Language Processing (NLP). We utilize a topic model to determine the topics the SERP displays for a given query (for a given domain). Such topics include advertising, tracking, CDN, and hosting information. We use those topics to identify the type of service a domain offers and then determine the relationship between a given domain and a smart home device. Figure 2 depicts the overall process for identifying support parties.

Dataset. We construct a text dataset utilizing Search Engine Result Pages (SERP). For each domain, we pull a SERP from Google and one from BING. We notice variations in organic search results for the same query on different search engines; therefore, we utilize both. We aggregate all the texts from all the search results. This creates a stream of text for each of the domains. We retrieve data for over 5,090 IoT-specific domains.

Data Pre-processing. At this stage, we clean the text stream from the SERP. We remove irrelevant words and retain only the essential keywords or tokens. Our target tokens are primarily nouns, pronouns, and adjectives. To retain target tokens, we utilize *spaCy* [75], a popular lemmatization tool used to identify the parts of speech (POS) of each word in a given text stream.

[5]We consider Roku a support party because it provides the operating system and network infrastructure for many non-Roku-branded TVs, such as TCL, Hisense, and Philips. See https://www.roku.com/products/roku-tv

Feature Extraction. The topic model cannot accept the token stream directly, as most NLP models prefer numeric input over strings. Therefore, we need a method for converting text input to meaningful numerical features. This conversion can be done in various ways, including word counting, term frequency — inverse document frequency (TF-IDF), binary encoding, etc. In all the above methods, a vocabulary is first made by looking at each unique token in the entire dataset (i.e., corpus). If the token in the given token stream exists in the vocabulary, the corresponding vector element in binary encoding is set to 1. In the counting approach, in addition to determining whether a word exists, it examines the frequency with which it appears. This method assigns weights to words based on frequency, and words that occur more frequently will have greater weights. Contrarily, TF-IDF assigns greater weight to infrequent words than frequent words [3]. It is based on the assumption that less common words are more significant. Counting and TF-IDF represent two extremes; we need something in between for our dataset. In our feature extraction phase, we employ both counting and TF-IDF and feed to the topic model separately.

Topic Modeling. We employ the topic model to determine the topic of a domain's search result. A topic is simply a group of words that describe the overall theme. BERT [16] and Non-Negative Matrix Factorization (NMF) [42] are popular topic modeling methods. NMF has been widely used as a clustering method, particularly for document data, because it produces semantically meaningful results easily interpretable in clustering applications. We build two separate NMF topic models with different feature extraction techniques: Counting and TF-IDF. We set the number of components in each topic model, `n_components`, to 30 empirically. A lower value of `n_components` combines two or more different types of clusters into one, while a higher `n_components` splits a cluster into many. We set `n_components` in a way so that the model starts splitting the same category cluster into multiple groups. As NMF is an unsupervised technique yielding unlabeled clusters, manual effort is required to label the clusters. However, this is a one-time offline process, and once the labels are determined, this model can be used to infer the topic for future input. After obtaining labels for topic clusters, we identify clusters that align with the definition of support parties mentioned earlier. We extract two sets of support parties from these two topic models and take the intersection of these two sets. The rationale for this is that overlapping domains more accurately represent support parties. A complete list of cluster labels is shown in Table 10 and 11 of Appendix B.1.

We manually vet 100 support party domains determined by this framework and find that 97 are `True` support parties. The remaining 3 domains are actually third parties misclassified as support parties. One such example is YaleHome <`yalehomesystem.co.uk`>, a home security system company. The SERP for this domain includes many IoT product keywords (such as '`smart door locks`', '`smart home alarms`', '`CCTV systems`'), leading to misclassification. We discuss the limitations of this approach in the discussion section.

5.3 Finding Third Parties

Contacted domains that are neither classified as first nor support parties are categorized as third parties. Many high-end IoT devices like smart TVs come with internet browsers, and users may browse

domains that are not in the first-party or support-party list. Additionally, third parties may represent any third-party app or skill for high-end gadgets such as voice assistance or smart TV. The most interesting third parties are advertising and tracking companies. We randomly choose 100 third-party domains and manually vet the relationship with the corresponding vendors. 98 of them were categorized correctly. The remaining two domains were support parties misclassified as third parties. These misclassifications are due to the limitations of the tools used to identify support parties. The limitations are described in the discussion section.

5.4 Design Choices

LLM vs SERP. One alternative approach to classifying remote endpoints into first, support, and third parties is employing LLM, such as ChatGPT. We prefer using SERP over LLM for endpoint categorization due to SERP's superior performance. To compare the performance of the LLM approach with the SERP approach, we randomly sample 100 vendor-endpoint pairs (33 ~34 samples per endpoint type) and use OpenAI-GPT-3.5 API to classify the endpoints. We find that LLM correctly categorizes only 57 out of 100 samples. In contrast, we found that SERP can categorize 97 domains out of 100, shown in §5.2. For example, OpenAI-GPT-3.5 categorizes sbixby.com as a support party for a given Samsung device (with explanation: *The domain sbixby.com does not match the company name Samsung. It appears to be related to Samsung's virtual assistant Bixby, indicating a support party relationship*), whereas it should have been the first party (by looking into Whois). Contrarily, regarding device identification, SERP typically fails to identify devices, whereas LLM understands context and offers better results. For example, DHCP hostnames can be used to identify devices. Google searches with full DHCP hostnames typically yielded useless or blog-based network traffic analysis websites. However, OpenAI-GPT-3.5 often tokenizes DHCP hostnames and infers vendor names correctly.

Manual Efforts Required for Incorporating New Data. We claim that our endpoint categorization is automated if we have device identity (vendor name/type). However, we manually identified some devices because the dataset is overly sparse. When a device has more than two metadata entries, language models such as ChatGPT can consistently recognize the device in most cases. Manual device identification will no longer be needed if users voluntarily provide device information in their home network or use IoT Inspector for sufficient time to let it collect necessary meta-information. Again, the endpoint categorization method can be applied at the user end of IoT Inspector once the device information is identified. About constructing the "device-org" list (§5.1), we can create the list scalably using SERP. We utilize a script for scraping parent and subsidiary company names based on a vendor name. As for cluster labeling (§5.2), it's a one-time process, and we can reuse the cluster model for new endpoints.

6 RESULTS ANALYSIS

Using the recent IoT inspector dataset, we analyze the network traffic statistics of IoT devices shown in Table 3 to answer the research question mentioned in section 1. The following subsection highlights the analysis results of each of the research questions.

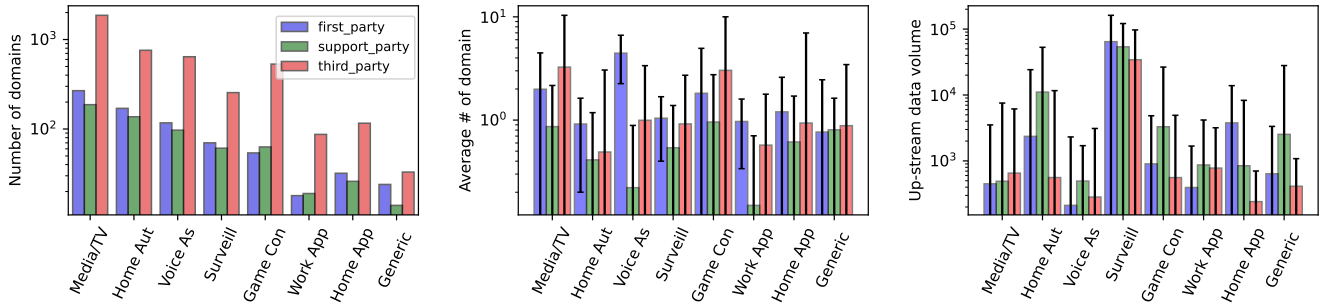
6.1 Distribution of Endpoint Type

In this section, we address the research question **RQ1:** *How does the distribution of different types of contacted endpoints vary across various categories of IoT devices?* To shed light on the backend infrastructure for IoT products, we analyze the different domains IoT devices communicate with and group them into first, support, and third-party endpoints. This analysis will help us understand how information could flow between IoT devices and various endpoints, and provide an upper bound on *potential* privacy risks. This is an upper bound because IoT Inspector does not capture the traffic payload, and we have no knowledge of whether/which sensitive data is shared between IoT devices and remote endpoints. Figure 3 highlights the distribution of first, support, and third-party endpoints for IoT devices of various categories. We present the normalized numbers across device categories in Table 17 in Appendix F.1.

In Figure 3, we see that the total and the average number of third-party domains communicated by devices in the `Media/TV category` is higher compared to other categories. This is reasonable, considering that smart TV users access various streaming services. However, this number also indicates that TV devices are the most susceptible to exposure to third parties. Smart voice assistant devices rank top in terms of the average number of first-party domains contacted by each device. Popular voice assistant devices (e.g., Amazon Alexa and Google Assistant) include third-party applications, often requiring them to communicate with third-party remote endpoints. Among those third-party endpoints, we notice 139 advertising domains (based on Disconnect list [31], Easylist [52], and DuckDuckGo tracker radar [17]). Wireless speakers are becoming popular, and many also support voice assistant services. This type of device also communicates with third-party domains, including tracking-based analytic services like `google-analytics.com`.

Figure 3c reveals that Surveillance devices, such as IP Cameras, smart doorbells, and indoor cameras, rank top in terms of upstream data volume per second. This is reasonable as these camera devices transmit video data to servers. However, the concerning aspect is that surveillance devices also transmit a considerable amount of data to third parties, potentially including private video records [20]. We observe that `Surveillance` devices communicate with third parties, including advertising and analytic service parties like `doubleclick.net`, `pubmatic.com`, etc. Devices across various categories (namely Home Automation, Voice Assistant, Game console, Work Appliance, and Generic) tend to send more data to support parties than the first party. This suggests that devices in these categories rely more on support parties for functionalities than on the first parties, potentially making support parties more significant in terms of data harvesting. Notably, Media/TV devices stand out for uploading more data to third parties than to either first or support parties. Most smart TVs can collect audio (via voice activation features) and Usage Data, which involves monitoring how users interact with the TV for advertising and marketing purposes. There is a suspicion that this information is shared with third parties, posing a significant threat to individual privacy [74].

Furthermore, we see a lot of third-party domains being contacted by `Game Console` devices. As many as 158 third-party advertising /tracking domains are particularly interested in gaming devices. These include advertisers like `doubleclick.net`, `adsrvr.org`,



(a) Total number of domains (b) Avg. number of domains (c) Average upstream data volume

Figure 3: The distribution of contacted endpoints and upstream data volume. (a) shows the number of domains accessed by all the devices in each category. (b) presents the average and standard deviation of the number of domains contacted by each device. (c) presents the average and standard deviation (shown as whiskers) of the volume of upstream data a device sends in a second (byte/sec).

`googleadservices.com`, `ad-delivery.net`, `advertising.com`, `amazon-adsystem.com`, etc. Interestingly, many simple IoT devices (smart plugs, bulbs, etc.) also communicate with third-party services. These services include advertising services like `segment.io`, `newrelic.com`, `adzerk.net`, `hotjar.com`, etc.

Given that the dataset is imbalanced regarding the number of devices across categories, it can potentially introduce biases in the outcome shown in Figure 3 (e.g., more domains for media TV, home automation, voice assistance). We, therefore, present the average number of domains a device communicates with and the average volume of upstream data to alleviate that issue. To demonstrate the generalizability of our analysis, we also randomly sample an equal number of devices from each category and conduct a comparable analysis on this balanced dataset. We observe nearly identical distributions for both the balanced set and the entire set of devices. These results are shown in Figure 10 and Figure 11 in the Appendix F.1. Also, some device categories exhibit a high standard deviation in the number of domains contacted and upstream data volume. This is due to the voluntary nature of participation of organic users.

