How IoT Re-using Threatens Your Sensitive Data: Exploring the User-Data Disposal in Used IoT Devices

Peiyu Liu[†‡], Shouling Ji[†](✉), Lirong Fu[†], Kangjie Lu[§],
Xuhong Zhang[†], Jingchang Qin[†], Wenhai Wang[†‡], Wenzhi Chen[†](✉)
[†]Zhejiang University, [‡]Zhejiang University NGICS Platform, [§]University of Minnesota
{liupeiyu, sji, fulirong007}@zju.edu.cn, kjlu@umn.edu, {zhangxuhong, qjc1999, zdzzlab, chenwz}@zju.edu.cn

***Abstract*—With the rapid technology evolution of the Internet of Things (IoT) and increasing user needs, IoT device re-using becomes more and more common nowadays. For instance, more than 300,000 used IoT devices are selling on Craigslist. During IoT re-using, sensitive data such as credentials and biometrics residing in these devices may face the risk of leakage if a user fails properly dispose of the data. Thus, a critical security concern is raised: do (or can) users properly dispose of the sensitive data in used IoT? To the best of our knowledge, it is still an unexplored problem that desires a systematic study.**

In this paper, we perform the first in-depth investigation on the user-data disposal of used IoT devices. Our investigation integrates multiple research methods to explore the status quo and the root causes of the user-data leakages with used IoT devices. First, we conduct a user study to investigate the user awareness and understanding of data disposal. Then, we conduct a large-scale analysis on 4,749 IoT firmware images to investigate user-data collection. Finally, we conduct a comprehensive empirical evaluation on 33 IoT devices to investigate the effectiveness of existing data disposal methods.

Through the systematical investigation, we discover that IoT devices collect more sensitive data than users expect. Specifically, we detect 121,984 sensitive data collections in the tested firmware. Moreover, users usually do not or even cannot properly dispose of the sensitive data. Worse, due to the inherent characteristics of storage chips, 13.2% of the investigated firmware perform "shallow" deletion, which may allow adversaries to obtain sensitive data after data disposal. Given the large-scale IoT re-using, such leakage would cause a broad impact. We have reported our findings to world-leading companies. We hope our findings raise awareness of the failures of user-data disposal with IoT devices and promote the protection of users' sensitive data in IoT devices.

I. Introduction

With the development of the Internet of Things (IoT), IoT devices are projected to amount to 30.9 billion worldwide by 2025 [8]. Meanwhile, along with the frequent upgrading of IoT devices, users periodically buy new IoT devices and resell/discard old ones for environmental protection or economic benefits. Specifically, in 2019, the world generated 53.6 million metric tons used IoT devices [6]. During the usage, an IoT device may collect and carry various categories of sensitive information, e.g., user portrait, third-party accounts, etc., to maintain normal utilities, improve service quality, and achieve many other goals [32], [44]. For example, a smart camera may require the password of an FTP service to upload surveillance video for backup. Thus, if a user does not properly dispose of the sensitive data in a used IoT device before reselling/discarding it, the user may face the risk of leaking sensitive data.

The data leakage caused by improper disposal of used IoT devices may lead to severe consequences. Prior works already show that adversaries have incentives to obtain the residual sensitive data to launch various attacks [3], [7], [29], [34], [35], [37]. For instance, by obtaining a user's email address and web browsing history from a used IoT device, an adversary could launch a phishing attack by emailing a fake login page of a website that the user is interested in. Therefore, proper data disposal of used IoT devices is important to prevent users from being exposed to potential security risks.

In recent years, researchers pointed out that IoT vendors should facilitate customers to remove sensitive data from used IoT devices [19]. Additionally, many communities call for users to erase personal information after their usages. For instance, Hong Kong's privacy commissioner provides a suggestion for IoT users—"before you resell/discard your IoT devices, purge the user account information and other personal data stored in the IoT devices" [13]. Moreover, to decrease the security risks of data leakage brought by improper disposal, prior approaches suggest that compared to storing sensitive data locally, user's data should be transmitted to the cloud of things (the integration of cloud computing and the internet of things) [16], [39]. Although existing efforts attempted to decrease the security risks of data leakage, a critical question remains—do (or can) users actually properly dispose of their data in used IoT devices?