We also see that the distribution of contacted endpoints varies across devices that belong to the same category but are produced by different vendors. For example, Google Chromecast, Samsung, Sony, and Apple TV devices communicate with multiple first-party domains, whereas TCL, Hisense, Onn, and Sharp TV devices communicate with no first-party domains. Instead, they rely on support and third-party domains to function. Similarly, the Samsung camera sends data to support and third parties, whereas SimpliSafe camera only shares data with the first-party domains.

Common Infrastructure Across Vendors. We would like to understand common infrastructure across IoT vendors such that the infrastructure does not appear to be vendor specific. This would allow us to understand how this common infrastructure may have amassed data from various IoT vendors and also potentially impose software supply chain risks. The software supply chain is likely at play here because if devices from two different vendors communicate with the same infrastructure, these devices potentially share the same (third-party) libraries to achieve similar functionalities, although we cannot verify as we lack access to the raw firmware.

We use a subset of IoT Inspector's hostname dataset, which shows the remote hostnames (FQDNs) that each device communicated with. According to the IoT Inspector paper [30], these hostnames were extracted from DNS queries, SNI fields within ClientHellos, and Host fields within HTTP requests; they were also inferred based on reverse DNS (i.e., PTR records) and passive DNS (i.e., based on FarSight's data) records. In this way, we obtain a list of hostnames and domains, along with the device IDs and the associated user IDs.

We restrict our analysis to devices labeled and validated per the method described in Section 3. We further restrict our analysis to remote domains contacted by at least *three* distinct users (based on the user IDs) to reduce the probability of mislabeled devices and of looking at short-lived or temporary connections. To map out the common infrastructure across vendors, we also restrict to domains that were contacted to by devices from at least two vendors. Finally, we remove domains that are exclusively contacted by media-related devices, such as TVs and speakers. We remove these cases because the domains are likely a result of the users' interaction with third-party content rather than built into the devices themselves. All the above restrictions further narrow us down to 19,470 devices from 159 products across 129 vendors. We manually assign a label to each domain. These labels are mutually exclusive, including the following:

- Company-specific, such as Heroku, Facebook, Apple, Microsoft, Google, and Amazon. These correspond to generic hostnames hosted by the said companies' infrastructure, which could be first-party (Apple's iCloud) or third-party (AWS).
- Functionality-specific, including (i) *analytics*, for logging and performance measurement; (ii) *IoT*, for infrastructure related to controlling IoT devices; and (iii) *time servers*, for infrastructure for devices to ask for the time (e.g., through the Network Time Protocol).
- Network: Domains related to DNS and CDNs.
- Others: All other cases.

After applying the above method, we show the result of our analysis in Table 4. We highlight a few cases of shared infrastructure and likely shared functionalities across vendors. First, we examine domains with the "IoT" label.

Table 4: Infrastructure shared across IoT devices of various vendors

Label	Vendor count	Product count	Device count	User count	Hostname count	Domain Count
Amazon	79	97	8705	3052	32928	16
Google	72	92	9187	3486	9317	17
Others	71	88	9213	3580	293253	351
Time Server	64	72	3411	1447	211	3
Network	46	56	4261	2103	111444	18
IoT	21	28	2854	1855	1957	13
Microsoft	16	20	302	253	1175	7
Apple	15	19	735	541	1738	3
Analytics	12	13	303	236	17	2
Facebook	12	17	445	362	85	2
Heroku	5	5	30	28	88	1

- `meethue.com` is shared across 1,143 Philips hubs, 241 Amazon voice assistants, 15 Amazon TV. Apparently, these devices all support capabilities for interacting with Philips Hue lights [29].
- `pubnub.com` is shared across 351 Logitech hubs and 21 Wink hubs. This service allows the control of IoT devices through various messaging services, such as MQTT [53]. Similarly, `pndsn.com` is shared across the above devices as well; this is a Data Stream Network for PubNub [35].
- `dropcam.com` is shared across 169 Nest cameras and 14 Google Home voice assistants. Interestingly, DropCam was acquired by Google; this evidence potentially shows some of Google's Nest Cameras still use DropCam's infrastructure [40].
- `netgear.com` is shared across 20 Netgear cameras, 9 Netgear hubs, and 3 Arlo cameras. Arlo was a spin-off company of NetGear [33]. Similarly, `arlo.com` is shared across 23 Netgear hubs, 6 Arlo cameras, and 19 Netgear cameras, while `arloxcld.com` (Arlo Cloud) is shared across 31 Netgear hubs, 32 Netgear cameras, and 6 Arlo cameras.

For domains labeled as "Analytics", we see `domotz.com`, a network monitoring service, being shared across 216 Roku TVs, 21 TCL TVs, 8 Amcrest cameras, 5 Insignia TVs, and 3 Sharp TVs. Additionally, `mixpanel.com`, another analytics company, is shared across 13 Sony TVs, 7 Wink hubs, 8 Amazon TVs, 5 Nvidia TVs, and 3 Vizio TVs. A comprehensive list of common infrastructures shared across IoT vendors is presented in Table 12 in Appendix C.

Summary. Our analysis shows that streaming devices like Smart TVs, AVR/DRV, STBs, and game console devices are more exposed to third-party domains as the average number of domains connected by devices in these categories are higher, and as much as 23% of the third-party domains are advertiser or tracker. Users need to pay close attention while using such devices, as information could potentially be flowing from these devices to the said third parties. In addition, low-end smart home devices like plugs, switches, and IP cameras are also exposed to third-party domains. Regarding support parties, we see many devices sharing similar back-end infrastructure. Knowing the distribution of back-end can help gauge the impact of service disruption or even denial-of-service attacks. It also paves the way for future work looking into potentially shared software supply chains for some of these devices.

6.2 Impact of User Location

In this section, address the following research question. **RQ2:** *Do the contacted endpoints significantly change based on the user's location?* We investigate if the contacted first, support, and third parties vary based on the user's location. This analysis will help us measure how users' privacy risks (e.g., exposure to third parties) are potentially varied depending on the geographical location. It would be intriguing to understand how often user data crosses national borders and how this affects varying privacy regulations and expectations. However, the IoT-inspector dataset lacks country-specific information but contains users' time zones. Therefore, we use time zone as a more coarse-grained resolution of users' locations.

We first divide smart home devices into three groups to answer this question based on the users' time zone. The first group consists of IoT devices discovered in either North or South America. The second group includes devices from Europe and Africa. The final group comprises devices tracked down in the Asia Pacific and Australia. In the dataset, there is an imbalance in the number of devices (76%, 20%, and 4%) across various regions (Table 9 in Appendix A.3 highlights the number of users and devices across different regions). Most of the IoT devices in the dataset come from the North American region. European and African regions are second regarding the total number of devices. There could be two possible reasons behind this imbalance. First, consumers use more smart home devices in the American region. Second, more users from the American region used the IoT Inspector software. This imbalance in the number of devices across various regions makes it difficult to compare the number of different parties contacted by the same number of devices across various regions. The number of different parties positively varies with the number of devices in each region. In Figure 4, we, therefore, show the average number of different parties communicated by each device in various regions. We see a mixed variation in the average number of domains communicated by each device across various regions. For example, the average number of third party domain communicated by `Home Appliance` devices is much higher in Asia Pacific region than other two regions. `Media/TV` devices communicate with more different types of remote endpoints in American regions, whereas `Voice Assistant` devices communicate more in the European and African regions. Additionally, across Europe and Africa, devices from 4 out of 8 categories show the highest average communication with third-party domains. Conversely, devices in American regions generally engage more with first-party and support entities on average.

We also look at **RQ3:** *How frequently does user data cross borders?* Here, we want to shed light on how frequently data from IoT devices cross regulated borders, which is an interesting question to investigate given that different cross-border data transfer regulations (e.g., GDPR [76], CBPR [70], CCPA [57], VCDPA [25], etc.) impose particular requirements about data transfer across borders. Furthermore, we found that IoT devices often transmit data with remote hosts in other countries. We used country-level information of remote endpoints from the IoT-Inspector dataset for this analysis. The IoT-Inspector team used the Maxmind geolocation (99.8% accuracy) database [32] to locate remote endpoints. Since we clustered users into three regions, we also analyzed to what extent data cross these three regions. Figure 5 represents a heatmap of the outgoing data flow (i.e., considering only upstream traffic) percentage going to different regions. In each subfigure, the rows refer to where the users are located, and the columns indicate where the remote endpoints are located. The diagonal indicates how much data stays inside the respective regional boundary. Here, we see that the first

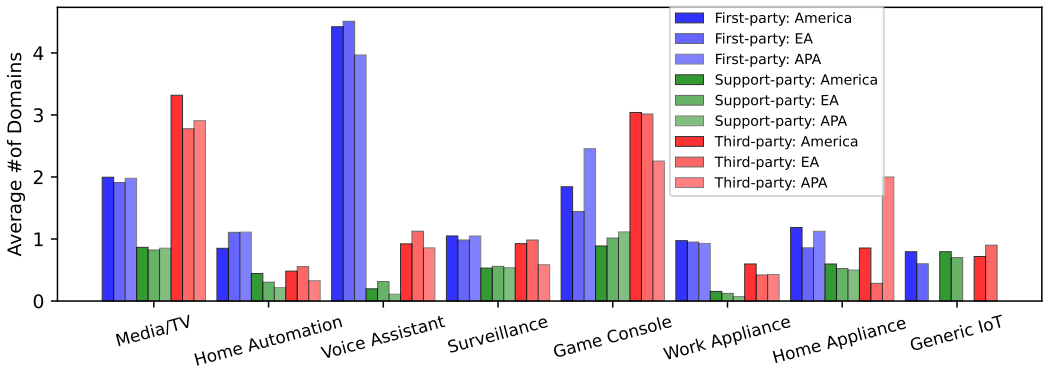
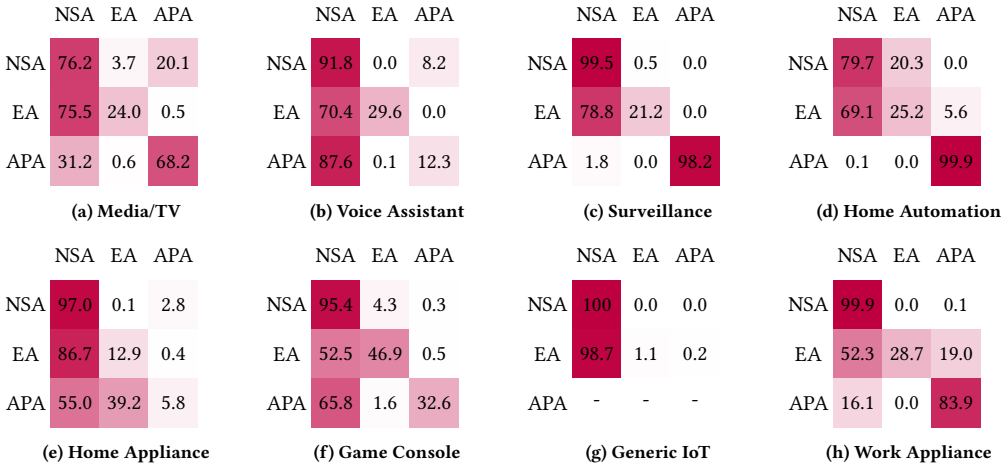


Figure 4: Average number of domains communicated by each device of different party type in different region.



(a) Media/TV **(b) Voice Assistant** **(c) Surveillance** **(d) Home Automation**

(e) Home Appliance **(f) Game Console** **(g) Generic IoT** **(h) Work Appliance**

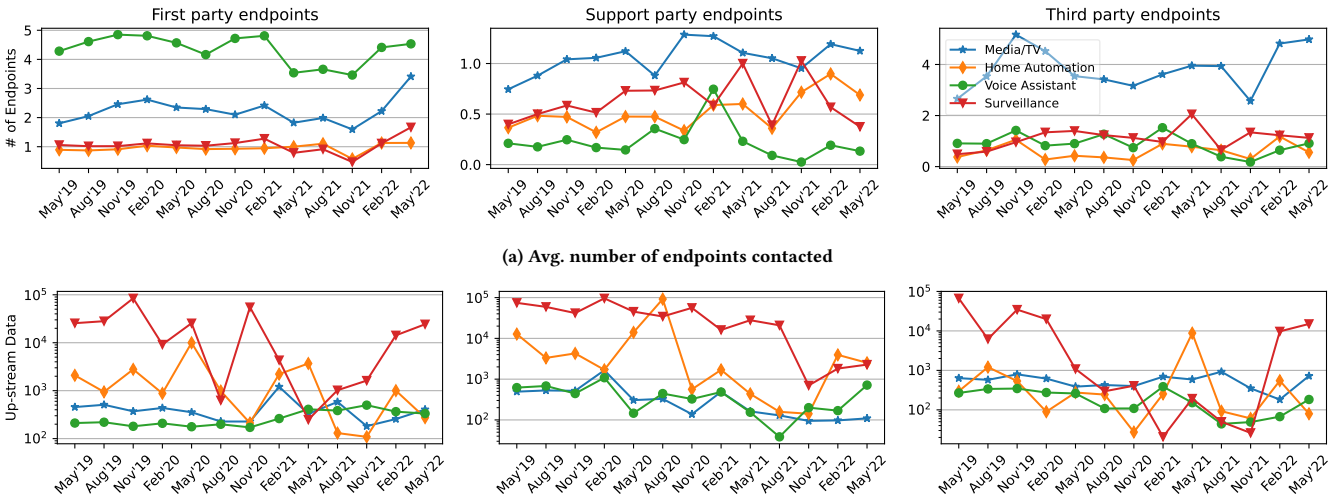
Figure 5: Percentage of outgoing data flow crossing borders. Row: User Location, Column: Destination location. NSA: North & South America, EA: Europe & Asia, APA: Asia Pacific & Australia.

column stands out for almost all the device categories. This means that these devices send most of their payload to the North and South American regions, no matter where they are located. Such analysis can potentially shed light on non-compliance with privacy regulations. For example, GDPR mandates companies to follow strict data transfer policies (e.g., the privacy shield program [44]) from the European Union region. The data flow distribution to different types of endpoints across regions is shown in Figure 8 of Appendix D. There, we observe that most data payload crossing borders go to first parties, with a substantial portion going to third parties. Moreover, the significant volume of payload crossing the border from American regions mostly goes to third-party endpoints.

We also investigated if the privacy policies of remote endpoints address cross-border data transfer regulations. We run MAPS [81] crawler to crawl privacy policies from the effective domains (eTLD) of remote endpoints. This crawler utilize a logistic regression classifier (99.0% accuracy and a 99.2% F1 score) to detect if a given document is privacy policy or not. With this crawler, we find 59% of the domains in the dataset has at least one privacy document in the web. Thereafter, we want to see if those privacy policies talks about international audience. We use PrivBERT [69], a pre-trained privacy policy language model to build a binary classifier that classify each sentence of a privacy policy document whether it addresses data practice of *'International & Specific Audiences'* or not. In this context, we are referring to a case where a privacy policy includes guidelines that apply exclusively to a particular set of users, such as children, individuals from Europe, or residents of California [79].

We train the classifier model with OPP-115 dataset of privacy policy annotations [79] and achieve 99% accuracy. By using this classifier, we found that among the domains which has at least one privacy policy document in the web, only 46% of them talks about international audience. Please be aware that merely examining privacy policies to find mentions of international audience does not constitute a rigorous compliance assessment. Instead, it represents a preliminary effort to determine whether remote endpoints recognize the data transfer across international borders. The assessment of privacy compliance with regulations requires specialised subject expertise; therefore, we exclude it from the scope of this paper.

Summary. We see that based on the users' location, IoT devices contact not only different types of endpoints but also the average number of contacted endpoints varies. As such, the level of privacy risks could potentially depend on the users' geographical location. Moreover, data from the IoT devices cross borders where most endpoints end up in North and South American regions. This has an interesting privacy implication from regulations such as GDPR



(a) Avg. number of endpoints contacted

(b) Avg. upstream data volume

Figure 6: Fine-grained temporal variation in (a) the avg. number of endpoints and (b) the volume of upstream data over a three-month window.

and CBPR. Our data shows that about 59% of remote endpoints have online privacy policies, but just 46% of them address data practice of international and specific audiences. This finding paves the way for future research into whether this pervasive transnational data traffic violates regional data privacy regulations.

6.3 Longitudinal Analysis

This section dives deep into **RQ4:** *Does the distribution of different types of contacted endpoints vary over time?* We want to understand if the endpoints associated with a device shift temporarily. This analysis will help us grasp the fluctuations in privacy risks for a device category over time and identify potential influencing factors.

The dataset includes network traffic from smart home devices between April 2019 and July 2022. We analyze the fine-grained changes in terms of the average number of endpoints IoT devices communicate and the volume of upstream data per second. We analyze these changes over a three-month window, and the outcome is shown in Figure 6. From the figure, it becomes evident that fluctuations occur over the temporal durations. Additionally, a plausible correlation between the utilization of IoT devices and the occurrence of the COVID-19 waves comes to light. The data indicates an escalation in both the average number of endpoints communicated with by IoT devices and the corresponding volume of upstream data from the pre-COVID period (prior to February '20) to the first wave of the COVID-19 outbreak (between March '20 and August '20). We also observe a comparable shift within the timeframe of the second surge of the COVID-19 outbreak, spanning from September '20 to April '21. However, our speculations are hypothetical as we lack ground truth (the IoT inspector team did not consult end-users on device usage changes).

Fine-grained temporal variation in the average number of endpoints across various regions is shown in Figure 9 of Appendix E. A temporal analysis using a more coarse-grained time interval is outlined in Appendix E. The findings indicate a progressive growth in the support-party back-end infrastructure for IoT devices, with North America and Europe taking the lead in this development.

We also wanted to understand how the data flow across borders has evolved over time. To analyze this, we split the dataset into two intervals to gain a broader perspective on these changes. The first segment of the dataset includes network flow data collected between April 2019 and December 2020. The second portion of the dataset contains data collected between January 2021 and July 2022. We then separately repeat our analysis for **RQ3** on these two subsets. In Figure 7, we show the temporal shift of remote endpoints' locations. The cross-border statistics shown here exhibit a changing trend both before (the row on top) and after (the row on the bottom) January 2021. We present statistics for categories for which we have at least three devices in both periods. From Figure 7, it is evident that the data transmission practice has shifted towards the North and South American regions for most of the device categories, which means that in the latter period, devices are sending more data payload to the western region than other regions.

Summary. We see a significant change in the communication pattern over time. Our analysis reveals that the average number of support parties contacted by each device has grown, suggesting that the supporting back-end infrastructure for IoT devices is expanding over time. Also, IoT infrastructure evolves faster in the American and European regions than Asia Pacific region. We also observe a gradual shift of data towards western regions. This transition has implications for cross-border data privacy regulations such as the EU-US Privacy Shield, CCPA, VCDPA, etc.

7 DISCUSSION

Ethical Considerations. Our institution's IRB approved our use of IoT Inspector's dataset through a reliance agreement. We follow industry-standard security and privacy practices to safeguard the data and limit access only to the co-authors. Crowdsourcing IoT traffic is covered by the original IoT-Inspector paper [30].

Limitations. (i) Although the dataset comprises real-world data from April 2019 to July 2022, the data population at the former timespan is larger than that at the later timespan. This is because

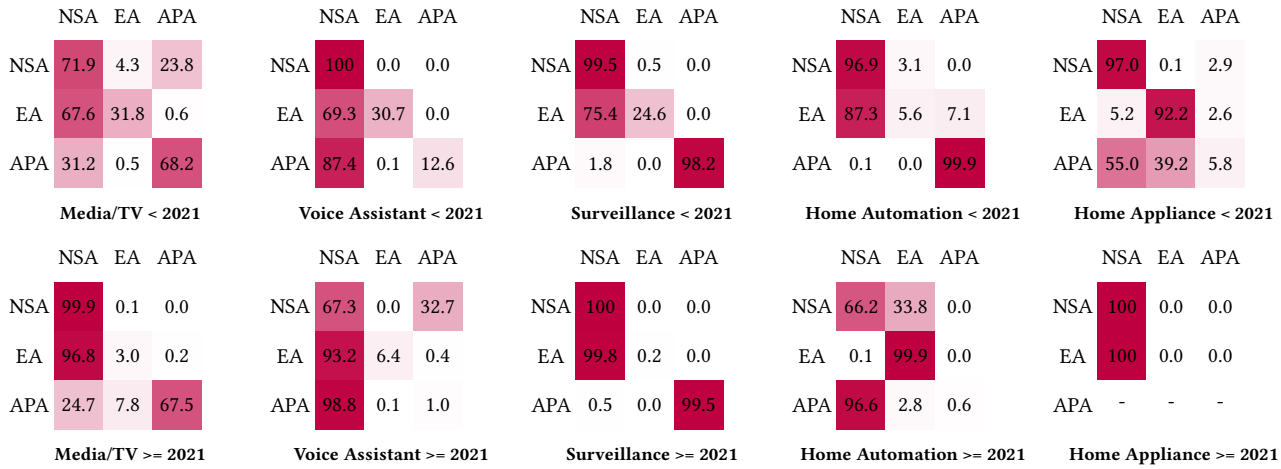


Figure 7: Temporal change in percentage of outgoing data flow crossing borders. Row: User Location, Column: Destination location. NSA: North & South America, EA: Europe & Asia, APA: Asia Pacific & Australia

users adopted IoT Inspector software more widely at the start of its release, but its popularity faded with time [30], hence presenting challenges (e.g., more recent sample sizes are smaller) for our longitudinal analysis. **(ii)** The network traffic we observe in the dataset is a result of user interactions (or the lack thereof) with individual devices. The amount and duration of traffic a device generates could vary depending on the user's interaction with that device. Also, the sample sizes for less popular devices are naturally smaller. This makes it difficult to gather more controlled information from this analysis (unlike in-lab experiments). **(iii)** The device information is not well-organized; some devices do not have user annotations or mDNS/UPnP information, which are essential for product identification. Even if a device provides mDNS/UPnP or User Agent, the information may not be indicative of the actual product [59]. As such, we have been unable to catalog more than 9,000 gadgets. We *exclude* such a large number of devices from our research because we lack the means to identify them properly. This privacy analysis can be expanded once we are able to identify more smart devices in the wild. **(iv)** In the methodology, we employ non-domain-specific, generic topic modeling to identify support parties. For feature extraction, we employ spaCy tagger, which has a 97.8% accuracy rate. With a domain-specific tagger and a domain-specific topic model, the performance of our method can be enhanced. **(v)** We analyze the endpoints that IoT devices communicated with, although we have no knowledge what data is shared between IoT devices and these endpoints, because the IoT Inspector dataset does not capture the traffic payload. We keep these as our future works.