To the best of our knowledge, the user-data disposal problem with used IoT devices has not yet been systematically investigated. For the moment, a comprehensive and in-depth study of this problem is highly demanded to help users and related parties understand the potential risks of data leakage. Intuitively, to address the above problem, we need to answer the following three research questions: **RQ1**: *which kinds of sensitive data reside in used IoT devices?* **RQ2**: *which methods can be used to dispose of sensitive data?* and **RQ3**: *are existing*

✉ Shouling Ji and Wenzhi Chen are co-corresponding authors.

disposal methods effective in erasing the sensitive data?

To dive into the details of these research problems, in this paper, we integrate multiple research methods. First, we conduct a user study (with 277 users) to investigate the three research questions from a user's perspective (see §III-A). The user study enables us to understand the user awareness of the data disposal methods and the security risks of data leakages with used IoT devices. Our user study shows that (1) the re-using of IoT devices is quite common. (2) During IoT re-using, many users lack the awareness and technical knowledge to dispose of their sensitive data properly. Specifically, even though 80.2% of the users are concerned about the leakage of their sensitive data, 51.1% of the users do not erase personal data before the disposal of used IoT devices (see §IV). This finding by itself shows that the current security awareness and disposal are inadequate.

Motivated by this user study, we then try to figure out whether the reality is consistent with the users' expectations. Therefore, we conduct a large-scale analysis on 4,749 IoT firmware from 11 worldwide leading vendors to investigate **RQ1** (see §III-B). To achieve this, we design a system to perform sensitive data analysis on firmware images. This system allows us to provide a real-world view of user-data collection by IoT devices. After that, to investigate **RQ2** and **RQ3**, we conduct an empirical study on various types of IoT devices (see §III-C). In this study, we first explore the disposal methods provided by different IoT devices. Then, we perform forensic analysis on real-world IoT devices to understand the effectiveness of each disposal method. Specifically, according to our user study, network equipment, such as routers and access points, is the most common IoT device. Thus, we investigated more network equipment. Meanwhile, to enable a more comprehensive understanding, we also investigated other device categories. Our evaluation allows us to paint an unprecedented picture of the real-world user-data disposal of used IoT devices, which reveals that (1) IoT devices hold more sensitive data than users expect. Specifically, our system detects 121,984 sensitive data collections in the tested firmware. Besides, 63.8% of the sensitive data is stored in plain text in tested IoT devices, which is opposite to user expectation (see §V). (2) The used data disposal methods (including data overwriting and device resetting) oftentimes cannot effectively erase sensitive data (see §VI). Indeed, 9 out of the 33 tested devices and 13.2% of the tested firmware face the risk of ineffective data disposal.

Our findings show that the current data disposal of used IoT devices is insufficient. Due to the lack of user awareness, inadequate data protection, and hardware characteristics, users are highly likely to suffer from data leakage risks. We have reported our findings to IoT vendors. Three vendors indicate that this is an important issue. We are in contact with these vendors to alleviate the potential leakage of users' data. Meanwhile, we also reported our findings to local companies that use IoT devices. Specifically, one world-leading industry control company acknowledged our report and invited us to help them identify risks in the devices deployed in the company. We hope our study would encourage more researchers, policy makers, manufacturers, and users to carefully protect sensitive data in used IoT devices with effective disposal methods.

In summary, our study makes the following contributions.

- We integrate multiple research methods to conduct the first systematical investigation on the user-data disposal of used IoT devices, which uncovers a serious (but without sufficient attention) risk of high-volume data leakage in the IoT ecosystem.
- We propose a system to detect user-data collections in IoT firmware. The experimental results on 4,749 IoT firmware images show that IoT devices collect much more sensitive data (e.g., home WiFi information, email account passwords, and browsing history) than users expect. Besides, Our evaluations on 33 representative devices indicate that 63.8% of the sensitive data in used IoT devices is stored in plain text.
- Our findings reveal the root causes of privacy leakage in used IoT devices are multilayered. First, many users lack the awareness and technical knowledge for data disposal. Second, IoT vendors do not provide adequate data protections, such as data encryption. Third, the inherent characteristics of storage chips make the "shallow" deletion cannot effectively erase user data. To alleviate this situation, we propose multiple suggestions for both users and IoT vendors.

II. Problem Statement

To better understand the problem studied in this paper, we first elaborate on the real-world re-using of IoT devices. Then, we discuss how user data leaks during IoT re-using and the severe consequences of such leakage.

The re-using of IoT devices. In the era of "interconnection of all things", more and more IoT devices are deployed in people's daily life. The re-using of IoT devices becomes common, happening every day. For instance, in 2020, Total Green Recycling [15] recycled more than 300,000 used IoT devices, including routers, printers, and smart TVs, in Australia.

Generally, the old IoT devices are re-used mainly in two ways. (1) Users often resell or lend IoT devices. Increasing online trading platforms boost the re-using of IoT devices, which are sold or leased on popular e-commerce sites (such as Craigslist [4], Amazon [1], and eBay [5]). For instance, more than 300,000 used IoT devices are selling on Craigslist. One can purchase used IoT devices on these online platforms conveniently. (2) Users may discard their used IoT devices. Anyone can collect discarded devices from dustbins. Moreover, governments and communities have built up more and more recycling collection sites to protect the environment and save energy, gathering many IoT devices discarded by users. Many collected IoT devices could be resold after viably repair [14]. In summary, with the concept of environmental protection deeply rooted in people's minds, more and more users become willing to participate in the re-using of IoT devices. An IoT device might be used by many users during device re-using.

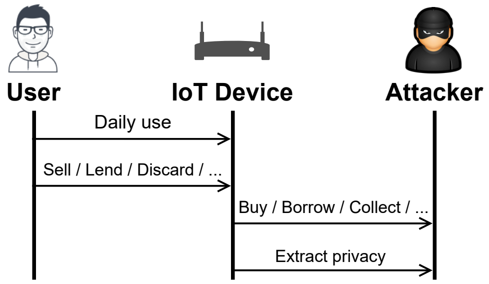

Fig. 1: The threat model of our measurement study.

The sensitive data residing in devices might be obtained by others, including an adversary.

During the usage, compared with the data (i.e., visible files) on PCs and mobiles, the data stored in IoT devices is invisible to users. Thus, many users are unaware of the data remanence concerns in IoT devices. Besides, there exist many tools that can be used to perform secure data encryption, detection, deletion, and recovery on PCs and mobiles. By contrast, due to the limited computational capabilities, there is still a lack of tools for protecting data on IoT devices. Therefore, while re-using IoT devices boosts sustainable development, it also brings new security risks to users.

User-data leakage and consequences. Briefly, as shown in Figure 1, a user's sensitive data is collected and stored in an IoT device. With the re-using of this device, it may be sold, lent, or discarded by the user. Then, an adversary may collect this used device by purchasing, borrowing, or picking it up from the corresponding recycling collection site. Once obtaining this device, the adversary can leverage various methods to dump the user's sensitive data from the IoT device. Such leakage caused by the large-scale IoT re-using would have a broad impact. For example, the adversary can collect a large number of third-party user accounts from used IoT devices to conduct credential stuffing attacks to control more critical accounts, such as financial accounts. Moreover, the adversary can conduct social engineering attacks through the trading platform, such as obtaining the user's home address when purchasing used IoT through Craigslist [4]. Then, the adversary can further use the user data obtained from the used IoT to conduct severe attacks. For example, suppose that the adversary gets the physical security setting in the user's house, such as monitoring areas. He/she may break into the house without being monitored.