Recommendations. According to our findings, home IoT devices communicate with a variety of remote endpoints located in various locations. Consumers should choose an IoT device carefully, possibly preferring one that shares data with a small number of tracking endpoints. Users should pay closer attention to a home device's data-sharing practices. Furthermore, as previously demonstrated in research, users can block all third-party domains without interfering with device functionality [47]. We recommend that users examine the settings of smart home device apps to determine whether they have any control over data sharing. Furthermore, device manufacturers should highlight what information is collected and where it is sent. Arlo, for example, defines a list of external services or providers that their devices use explicitly. In addition, data protection regulatory organizations should investigate the IoT ecosystem's data-sharing practices more closely. Consumer privacy laws (GDPR, CCPA, etc) can be revised to focus on the IoT ecosystem, which will benefit users' privacy the most. We plan to release a comprehensive report summarizing our analysis for public use.

8 CONCLUSION

Smart home devices collect user data and transmit it to various remote endpoints. There is little understanding of where the information of the users is going. In this paper, we conducted a comprehensive study of network traffic generated by a considerable sample of smart home devices belonging to users in different parts of the world. We analyzed where data from home devices ended up and the relationship between device manufacturers and remote endpoints. First, we identified and depicted organizations corresponding to the device manufacturer and the remote endpoints. We discovered that most remote endpoints corresponded to third parties or support parties, with only a few domains representing the device manufacturer. We categorized IoT devices based on the functionality these devices provide and then compared the communication patterns across these categories. Then, we identified the major players in the IoT ecosystem support parties group and determined which portion of the ecosystem system will be impacted if one of these major players is compromised. We also investigated cross-border data-sharing practices and how they evolved over time.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their valuable feedback. This material is based upon work supported in parts by the National Science Foundation (NSF) under grant number CNS-2219866, CNS-2219867 and the Center for Accelerated Real Time Analytics (CARTA) - NCSU Research Site. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Abbas Acar, Hossein Fereidooni, Tigist Abera, Amit Kumar Sikder, Markus Miettinen, Hidayet Aksu, Mauro Conti, Ahmad-Reza Sadeghi, and Selcuk Uluagac. 2020. Peek-a-boo: i see your smart home activities, even encrypted!. In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. Association for Computing Machinery, New York, NY, USA, 207–218.

[2] Dilawer Ahmed, Anupam Das, and Fareed Zaffar. 2022. Analyzing the Feasibility and Generalizability of Fingerprinting Internet of Things Devices. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2022, 2 (2022), 578–600.

[3] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[4] N Alhalafi and Prakash Veeraraghavan. 2019. Privacy and Security Challenges and Solutions in IOT: A review. *IOP Conference Series: Earth and Environmental Science* 322, 1 (2019), 012013.

[5] Omar Alrawi, Chaz Lever, Manos Antonakakis, and Fabian Monrose. 2019. SoK: Security Evaluation of Home-Based IoT Deployments. In *2019 IEEE Symposium on Security and Privacy (SP)* (San Francisco, CA, USA). IEEE Computer Society, New York, NY, USA, 1362–1380.

[6] Noah Apthorpe, Danny Yuxing Huang, Dillon Reisman, Arvind Narayanan, and Nick Feamster. 2019. Keeping the smart home private with smart(er) IoT traffic shaping. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2019, 3 (2019), 128–148.

[7] Noah Apthorpe, Dillon Reisman, and Nick Feamster. 2017. Closing the Blinds: Four Strategies for Protecting Smart Home Privacy from Network Observers. *CoRR* abs/1705.06809 (2017).

[8] Noah Apthorpe, Dillon Reisman, and Nick Feamster. 2017. A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic. arXiv:1705.06805

[9] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. 2017. Spying on the Smart Home: Privacy Attacks and Defenses on Encrypted IoT Traffic. *CoRR* abs/1708.05044 (2017).

[10] Leonardo Babun, Kyle Denney, Z. Berkay Celik, Patrick McDaniel, and A. Selcuk Uluagac. 2021. A survey on IoT platforms: Communication, security, and privacy perspectives. *Computer Networks* 192 (2021), 108040.

[11] Sharu Bansal and Dilip Kumar. 2020. IoT ecosystem: A survey on devices, gateways, operating systems, middleware and communication. *International Journal of Wireless Information Networks* 27, 3 (2020), 340–364.

[12] Bruhadeshwar Bezawada, Maalvika Bachani, Jordan Peterson, Hossein Shirazi, Indrakshi Ray, and Indrajit Ray. 2018. Behavioral Fingerprinting of IoT Devices. In *ACM Workshop on Attacks and Solutions in Hardware Security*. Association for Computing Machinery, New York, NY, USA, 41–50.

[13] Suman Sankar Bhunia and Mohan Gurusamy. 2017. Dynamic attack detection and mitigation in IoT using SDN. In *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–6.

[14] Stuart Cheshire and Marc Krochmal. 2013. Multicast DNS. https://www.ietf.org/rfc/rfc6762.txt.

[15] Bogdan Copos, Karl Levitt, Matt Bishop, and Jeff Rowe. 2016. Is anybody home? Inferring activity from smart home network traffic. In *IEEE Security and Privacy Workshops (SPW)*. IEEE, San Jose, CA, USA, 245–251.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[17] DuckDuckGo. 2023. DuckDuckGo Tracker Radar. https://duckduckgo.com.

[18] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the experts: What should be on an IoT privacy and security label?. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 447–464.

[19] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2021. Which privacy and security attributes most impact consumers' risk perception and willingness to purchase IoT devices?. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 519–536.

[20] Lesley Fair. 2023. Not home alone: FTC says Ring's lax practices led to disturbing violations of users' privacy and security. https://www.ftc.gov/business-guidance/blog/2023/05/not-home-alone-ftc-says-rings-lax-practices-led-disturbing-violations-users-privacy-security.

[21] Xuan Feng, Qiang Li, Qi Han, Hongsong Zhu, Yan Liu, Jie Cui, and Limin Sun. 2016. Active Profiling of Physical Devices at Internet Scale. In *25th International Conference on Computer Communication and Networks*. IEEE, Waikoloa, HI, USA, 1–9.

[22] FingerBank. 2023. FingerBank. https://www.fingerbank.org/

[23] A. Gilchrist. 2017. *IoT Security Issues*. De Gruyter, Incorporated, Berlin, Germany.

[24] Tomer Golomb, Yisroel Mirsky, and Yuval Elovici. 2018. CIoTA: Collaborative IoT Anomaly Detection via Blockchain. arXiv:1803.03807

[25] Virginia Gov. 2023. Virginia Consumer Data Protection Act (VCDPA). https://law.lis.virginia.gov/vacode/title59.1/chapter53/.

[26] Ibbad Hafeez, Aaron Yi Ding, Lauri Suomalainen, Alexey Kirichenko, and Sasu Tarkoma. 2016. Securebox: Toward Safer and Smarter IoT Networks. In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking* (Irvine, California, USA). Association for Computing Machinery, New York, NY, USA, 55–60.

[27] Ibbad Hafeez, Aaron Yi Ding, and Sasu Tarkoma. 2017. IOTURVA: Securing Device-to-Device (D2D) Communication in IoT Networks. In *Proceedings of the 12th Workshop on Challenged Networks* (Snowbird, Utah, USA). Association for Computing Machinery, New York, NY, USA, 1–6.

[28] Gilbert Held. 1998. Cinco Network's WebXRay. *International Journal of Network Management* 8, 4 (1998), 254–261.

[29] Signify Holding. 2023. Amazon Alexa is a perfect sidekick to Philips Hue, giving you hands-free voice control of your smart lights. https://www.philips-hue.com/en-us/explore-hue/works-with/amazon-alexa.

[30] Danny Yuxing Huang, Noah Apthorpe, Frank Li, Gunes Acar, and Nick Feamster. 2020. IoT Inspector: Crowdsourcing Labeled Network Traffic from Smart Home Devices at Scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 46 (2020), 21 pages.

[31] Disconnect Inc. 2023. Disconnect filter list. https://disconnect.me/.

[32] MaxMind Inc. 2023. Maxmind geolocation database. http://www.maxmind.com.

[33] NETGEAR Inc. 2023. Legacy Arlo Products. https://www.netgear.com/about/legacyarloproducts/.

[34] OpenAI Inc. 2023. OpenAI's TextCompletion API. https://platform.openai.com/docs/guides/text-generation/completions-api.

[35] PubNub Inc. 2023. PNDSN.COM - Domain Info. https://www.netify.ai/resources/domains/pndsn.com.

[36] Fortune Business Insights. 2023. Internet of things [IOT] market size, share & trends, 2029. https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-100307.

[37] Alex Koohang, Carol Springer Sargent, Jeretta Horn Nord, and Joanna Paliszkiewicz. 2022. Internet of Things (IoT): From awareness to continued use. *International Journal of Information Management* 62 (2022), 102442.

[38] J. Nick Koston, Pascal Vizeli, and Patrik Lindgren. 2023. Netdisco. https://github.com/home-assistant-libs/netdisco

[39] Deepak Kumar, Kelly Shen, Benton Case, Deepali Garg, Galina Alperovich, Dmitry Kuznetsov, Rajarshi Gupta, and Zakir Durumeric. 2019. All Things Considered: An Analysis of IoT Devices on Home Networks. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 1169–1185.

[40] Greg Kumparak. 2014. Google and Nest Acquire Dropcam For $555 Million. https://techcrunch.com/2014/06/20/google-and-nest-acquire-dropcam-for-555-million/.

[41] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 260–270.

[42] Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems* (Denver, CO) *(NIPS'00)*. MIT Press, Cambridge, MA, USA, 535–541.

[43] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Physics Doklady, Soviet Union, 707–710.

[44] Emily Linn. 2017. A Look into the Data Privacy Crystal Ball: A Survey of Possible Outcomes for the EU-US Privacy Shield Agreement. *Vand. J. Transnat'l L.* 50 (2017), 1311.

[45] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of machine learning research* 2, Feb (2002), 419–444.

[46] Stan Entertainment Pty Ltd. 2023. Stan Website. https://www.stan.com.au/.

[47] Anna Maria Mandalari, Daniel J. Dubois, Roman Kolcun, Muhammad Talha Paracha, Hamed Haddadi, and David Choffnes. 2021. Blocking Without Breaking: Identification and Mitigation of Non-Essential IoT Traffic. *Proceedings on Privacy Enhancing Technologies* 2021, 4 (2021), 369–388.

[48] M. Mazhar and Z. Shafiq. 2020. Characterizing Smart Home IoT Traffic in the Wild. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE Computer Society, Los Alamitos, CA, USA, 203–215.

[49] Samuel Mergendahl, Devkishen Sisodia, Jun Li, and Hasan Cam. 2017. Source-End DDoS Defense in IoT Environments. In *Proceedings of the 2017 workshop on internet of things security and privacy* (Dallas, Texas, USA). Association for Computing Machinery, New York, NY, USA, 63–64.

[50] Markus Miettinen, Samuel Marchal, Ibbad Hafeez, N Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. 2017. IoT SENTINEL: Automated device-type identification for security enforcement in IoT. In *Proceedings of the 37th IEEE International Conference on Distributed Computing Systems (ICDCS)*. IEEE, Atlanta, GA, USA, 2177–2184.

[51] Brent A Miller, Toby Nixon, Charlie Tai, and Mark D Wood. 2001. Home networking with universal plug and play. *IEEE Communications Magazine* 39, 12 (2001), 104–109.

[52] Khrin MonztA, Famlam. 2023. Easylist. https://easylist.to/easylist/easylist.txt.

[53] MQTT.org. 2024. MQ Telemetry Transport. https://mqtt.org/

[54] Mehdi Nobakht, Vijay Sivaraman, and Roksana Boreli. 2016. A host-based intrusion detection and mitigation framework for smart home IoT using OpenFlow. In *11th International conference on availability, reliability and security (ARES)* (Salzburg, Austria). IEEE, New York, NY, USA, 147–156.