III. Methodology

In this paper, we conduct the first in-depth study of the user-data disposal problem with used IoT devices. To provide the most comprehensive view of this problem to date, we integrate multiple research methods and investigate the following three key research questions: **RQ1**: *which kinds of sensitive data reside in used IoT devices?* **RQ2**: *which methods can be used to dispose of sensitive data?* and **RQ3**: *are existing disposal methods effective in erasing the sensitive data?*

In the rest of this section, we first detail our investigation methods used in this paper (§III-A, §III-B, and §III-C). Then, we elaborate on our dataset (§III-D). Finally, we explain our ethical consideration (§III-E).

A. A User Study

The security risks of user-data leakage in used IoT devices depend, to a large extent, on user awareness and their disposal methods. When a user knows which categories of sensitive data may potentially leak and how to dispose of them in IoT devices properly, an attacker may have less chance to collect the user's sensitive data and launch consequent attacks. By contrast, data leakages likely happen when the user is unaware of the sensitive data or how to dispose of data. Thus, we first perform a user study to understand the user awareness of sensitive data and data disposal methods of used IoT devices.

Method. In this user study, we collect the following information. (1) Personal information (e.g., ages, professions, etc.), for understanding different user awareness from various perspectives. (2) User experience information (including how to deal with used IoT devices, dispose of sensitive data before device re-using, and set/update user passwords), for understanding how the surveyed users use IoT devices in their daily lives. (3) User understandings (involving what sensitive data users think may store in a used IoT device, do users trust the disposal methods they use, and so on), for learning how users understand the security risks and protection methods of IoT devices and their expectations.

We then design and send online questionnaires[1] to diverse participants to obtain user replies. Specifically, our online questionnaire was sent to our colleagues and classmates and further spread by them. To ensure a balanced assessment, the surveyed users cover 321 participants with different professions (including 39.2% of non-CS background), ages, educational backgrounds, genders, and regions (the distribution of the surveyed users is deferred to Appendix §A-A). To obtain high-qualified and unbiased user replies and avoid invalid questionnaires that a user randomly returns, we designed a control question with a provided answer in it. If the user provides a wrong answer, we exclude all his/her replies. In total, we received 277 valid questionnaires (the control question excludes the rest questionnaires).

B. A System for Analyzing Sensitive Data

After the user study, we investigate whether the reality is consistent with the users' expectations. First, we want to answer **RQ1**, i.e., which kinds of sensitive data reside in used IoT devices? Intuitively, one may collect used IoT devices from users and detect sensitive data in these devices. However, this method has several inherent limitations. First, this method is insufficient for a large-scale study since it is impractical to collect a large number of user devices. Second, it may introduce bias to the analysis result since, under different user

[1] https://docs.google.com/document/d/1dKcWdU6fG4CvS92qpXMOkaSekObLL1QhqgzLxeJH9us.

```
# /etc/config/ddns
config global 'ddns'
    option username 'AA@BB.com'
    option password 'CCDDEE'
    option domain 'FF.com'
```

(a)

```
# /usr/lib/lua/luci/controller/api/xqnetwork.lua
function ddnsEdit()
    local XQDDNS = require("xiaoqiang.module.XQDDNS")
    local domain = LuciHttp.formvalue("domain")
    local username = LuciHttp.formvalue("username")
    local password = LuciHttp.formvalue("password")
    XQDDNS.editDdns(username, password, domain)
```

(b)

Fig. 2: An example of (a) sensitive data in a router and (b) the corresponding data collection code in the firmware (more details are deferred to §A-B).

preferences, the same IoT devices may store different types of sensitive data. Thus, we design a new device-free method to understand **RQ1** comprehensively.

Intuition. The intuition of our design is that the store behavior of sensitive data always accompanies a data collection code logic in IoT firmware. For example, the user's DDNS service information residing in a router (shown in Figure 2a) is collected by the code of the router's firmware (shown in Figure 2b). Therefore, we can translate the problem of "detecting sensitive data in IoT devices" to "detecting user-data collection in IoT firmware". This method has multiple advantages. 1) It does not introduce ethical concerns since it works without user devices or personal data. 2) This method is scalable since one can collect a large amount of IoT firmware images. 3) This method can provide a comprehensive view since it reports all the user-data collection behavior in IoT firmware.

Challenges. Although our method seems straightforward for the motivating example, applying it to real-world IoT firmware is still challenging. First, accurately identifying user-data collection codes, among a large number of other codes in an unpacked IoT firmware image, is challenging (**C1**). For example, the firmware discussed in Figure 2 contains more than 1,200 files and over 120,000 code lines. Second, it is challenging to understand the semantics of each vendor-defined API (such as "XQDDNS.editDdns" in Figure 2b) without domain knowledge (**C2**).

Method. We develop a new sensitive data analysis system by addressing the above challenges. For **C1**, the key point of our system is detecting the pairs of source and sink APIs (such as "LuciHttp.formvalue" and "XQDDNS.editDdns" in Figure 2) that reveal obtaining and storing user inputs. However, we still face **C2** when collecting the calls of the source/sink APIs (SAPIs). To address this challenge, we propose a two-layer API inferring method. First, many vendor-defined APIs are implemented by encapsulating library APIs. We can infer them based on the library APIs they encapsulate. Second, vendors also implement vendor-defined APIs in binary libraries. Since they are closed-source, we cannot identify these APIs by analyzing encapsulating APIs. Fortunately, we find that vendors tend to develop vendor-defined APIs for sink/source usage rather than other usages of the source/sink data. Thus, we can collect all the uses of a known source/sink data. Then, if one use is not a predefined known common API, we infer it is a sink/source usage (the predefined API list and more details are deferred to §A-C).

Based on the above methods, our system takes as input a firmware image and reports user-data collections. It first unpacks the firmware image with `binwalk` [2]. It then analyzes Lua files statically to collect the calls of predefined SAPIs and identifies vendor-defined SAPIs by analyzing function encapsulation. Once the system identifies a vendor-defined SAPI, we also collect its calls. Then, for each collected source API call c, we further collect the uses of its source data. Suppose that we find a use u is a call of a known sink API. In this case, we report a user-data collection since a source API call (c) pairs with a sink API call (u). Suppose none use is a known sink API call. We try to find unknown API calls in the uses and report a user-data collection once we find one. We do the same analysis for each sink API call.

Note that, according to the analysis, more than 76.0% of the tested firmware images use Lua (exists in source code form) to process user inputs. Thus, our system only analyzes Lua for now. Supporting other programming languages, such as PHP and ASP.Net, could be a future research direction. The experiment results reveal that our system enables a conservative analysis and finds high-volume data leakage.

C. An Empirical Study

The firmware analysis achieves good performance when investigating **RQ1** (provided in §V). Based on the analysis results, we conduct an empirical study on real-world IoT devices to investigate **RQ2** and **RQ3**. In this study, we purchase several representative IoT devices (detailed in §III-D). We set up each device and enable as many utilities supported by the device as possible according to its instructions manual and related online discussions. Note that we use magic strings when configuring each utility. For instance, we use "Oakland23WifiSsid" and "Oakland23WifiPwd" as a router's WiFi SSID and password, respectively. Then, we conduct a forensic analysis to discover the magic strings in each device before and after we perform data disposal. Suppose that a magic string exists before data disposal and disappears after disposal. In this case, we report that the data disposal method effectively erases a sensitive item. Otherwise, we report a data disposal method as ineffective if a magic string still exists after the disposal. Three authors perform the empirical study independently to mitigate the possible bias caused by manual analysis. Finally, we report the unanimous result. Next, we introduce the forensic analysis in this study.