[55] TJ OConnor, Reham Mohamed, Markus Miettinen, William Enck, Bradley Reaves, and Ahmad-Reza Sadeghi. 2019. HomeSnitch: behavior transparency and control for smart home IoT devices. In *Proceedings of the 12th conference on security and privacy in wireless and mobile networks*. ACM, New York, NY, USA, 128–138.

[56] Muhammad Talha Paracha, Daniel J. Dubois, Narseo Vallina-Rodriguez, and David Choffnes. 2021. IoTLS: understanding TLS usage in consumer IoT devices. In *Proceedings of the 21st ACM Internet Measurement Conference* (Virtual Event) *(IMC '21)*. Association for Computing Machinery, New York, NY, USA, 165–178.

[57] Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y* 23 (2018), 68.

[58] Roberto Perdisci, Thomas Papastergiou, Omar Alrawi, and Manos Antonakakis. 2020. IoTFinder: Efficient Large-Scale Identification of IoT Devices via Passive DNS Traffic Analysis. In *European Symposium on Security and Privacy (EuroS&P)* (Genoa, Italy). IEEE, New York, NY, USA, 474–489.

[59] Vijay Prakash, Sicheng Xie, and Danny Yuxing Huang. 2022. Inferring Software Update Practices on Smart Home IoT Devices Through User Agent Analysis. In *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses* (Los Angeles, CA, USA) *(SCORED'22)*. Association for Computing Machinery, New York, NY, USA, 93–103.

[60] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, Phillipa Gill, et al. 2018. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem, In The 25th Annual Network and Distributed System Security Symposium (NDSS 2018) (San Diego, California). *Network and Distributed Systems Security (NDSS) Symposium* 2018.

[61] Jingjing Ren, Daniel J. Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. 2019. Information Exposure From Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 267–279.

[62] Said Jawad Saidi, Srdjan Matic, Oliver Gasser, Georgios Smaragdakis, and Anja Feldmann. 2022. Deep dive into the IoT backend ecosystem. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 488–503.

[63] Armin Sarabi and Mingyan Liu. 2018. Characterizing the Internet Host Population Using Deep Learning: A Universal and Lightweight Numerical Embedding. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 133–146.

[64] Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Parallel Corpus Filtering Based on Fuzzy String Matching. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Association for Computational Linguistics, Florence, Italy, 289–293.

[65] Zain Shamsi, Ankur Nandwani, Derek Leonard, and Dmitri Loguinov. 2016. Hershel: Single-Packet OS Fingerprinting, In ACM SIGMETRICS Conference. *IEEE/ACM Transactions on Networking* 24, 4, 2196–2209.

[66] Arunan Sivanathan, Hassan Habibi Gharakheili, Franco Loi, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. 2018. Classifying IoT devices in smart environments using network traffic characteristics. *IEEE Transactions on Mobile Computing* 18, 8 (2018), 1745–1759.

[67] Arunan Sivanathan, Daniel Sherratt, Hassan Habibi Gharakheili, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. 2017. Characterizing and classifying IoT traffic in smart cities and campuses. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Atlanta, GA, USA, 559–564.

[68] Arunan Sivanathan, Daniel Sherratt, Hassan Habibi Gharakheili, Vijay Sivaraman, and Arun Vishwanath. 2016. Low-cost flow-based security solutions for smart-home IoT devices. In *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)* (Bangalore, India). IEEE, New York, NY, USA, 1–6.

[69] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6829–6839.

[70] Cross Border Privacy Rules System. 2023. The APEC Cross-Border Privacy Rules (CBPR) System. http://cbprs.org/.

[71] Lo'ai Tawalbeh, Fadi Muheidat, Mais Tawalbeh, and Muhannad Quwaider. 2020. IoT Privacy and security: Challenges and solutions. *Applied Sciences* 10, 12 (2020), 4102.

[72] Rahmadi Trimananda, Janus Varmarken, Athina Markopoulou, and Brian Demsky. 2020. Packet-Level Signatures for Smart Home Devices, In Proceedings of the 27th Annual Network and Distributed System Security Symposium (NDSS). *Network and Distributed Systems Security (NDSS) Symposium* 2020.

[73] Shafiq Ul Rehman, Parminder Singh, Selvakumar Manickam, and Supriyanto Praptodiyono. 2020. Towards Sustainable IoT Ecosystem. In *2020 2nd International Conference on Industrial Electrical and Electronics (ICIEE)*. IEEE, Lombok, Indonesia, 135–138.

[74] Janus Varmarken, Hieu Le, Anastasia Shuba, Athina Markopoulou, and Zubair Shafiq. 2020. The tv is smart and full of trackers: Measuring smart tv advertising and tracking. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 129 – 154.

[75] Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, San Francisco, CA, USA.

[76] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[77] Sean Whalen. 2001. An introduction to arp spoofing. https://priv.gg/e/arp_spoofing_intro.pdf

[78] Who.is. 2023. WHOIS Search. https://who.is/.

[79] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1330–1340.

[80] Tianlong Yu, Vyas Sekar, Srinivasan Seshan, Yuvraj Agarwal, and Chenren Xu. 2015. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the Internet-of-Things. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks* (Philadelphia, PA, USA). Association for Computing Machinery, New York, NY, USA, 7 pages.

[81] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2019 (2019), 66.

APPENDIX

A DATASET

A.1 Device Selection and Categorization

Table 5: Device categorization

Generic Category	Example
Computer	Laptop/computer
Router	Any networking device for e.g. router, access point, extender, mesh router, internet gateway
Mobile Device	Phone, Tablet, E-readers
Storage	NAS, or anything that could store media which could be streamed somewhere over the network, DLNA DMS (Digital Media Server)
Wearable	Smartwatch, hand ring, or band
Other	Can't say if it's IoT
Home Automation	Anything related to smart lights, switches, plugs, strips, window sensors, weather sensor, thermostat, garage door opener, lock, air purifier, vacuum etc.
Surveillance	Any smart/IP camera, or smart doorbell
Work Appliance	Printer, fax, VoIP etc.
Voice Assistant	Any speaker based voice assistant, Nest Hub and Echo show
Vehicle	Cars, and other vehicles
Media/TV	Streamers, smart TVs, and voice assistant with display for e.g. DLNA DMR (Digital Media Renderer), also the case of both DMS and DMR, and any Receiver, smart speakers, and audio streamers, Internet radio device
Home Appliance	Smart refrigerator, coffee maker, AC, Purifier, etc.
Generic IoT	toothbrush, medical devices, and anything that doesn't fit in the above categories but it's and IoT.
Game Console	Xbox, Play station

Table 6: Generic category mapping. Source category: *category identified by ChatGPT or manual labeling,* Generic category: *mapped category according to Table 5,* Example vendor: *some examples within that category.*

Source Category	Generic Category	Example Vendor
Laptop	Computer	microsoft, lenovo, google, dell, apple, intel, samsung, toshiba
Video Game Console	Game Console	sony, microsoft, apple
Smart Clock	Generic IoT	lenovo, insignia, lametric
Washing Machine	Home Appliance	samsung, lg
Smart Home Controller	Home Automation	brilliant, philips, amazon, lutron, logitech, wink, samsung
AV Receiver	Media/TV	denon, onkyo, yamaha, marantz, pioneer, apple, tivo, sony
HDMI Switch	Other	caavo
Wireless Access Point	Router	dlink, netgear, ubiquiti, tplink, devolo
Network Storage Device	Storage	western digital, synology, buffalo, plex
Doorbell	Surveillance	skybell, ring, wyze, nest, hikvision
Vehicle IoT Device	Vehicle	subaru, tesla
Echo Dot	Voice Assistant	amazon
Smartwatch	Wearable	apple, samsung, xiaomi
Printer and Scanner	Work Appliance	hp, brother

Table 7: A sample list describing mismatch between chatgpt and manual device identification.

GPT Vendor Name	GPT Category	Manual Vendor Name	Manual Category	Device Count
roku	Media/TV	tcl	Media/TV	38
google	Media/TV	vizio	Media/TV	22
sensi	Home Automation	emerson	Home Automation	11
spotify	Media/TV	roku	Media/TV	7
connect	Home Automation	connectsense	Home Automation	5
vocol	Home Automation	vocolinc	Home Automation	5
google	Media/TV	nvidia	Media/TV	4
microsoft	Storage	intel	Computer	4
google	Media/TV	xiaomi	Media/TV	4
amc	Surveillance	amcrest	Surveillance	4
roku	Media/TV	insignia	Media/TV	4
roku	Media/TV	sharp	Media/TV	3
wyze	Surveillance	ismart	Surveillance	3
pioneer	Media/TV	onkyo	Media/TV	3
asus	Computer	asustek	Computer	3
devolo	Router	ubiquiti	Router	3
naimaudio	Media/TV	naim	Media/TV	3
echostar	Media/TV	dish	Media/TV	3
plex	Media/TV	synology	Storage	3
microsoft	Storage	asustek	Computer	3
linksys	Router	belkin	Router	3
..	..	..	..	..

A.2 Validation of Device Identification Method

Table 8: A list showcasing the presence of metadata for unidentified devices in the auxiliary datasets.

Mac/IP Address	Ground Truth	dhcp response	mdns response	ssdp response	upnp response	user agent	hostnames
UNSW IoT Analytics [66] dataset							
00:62:6e:51:27:2e	Insteon Camera	✗	✗	✗	✗	✓	✗
14:cc:20:51:33:ea	TPLink Router	✓	✗	✓	✗	✓	✗
18:b4:30:25:be:e4	Nest Protect Smoke Alarm	✗	✗	✗	✗	✗	✗
30:8c:fb:2f:e4:b2	Dropcam	✗	✗	✗	✗	✗	✗
30:8c:fb:b6:ea:45	Dropcam	✗	✗	✗	✗	✗	✗
70:ee:50:03:b8:ac	Netatmo weather station	✗	✗	✗	✗	✗	✗
74:6a:89:00:2e:25	Blipcare Blood Pressure meter	✗	✗	✗	✗	✗	✗
74:c6:3b:29:d7:1d	iHome	✓	✓	✗	✗	✓	✗
d0:52:a8:00:67:5e	Smart Things	✓	✗	✗	✗	✗	✗
d0:73:d5:01:83:08	LiFX Smart Bulb	✓	✗	✗	✗	✗	✗
e0:76:d0:33:bb:85	PIX-STAR Photo-frame	✗	✗	✗	✗	✓	✓
f4:f2:6d:93:51:f1	TP-Link Camera	✓	✓	✗	✗	✓	✗
YourThings IoTFinder [58] datasets							
192.168.0.1	Gateway	✓	✗	✗	✗	✓	✗
192.168.0.10	NestCamera	✗	✓	✓	✗	✗	✗
192.168.0.13	LIFXVirtualBulb	✓	✓	✗	✗	✗	✗
192.168.0.16	WinkHub	✗	✗	✗	✗	✗	✗
192.168.0.17	NestProtect	✗	✗	✗	✗	✗	✗
192.168.0.19	RingDoorbell	✗	✗	✗	✗	✗	✗
192.168.0.2	GoogleOnHub	✗	✓	✗	✗	✗	✓
192.168.0.30	Canary	✓	✓	✓	✗	✓	✓
192.168.0.35	ChineseWebcam	✗	✓	✓	✗	✗	✗
192.168.0.4	SamsungSmartThingsHub	✓	✗	✓	✗	✓	✗
192.168.0.45	HarmonKardonInvoke	✓	✓	✓	✗	✗	✓
192.168.0.6	InsteonHub	✓	✓	✓	✗	✗	✗
192.168.0.8	SecurifiAlmond	✗	✓	✓	✗	✗	✗

A.3 Distribution of users and devices across various region

Table 9: Distribution of number of devices and participants across various regions

Category	North & South America		Europe & Africa		Asia Pacific & Australia		Unknown Location	
	#Device	#User	#Device	#User	#Device	#User	#Device	#User
Media/TV	6179	2401	1937	908	392	188	190	72
Voice Assistant	2808	1376	716	438	154	94	90	49
Surveillance	978	552	186	123	41	24	37	20
Home Automation	4976	1818	1181	692	260	105	149	51
Home Appliance	124	112	21	19	8	8	2	1
Game Console	499	443	133	120	35	27	20	14
Generic IoT	39	36	10	10	0	0	2	2
Work Appliance	375	339	81	78	14	12	6	5
Vehicle	14	13	0	0	0	0	1	1

B ENDPOINTS MAPPING

B.1 Support Party Mapping

In this section, we add cluster labeling described in subsection 5.2. In NMF model with TF-IDF approach (Table 10), we consider cluster ID 8, 12, 14 to be the support party. In NMF model with counting approach (Table 11), we consider cluster ID 4, 5, 15 to be the support party.

Table 10: Labeling clusters from NMF with TF-IDF approach. Topic: *most influencing words of a cluster*, Label: *Manual labels representing the clusters*

ID	Topic	Label
0	domain whois registrar dns lookup price information registration url registry	miscellaneous
1	tv channel television live streaming iptv cable content service app	tv/streaming services
2	movie free streaming series site movies tv website netflix online	video streaming services
3	cookie tracker directory betters script netifys directories cookiepedias site confection	ad/tracker
4	app mobile android user developer application device platform ios analytic	mobile apps services
5	radio music station france public fm audio stream npr streaming	radio/music
6	google search engine image webpage googlecom googles special llc world	search engines
7	browser adware firefox ad popup virus hijacker web website unwanted	adware/popup services
8	cdn delivery content network akamai edge server distribution cdns user	cdn/cloud
9	certificate ssl authority security ca comodo sectigo digital secure certificates	ssl service
10	ip address ovh dns sas nsipnet lookup ipv location country	dns service
11	email account mail imap webmail password smtp address desktop program	web-mail service
12	cloud storage service computing iot aws infrastructure device solution application	iot/cloud service
13	game games xbox gaming solitaire player developer playstation epic minecraft	gaming services
14	home smart device wifi camera control automation product security light	iot platforms
15	amazon aws prime amazoncom services alexa service music amazons customer	e-commerce service
16	samsung galaxy bixby electronics samsungs pihole blocklist device assistant smartphone	samsung services
17	company customer business platform datum software product management solution marketing	business/it
18	advertising ad publisher platform programmatic advertiser technology monetization digital mobile	advertising
19	podcast audio podcaster listener podcasting episode platform libsyn distribution spotify	audio/music
20	video streaming live content platform youtube medium player stream camera	video streaming
21	weather forecast radar api hyperlocal datum national local temperature station	weather forcast
22	server dedicated hosting minecraft provider servers dns client colocation seedbox	web
23	website visitor traffic rank scam site web unique day link	miscellaneous
24	news channel sport live television entertainment nbc cbs story fox	news/sports/social
25	torrent file bittorrent download eztv site seedbox search music pirate	torrent/file server
26	net api open aspnet framework foundation source development visual network	miscellaneous
27	proxy vpn hotspot luminati telegram xyz privacy internet private ip	vpn/proxy
28	internet broadband service speed cable provider phone network connection fiber	tv/internet/broadband
29	microsoft windows office azure file microsofts service onedrive msn version	microsoft services

Table 11: Labeling clusters from NMF with counting approach. Topic: *most influencing words of a cluster*, Label: *Manual labels representing the clusters*

ID	Topic	Label
0	service provider streaming streaming service customer service provider business subscription available	streaming service
1	tv channel television live content streaming smart tv android smart device	streaming service
2	network device world datum wifi large private security wireless access	miscellaneous
3	website web site user visitor link online browser traffic page	miscellaneous/web
4	home smart smart home device control automation product home automation security assistant	smart home/IoT
5	content cdn delivery network content delivery delivery network network cdn user cdn content server	cdn/cloud
6	ip address ip address ovh dns country sas ovh sas location nsipnet	dns
7	radio station music france radio station public podcast fm new public radio	radio/music
8	app mobile user android device mobile app application developer store ios	mobile app
9	video content platform streaming player software wistia video content medium prime	video streaming
10	google search engine search engine world image information feature google search web	search engines
11	email account mail imap address password access email address program desktop	web-mail service
12	domain whois information domain domain registrar dns lookup price url registration	miscellaneous
13	game solitaire card xbox video game mahjong games card game online gaming	gaming services
14	news channel live television sport entertainment cbs story world newsy	news/sports/social
15	cloud storage solution cloud storage datum cloud service service device computing cloud computing	cloud service/iot
16	microsoft file windows software office browser web device application computer	microsoft services
17	cookie tracker directory site business application script betters betters site site tracker	ad/tracker
18	platform customer datum business software product management user tool analytic	business/it
19	amazon aws web services music prime web services amazon web customer amazoncom	e-commerce service
20	server proxy dedicated provider vpn minecraft dedicated server web client datum	proxy/vpn
21	ad advertising platform mobile publisher technology digital advertiser solution marketing	ad/tracker
22	torrent site proxy file torrent site search download eztv website bittorrent	torrent/file server
23	movie free streaming movie tv series tv site online popular streaming service	miscellaneous
24	company technology business product world entertainment medium group large new	business/IT
25	samsung health device samsung cloud galaxy print samsung health cloud print phone bixby	samsung services
26	weather forecast datum information api weather forecast location local time national	weather service
27	internet broadband speed provider internet service connection cable phone service provider mobile	internet/cable
28	digital certificate ssl security authority solution certificate authority ssl certificate provider ca	proxy/vpn
29	sky new sport box york new york sky sky good sky box dark	miscellaneous

C BACK-END INFRASTRUCTURE FOR IOT ECOSYSTEMS

Table 12: Infrastructure shared across IoT devices of various vendors

Infrastructure	Example
Amazon	amazon, elasticbeanstalk, amcs-tachyon, awsglobalaccelerator
Google	google, gstatic, android, nest, gmail
Apple	apple.com, icloud
Heroku	heroku
Time Server	ntp, nist, hshh
Network	one.one, comcast, cloudfront, akamai, opendns, rr.com, proxygo, hotspotproxy, apple-dns, hotspotproxy, proxy4
Facebook	facebook
Microsoft	microsoft, windows, azure
Analytics	domotz, mixpanel
IoT	meethue, pubnub, pndsn, smartthings, dropcam, ecobee, enphaseenergy, netgear, philips, arlo, arloxcld, aylanetworks, wink, telephony

Table 13: List of top vendors that exchange data with multiple support parties. The first column represents the vendor name; the second column represents a list of support parties with the number of devices communicating that support party domain in parenthesis. This table includes vendors with at least four support parties.

Vendor Name	List of Support Parties
sonos	Amazon(1699), Akamai(131), Google(36), Pandora(19), Facebook(9), Microsoft(7)
belkin	Amazon(1163), Google(12), Akamai(10), mtpfast.pw(7), alienvault.cloud(6), Facebook(5)
roku	Amazon(618), Akamai(119), Google(41), AMP(26), Lumen(15), TowerData(12), WordPress(9), Fastly(7), Limelight(7), Pandora(5)
amazon	Philips(257), Akamai(254), Google(39), Fastly(33), Roku(26), Samsung(22), Pandora(20), Limelight(13), Facebook(12), Lumen(11), TowerData(10), SiliconDust(8), Adobe(8), Snapchat(7), Cdk global(7), Microsoft(6), Heroku(6), TP-Link(6)
samsung	Amazon(382), Akamai(138), SmartThings(82), Google(70), Apple(60), Facebook(45), Roku(16), Microsoft(14), Lumen(11), Fastly(10), Limelight(9), AMP(9), Ring(7), CloudFlare(5), Edgecast(5), Synology(5)
google	Akamai(169), Amazon(128), Facebook(83), Fastly(39), CloudFlare(17), Samsung(8), Philips(8), Roku(6), Pandora(6), Xiongmai(5), Microsoft(5)
sony	Google(204), Akamai(160), Amazon(85), Limelight(28), Facebook(17), Fastly(10), Heroku(7), Vultr(5)
wyze	Amazon(287), homeassurednow.com(8), Akamai(6), tutk.com(6)
apple	Akamai(228), Amazon(26), Google(12), Lumen(9), Limelight(7), Fastly(6)
vizio	Amazon(139), Google(113), Apple(21), Akamai(17), CloudFlare(9), Limelight(7)
nvidia	Google(128), Akamai(39), Amazon(35), Facebook(15), Fastly(8), Heroku(8), SiliconDust(6), Edgecast(6), Roku(5)
lg	Amazon(73), Google(61), Akamai(39), Apple(6), TowerData(5), Philips(5)
logitech	Amazon(84), SmartThings(27), Nest(14), Heroku(8)
insignia	Ayla networks(66), Google(43), Amazon(10), Akamai(9), Roku(7)
wink	Amazon(109)
bose	Amazon(92), Google(8), Akamai(5)
microsoft	Akamai(71), Amazon(25), Google(11), GoDaddy(9)
tcl	Amazon(61), Roku(56), Akamai(8)
nintendo	Amazon(38), Akamai(36)
philips	Amazon(30), Akamai(25), Google(20)
mysa	Amazon(67)
lutron	Amazon(57), Nest(15)
amcrest	Amazon(56)
onkyo	Google(42), Streamunlimited engineering(24), Amazon(17)
xiaomi	Amazon(23), Google(19), Akamai(5)
irobot	Amazon(44)
netgear	Arlo(40), Amazon(20)
denon	Amazon(32), Akamai(5)
tivo	Limelight(23), Amazon(12), Akamai(9), Google(6)
directv	Lumen(16), Akamai(11), Amazon(11)
home	Amazon(10), nabu.casa(9)

Table 14: List of top support parties and their clients.The number of devices of a given vendor exchanging data with a given support organization is shown in the parenthesis. We exclude device vendors Vendors with only one device were excluded from the table.

Support Org	List of Clients
Amazon AWS	sonos(1699), belkin(1163), roku(618), samsung(382), wyze(287), vizio(139), google(128), wink(109), bose(92), sony(85), logitech(84), lg(73), mysa(67), tcl(61), lutron(57), amcrest(56), irobot(44), nintendo(38), ikea(37), nvidia(35), denon(32), tplink(30), philips(30), koogeek(29), yamaha(27), aurora(26), apple(26), microsoft(25), xiaomi(23), skybell(20), netgear(20), obihai(18), broadlink(18), onkyo(17), spotify(16), hikvision(14), ecobee(13), axis(13), august(13), rachio(12), leviton(12), tivo(12), directv(11), tesla(11), insignia(10), arris(10), osram(10), hp(10), home(10), sleep number(10), sharp(10), sense(10), nest(9), hunter douglas(8), canon(8), sky(8), aqara(8), first alert(7), neato(7), vocolinc(7), plex(6), sunpower(6), sleepnumber(6), enphase(6), dyson(6), humax(5), dish(4), eufy(4), nokia(4), polycom(4), wifiplug(4), insteon(4), panasonic(4), tablo(3), tado(3), vivint(3), reolink(3), lifx(3), epson(3), homey(3), caavo(3), bluesound(3), foscam(3), hubitat(3), carrier(3), grandstream(3), tuya(2), hdtv(2), rach(2), hikam(2), texas instruments(2), silicondust(2), freebox(2), firstalert(2), fdt(2), dlink(2), bryant(2), dahua(2), jbl(2), kuna(2), hunter(2), idevices(2), ihaper(2), lorex(2), phyn(2), wilife(2)
Akamai	amazon(254), apple(228), google(169), sony(160), samsung(138), sonos(131), roku(119), microsoft(71), nvidia(39), lg(39), nintendo(36), philips(25), vizio(17), directv(11), belkin(10), tivo(9), insignia(9), tcl(8), humax(7), wyze(6), sky(6), denon(5), bose(5), xiaomi(5), rainmachine(3), hikvision(3), panasonic(3), lifx(3), spotify(3), tplink(3), logitech(3), plex(2), netgear(2), onkyo(2), wink(2), tesla(2), yamaha(2), hdtv(2), cisco(2), lenovo(2), dish(2), ecobee(2)
Google	sony(204), nvidia(128), vizio(113), samsung(70), lg(61), insignia(43), onkyo(42), roku(41), amazon(39), sonos(36), philips(20), xiaomi(19), lenovo(18), jbl(17), apple(12), belkin(12), microsoft(11), bose(8), bang & olufsen(10), tivo(6), toshiba(4), panasonic(4), airtv(4), freebox(4), dlink(3), spotify(3), hikvision(3), vodafone(2), homeseer(2), tcl(2), telus(2)
Philips	amazon(257), google(8), homey(6), lg(5), hikvision(3), samsung(3), nvidia(2)
Roku	tcl(56), amazon(26), samsung(16), sharp(7), insignia(7), google(6), nvidia(5), nintendo(4), lg(2), denon(2), sony(2), spotify(2)
Fastly	google(39), amazon(33), samsung(10), sony(10), nvidia(8), roku(7), apple(6), lg(4), microsoft(4), sonos(3), belkin(2)
SmartThings	samsung(82), logitech(27), amazon(4), nvidia(4)
Limelight	sony(28), tivo(23), amazon(13), samsung(9), apple(7), roku(7), vizio(7), lg(4), sonos(4), hikvision(3), google(3), microsoft(3), nvidia(2)
Apple HomeKi	samsung(60), vizio(21), lg(6), amazon(2), roku(2)
Ayla Networks	insignia(66), hunter(20), shark(2)
Microsoft Azure	honeywell(27), samsung(14), sonos(7), amazon(6), panasonic(5), google(5), netgem(4), bang & olufsen(3), lennox(2), sony(2)
Lumen	directv(16), roku(15), amazon(11), samsung(11), apple(9), sony(4), microsoft(3), google(2), hikvision(2)
Pandora	amazon(20), sonos(19), google(6), roku(5), denon(2), logitech(2), samsung(2), sony(2)
AMP	roku(26), samsung(9), sony(4), apple(3), lg(3), nvidia(3), xiaomi(3), tivo(2), vizio(2)
Samsung	amazon(22), google(8), sony(4), roku(4), hikvision(3), lg(3), nintendo(2), philips(2), tplink(2)
CloudFlare	google(17), vizio(9), samsung(5), nvidia(4), sony(4), amazon(4), apple(2), home(2), philips(2)
Nest	lutron(15), logitech(14), apple(4), sony(4), amazon(3), lenovo(2), nvidia(2), samsung(2), sonos(2)
Heroku	logitech(8), nvidia(8), sony(7), amazon(6), samsung(3), apple(2), august(2)
Arlo	netgear(40)
TowerData	roku(12), amazon(10), lg(5), samsung(4), sony(4), tcl(2)
Synology	samsung(5), amazon(4), nintendo(3), hikvision(3), sonos(2), philips(2), nvidia(2)
StreamUnlimited	onkyo(24), bang & olufsen(2)
Tuya	belkin(4), google(4), amazon(3), shelly(3), tplink(3), xiaomi(3), samsung(2)
Ring	samsung(7), belkin(2), google(2), home(2), texas instruments(2)
Edgecast	nvidia(6), samsung(5), amazon(3), google(2)
QNAP	amazon(4), philips(3), google(2), nintendo(2), plex(2), roku(2)

D CROSS-BORDER DATA SHARING

In Figure 8, we presents the detailed cross-border data sharing statistics. Here we further split the remote endpoints within a region into first, support and third parties.

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	16.8	4.3	55.1	0.0	0.1	3.6	0.0	0.0	20.1
EA	51.8	17.8	5.9	8.4	2.3	13.3	0.1	0.0	0.4
APA	23.2	0.7	7.2	0.1	0.1	0.4	40.6	14.1	13.5

(a) Media/TV

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	84.5	0.6	6.7	0.0	0.0	0.0	0.0	0.0	8.2
EA	66.1	1.3	3.0	3.0	0.6	26.1	0.0	0.0	0.0
APA	87.1	0.1	0.5	0.0	0.0	0.0	10.9	0.2	1.3

(b) Voice Assistant

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	90.3	7.9	1.3	0.0	0.5	0.0	0.0	0.0	0.0
EA	32.4	4.1	42.3	15.8	0.2	5.1	0.0	0.0	0.0
APA	0.0	0.6	1.2	0.0	0.0	0.0	98.2	0.0	0.0

(c) Surveillance

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	73.9	2.6	3.2	6.5	0.0	13.8	0.0	0.0	0.0
EA	61.4	0.0	7.7	22.8	0.1	2.3	5.6	0.0	0.0
APA	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0

(d) Home Automation

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	93.2	3.0	0.8	0.1	0.0	0.0	2.7	0.1	0.0
EA	43.7	42.9	0.1	11.7	1.1	0.1	0.4	0.0	0.0
APA	0.1	0.8	54.2	38.9	0.2	0.1	0.3	0.0	5.6

(e) Home Appliance

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	67.4	4.7	23.4	1.8	0.1	2.4	0.2	0.0	0.1
EA	26.3	3.5	22.7	3.0	13.7	30.2	0.5	0.0	0.0
APA	61.9	3.0	1.0	1.4	0.0	0.2	3.7	0.0	28.9

(f) Game Console

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	97.5	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EA	5.8	83.2	9.7	0.1	1.0	0.0	0.2	0.0	0.0
APA	-	-	-	-	-	-	-	-	-

(g) Generic IoT

	NSA			EA			APA		
	F	S	T	F	S	T	F	S	T
NSA	35.3	11.0	53.6	0.0	0.0	0.0	0.1	0.0	0.0
EA	36.6	12.3	3.4	0.2	0.0	28.5	0.0	0.0	19.0
APA	10.8	0.0	5.3	0.0	0.0	0.0	0.1	0.0	83.9

(h) Work Appliance

Figure 8: Percentage of outgoing data flow to ANY party crossing borders. Row: User Location, Column: Destination location. NSA: North & South America, EA: Europe & Asia, APA: Asia Pacific and Australia, F: First party, S: Support party, T: Third party

E LONGITUDINAL ANALYSIS

The dataset includes network traffic from smart home devices between April 2019 and July 2022. We divide the dataset into two-time intervals and see if any previous distributions differ between these two intervals. The first segment of the dataset includes network flow data collected between April 2019 and December 2020. The second portion of the dataset contains data collected between January 2021 and July 2022. We then repeat our analysis on each of these two subsets separately.

Table 15 shows the temporal evolution of the distribution of various types of domains connected by different categories of smart home devices. The terms 'Avg.' and 'SD' refer to the average and standard deviation of the total number of domains contacted by each device during the given time frames. The top row of each category indicates distribution before January 2021, while the bottom row indicates inclusive distribution that occurs after January 2021. Different categories of devices exhibit a variety of behaviors. For instance, the average number of contacted first, support, and third-party domains increases in the `Media/TV`, `Home Automation`, and `Work Appliance` categories. This indicates a growing practice of exchanging more data with more third parties. Another aspect could be that smart home devices have expanded services over time, requiring them to contact more third parties. The average number of first-party domains for `Surveillance` devices has decreased over time, while the average number of support-party and third-party domains has grown. This indicates that camera device users use more third-party applications to operate their cameras. For instance, several camera manufacturers (such as Google Nest, Samsung, Sony, Vizio, Blink, LIFX, Logitech, Amcrest) manufacture cameras that are compatible with Apple HomeKit. Tuya Smart is another popular application that can be used to control a variety of popular home devices. In general, across most categories, the average number of support parties contacted by each device has grown, as seen in the Table 15. This suggests that the supporting back-end infrastructure for IoT devices is expanding over time. A detailed breakdown of temporal variation in the number of endpoints of various parties across various region is shown in Table 16 in Appendix E.

The average number of distant endpoints reached by each device varies by area, as seen in Table 16. The average number of third-party domains contacted drops for most device categories when the device is placed in the Asia Pacific and Australia region, whereas it rises when the device is located in any of the other two regions. We also see a rise in the average number of support-party domains contacted by each device over time for several types of devices when the device is located in North and South America, as well as European and African regions. This suggests that North America and Europe are leading the way in the development of the infrastructure necessary to support the rapid growth of IoT devices.

Table 15: Temporal variation in the number of endpoints. Avg. and SD stand for the average and standard deviation of the number of domains each device contacted. The *top row* in each category reflects the number of connections formed before January 2021, while the *bottom row* shows the number of connections made after January 2021.

Category	First party Avg. (SD)	Support party Avg. (SD)	Third party Avg. (SD)
Media/TV	1.98(2.47)	0.85(1.26)	3.21(7.01)
	2.20(2.85) ↑	1.14(1.76) ↑	4.01(8.25) ↑
Voice Assistant	4.45(2.19)	0.22(0.62)	1.00(2.37)
	4.31(2.25) ↓	0.28(1.08) ↑	0.87(2.22) ↓
Surveillance	1.05(0.63)	0.51(0.74)	0.81(1.66)
	0.98(0.71) ↓	0.74(1.01) ↑	1.49(1.76) ↑
Home Automation	0.91(0.68)	0.39(0.73)	0.47(2.53)
	0.97(1.12) ↑	0.67(1.16) ↑	0.76(3.05) ↑
Home Appliance	1.23(1.44)	0.62(1.12)	0.95(6.28)
	0.83(0.58) ↓	0.50(0.67) ↓	0.75(1.48) ↓
Game Console	1.82(3.15)	0.97(1.83)	3.15(7.18)
	1.72(2.92) ↓	0.68(1.20) ↓	1.23(2.27) ↓
Generic IoT	0.78(1.76)	0.83(0.85)	0.93(2.69)
	0.60(0.89) ↓	0.60(0.55) ↓	0.40(0.89) ↓
Work Appliance	0.97(0.63)	0.12(0.37)	0.55(1.10)
	1.00(0.62) ↑	0.67(1.73) ↑	0.96(2.36) ↑

Table 16: Temporal change in the average number of different type of domains communicated by a single device. First: Average number of first party domains, Support: Average number of support party domains, Third: Average number of first party domains, '-' Not applicable (no device seen).

Category	North & South America			Europe & Africa			Asia Pacific & Australia		
	First	Support	Third	First	Support	Third	First	Support	Third
Media/TV	1.99	0.85	3.28	1.85	0.80	2.63	2.01	0.87	3.00
	2.03 ↑	1.17 ↑	4.00 ↑	2.73 ↑	1.25 ↑	4.73 ↑	1.93 ↓	0.57 ↓	1.90 ↓
Voice Assistant	4.43	0.19	0.93	4.48	0.29	1.12	3.96	0.11	0.92
	4.23 ↓	0.23 ↑	0.82 ↓	4.64 ↑	0.48 ↑	1.13 ↑	3.73 ↓	0.07 ↓	0.13 ↓
Surveillance	1.08	0.50	0.81	0.94	0.52	0.88	0.94	0.41	0.53
	0.91 ↓	0.66 ↑	1.49 ↑	1.32 ↑	1.14 ↑	1.68 ↑	1.57 ↑	1.14 ↑	0.86 ↑
Home Automation	0.84	0.43	0.46	1.11	0.28	0.53	1.12	0.21	0.35
	0.97 ↑	0.67 ↑	0.76 ↑	0.96 ↓	0.78 ↑	0.96 ↑	1.00 ↓	0.30 ↑	0.09 ↓
Home Appliance	1.21	0.61	0.88	0.88	0.47	0.06	1.12	0.50	2.00
	0.88 ↓	0.38 ↓	0.50 ↓	0.75 ↓	0.75 ↑	1.25 ↑	-	-	-
Game Console	1.83	0.89	3.15	1.48	1.07	3.24	2.46	1.11	2.26
	1.94 ↑	0.75 ↓	1.44 ↓	1.00 ↓	0.45 ↓	0.55 ↓	-	-	-
Generic IoT	0.86	0.80	0.80	0.44	0.78	0.78	-	-	-
	0.25 ↓	0.75 ↓	0.00 ↓	2.00 ↑	0.00 ↓	2.00 ↑	-	-	-
Work Appliance	0.97	0.12	0.56	0.96	0.10	0.42	0.91	0.00	0.55
	1.05 ↑	0.75 ↑	1.20 ↑	0.75 ↓	0.50 ↑	0.50 ↑	1.00 ↑	0.33 ↑	0.00 ↓

Figure 9: Fine-grained temporal variation in the average number of endpoints across various geographic regions. Here, we only presents the analysis results for top three categories.

F DISTRIBUTION OF ENDPOINTS TYPE

F.1 Addressing the issue of the imbalanced device dataset

Here in Table 17, we present the analysis result of subsection 6.1 in tabular form. In this table, we add the normalized value for each outcome to better compare the distribution across categories.

Table 17: The distribution of endpoints contacted by smart devices. '#Dom' column shows the number of domains accessed by all the devices in each category. '#Avg.' represent the average of the number of domains contacted by each device. '↑Avg.' represent the average of the volume of upstream data shared by devices per second. The 'Nor' value represents the normalized values between 0 and 1 across all devices.

Category	First party			Support party			Third party		
	#Domain(Nor)	#Avg.(Nor)	↑Avg.(Nor)	#Domain(Nor)	#Avg.(Nor)	↑Avg.(Nor)	#Domain(Nor)	#Avg.(Nor)	↑Avg.(Nor)
Media/TV	268(0.36)	1.99(0.15)	449(0.01)	187(0.31)	0.86(0.19)	491(0.01)	1865(0.44)	3.26(0.29)	652(0.02)
Home Automation	170(0.23)	0.91(0.07)	2355(0.03)	137(0.23)	0.41(0.09)	11055(0.15)	758(0.18)	0.49(0.04)	559(0.01)
Voice Assistant	117(0.16)	4.45(0.34)	211(0.00)	97(0.16)	0.22(0.05)	494(0.01)	642(0.15)	0.99(0.09)	283(0.01)
Surveillance	70(0.09)	1.04(0.08)	64293(0.88)	61(0.10)	0.54(0.12)	53696(0.73)	255(0.06)	0.92(0.08)	34172(0.91)
Game Console	54(0.07)	1.82(0.14)	901(0.01)	63(0.10)	0.96(0.21)	3293(0.04)	531(0.12)	3.03(0.27)	558(0.01)
Work Appliance	18(0.02)	0.97(0.07)	394(0.01)	19(0.03)	0.15(0.03)	867(0.01)	87(0.02)	0.57(0.05)	778(0.02)
Home Appliance	32(0.04)	1.20(0.09)	3769(0.05)	26(0.04)	0.61(0.13)	844(0.01)	116(0.03)	0.94(0.08)	241(0.01)
Generic IoT	24(0.03)	0.76(0.06)	637(0.01)	14(0.02)	0.80(0.18)	2527(0.03)	33(0.01)	0.88(0.08)	412(0.01)

We concur with the notion that the dataset is imbalanced regarding the number of devices across categories. This problem appears to have the potential to introduce biases into the outcome shown in Figure 3 (e.g., more domains for media TV, home automation, voice assistance). We present the average number of domains a device communicates with and average volume of up-stream data to alleviate that issue. To illustrate that our analysis is bias-free, we first randomly sample 3,000 devices from the top 3 categories (in terms of number of devices) and do the same analysis. We do the same thing three times (with three different seeds) and take the mean of analysis results. Then we compare the mean of sampled analysis result with the results for all the devices in Figure 10. We see similar distribution in both the cases.

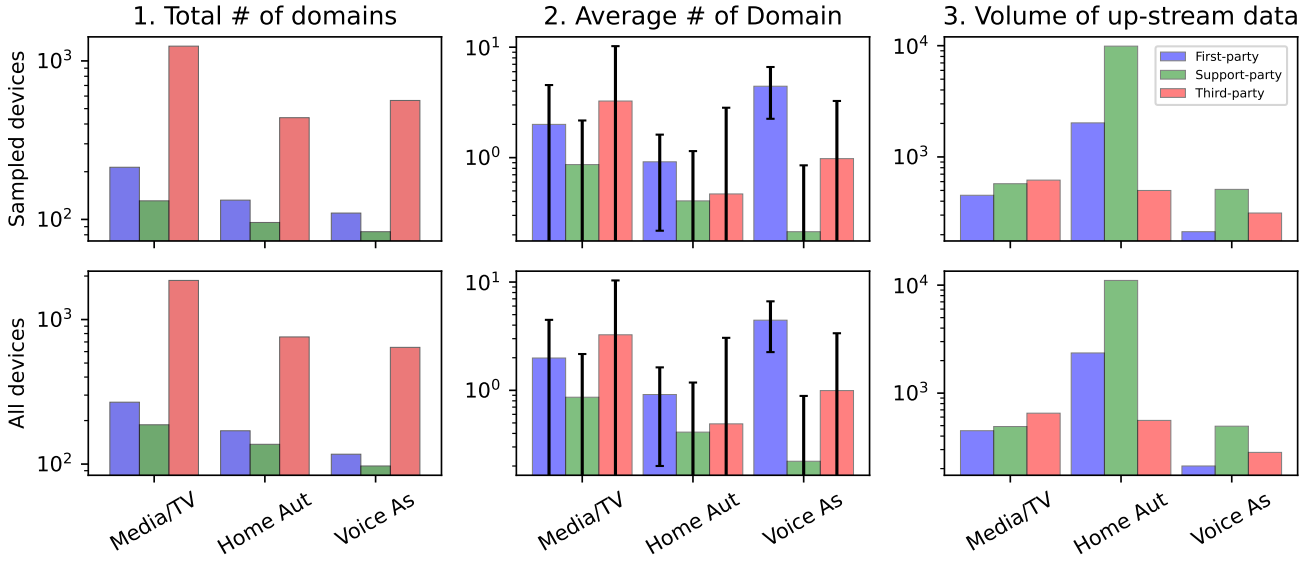


Figure 10: The distribution of contacted endpoints and upstream data volume. First row presents analysis mean outcomes on sampled devices with `N = 3,000`. Last row shows the distributions on all devices. Here we only show analysis results for top 3 categories.

We apply the same comparison methodology across all device categories. However, due to the limited number of devices in certain categories, we utilize a smaller sample size (N=50). Upon examination, we observe a nearly identical distribution with a slight variance. The findings are illustrated in Figure 11.

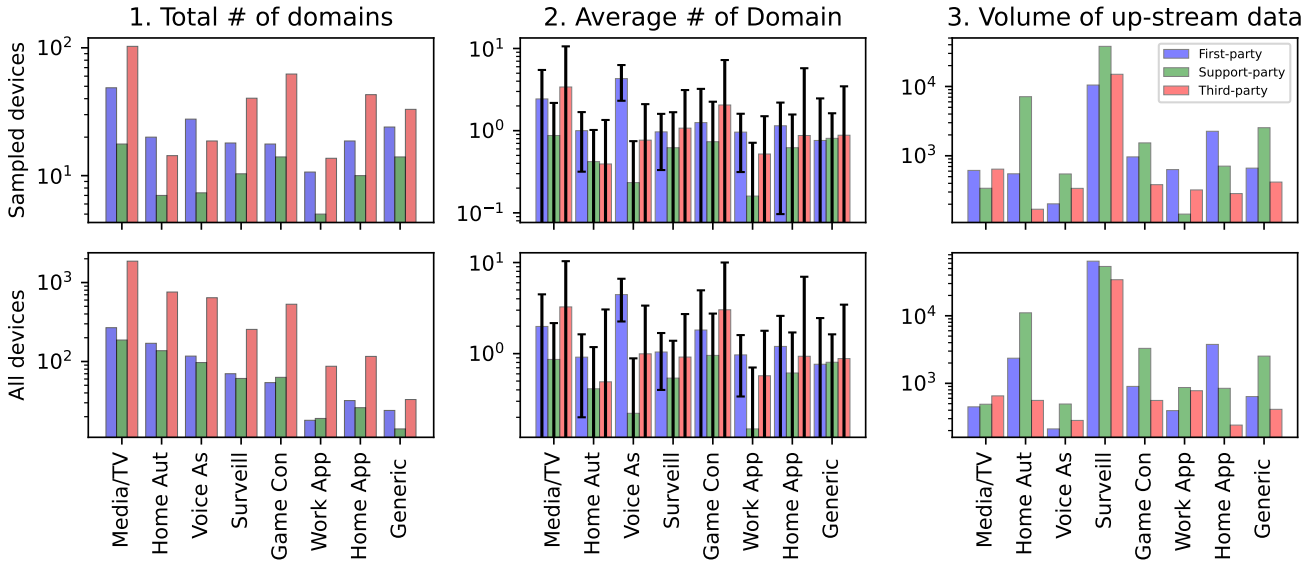


Figure 11: The distribution of contacted endpoints and upstream data volume. First row presents analysis mean outcomes on sampled devices with `N = 50`. Last row shows the distributions on all devices.

Table 6: Average number of different type of domains communicated by each device. 'First': Average number of first-party domains, 'Support': Average number of support-party domains, 'Third': Average number of third-party domains, '–': No device seen within that region. Numbers inside the parenthesis represent the normalized values between 0 and 1 across all devices.

Category	North & South America			Europe & Africa			Asia Pacific & Australia		
	First	Support	Third	First	Support	Third	First	Support	Third
Media/TV	2.00(0.05)	0.87(0.07)	3.32(0.07)	1.91(0.05)	0.82(0.07)	2.78(0.07)	1.98(0.05)	0.85(0.07)	2.91(0.07)
Home Automation	0.85(0.02)	0.44(0.04)	0.48(0.04)	1.11(0.03)	0.30(0.02)	0.55(0.02)	1.11(0.03)	0.22(0.02)	0.33(0.02)
Voice Assistant	4.42(0.12)	0.20(0.02)	0.92(0.02)	4.51(0.12)	0.31(0.03)	1.13(0.03)	3.97(0.10)	0.11(0.01)	0.86(0.01)
Surveillance	1.05(0.03)	0.53(0.04)	0.93(0.04)	0.98(0.03)	0.56(0.05)	0.98(0.05)	1.05(0.03)	0.54(0.04)	0.59(0.04)
Game Console	1.84(0.05)	0.89(0.07)	3.04(0.07)	1.44(0.04)	1.02(0.08)	3.02(0.08)	2.46(0.06)	1.11(0.09)	2.26(0.09)
Work Appliance	0.97(0.03)	0.15(0.01)	0.60(0.01)	0.95(0.02)	0.12(0.01)	0.42(0.01)	0.93(0.02)	0.07(0.01)	0.43(0.01)
Home Appliance	1.19(0.03)	0.60(0.05)	0.85(0.05)	0.86(0.02)	0.52(0.04)	0.29(0.04)	1.12(0.03)	0.50(0.04)	2.00(0.04)
Generic IoT	0.79(0.02)	0.79(0.06)	0.72(0.06)	0.60(0.02)	0.70(0.06)	0.90(0.06)	–	–	–