Forensic analysis. We perform both dynamic and static analysis when conducting the forensic analysis on a target device. For dynamic analysis, we try to discover the magic strings through the access interfaces of the device. For instance,

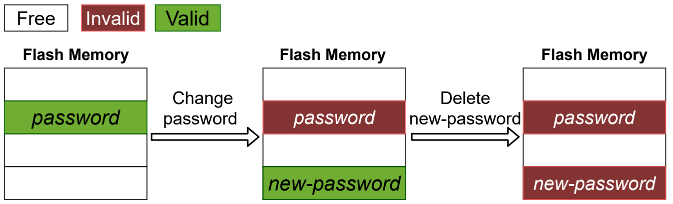

Fig. 3: An example of wear-leveling write. When *password* is changed or deleted, the corresponding storage unit of it will be marked as "invalid" but *password* still exists in the flash memory.

we can access the terminal of a device through its debug interface, such as telnet and ssh. By interacting with the terminal, we can scan the content of the files stored on the device. For static analysis, we try to discover the magic strings stored in the storage units of the device. For instance, we can extract the firmware image from the flash chip by leveraging a flash programmer [42]. Then, we can scan the content stored in the firmware image. Due to space limitations, we cannot present each analysis method in detail. Instead, in the rest of this section, we briefly summarize a technique we used in the static analysis, which allows us to discover magic strings after data disposal.

Insight. Currently, most IoT devices use a flash chip as the storage component [43], [27]. For a rewriting request, traditional in-place write always writes the same storage unit. However, the total number of written times of a flash storage unit is limited (typically 100K times). Thus, to extend the service life of a flash chip, IoT file systems widely adopt wear-leveling write [43], [27], which can balance the writing times of every storage unit. Figure 3 shows an example of wear-leveling write. Given a rewriting request, the wear-leveling write will find other free units to store the new data and then mark the old unit as invalid. Due to the wear-leveling write strategy used in flash chips, data overwriting and soft/hard resetting may not really erase the original sensitive data. The "deleted" sensitive data may still exist in the flash chip. Therefore, we may discover the "deleted" sensitive data by scanning the storage units in the flash chip.

To discover the "deleted" sensitive data in an invalid storage unit, we implement a customized filesystem scanner based on `jefferson` [9]. Specifically, it first scans the whole file system to find all storage units. Then, it collects file content from every unit no matter it is valid or not. Finally, it detects sensitive data by searching magic strings.

D. Two New Datasets

This paper aims to perform a systematical investigation on the data disposal of different IoT devices. Therefore, we consider covering more IoT devices from various vendors to construct a representative dataset.

IoT firmware (Dataset-1). When investigating **RQ1** through sensitive data analysis (§III-B), we construct a large-scale dataset by collecting IoT firmware images from online resources, such as official websites, FTP sites, and GitHub repositories. Table I summaries the firmware dataset, which includes 4,749 IoT firmware images from 11 vendors. According to our user study, network equipment, such as routers and access points, is the most common IoT device. Specifically, 86.4% of the users have used network equipment. Thus, we investigated more network equipment. Meanwhile, to enable a more comprehensive understanding, we also collected other firmware images with various device categories, such as surveillance, Bitcoin miner, and System-on-Module. This dataset helps us understand the status quo of data disposal along with real-world IoT devices.

TABLE I: Summary of the investigated firmware.

Vendor	Device Type	# Firmware Image
Avalon	Miner	278
Fastcom	Router, Others	11
GL.iNet	Edge Computing, Gateway, Router	117
MiCasaVerde	Gateway	1
OpenWRT	Access Point, Router, Others	2658
Phicomm	Router	10
ROOter	Router	567
TP-Link	Router, Switch, Others	1,043
Trendnet	Router, Surveillance	28
Xiaomi	Router, Others	20
8devices	System-on-Module	16
Total		**4,749**

IoT devices (Dataset-2). In the empirical study (§III-C), to investigate **RQ2** and **RQ3**, we analyze 33 real-world IoT devices. As shown in Table II, this dataset also covers various categories and vendors. Indeed, 19 (57.6%) devices are from 10 worldwide leading vendors [30]. The remaining 14 devices are also popular on online second-hand trading platforms.

E. Ethical Considerations

In this paper, we conduct a user study to understand how users dispose of the used IoT devices. Moreover, we conduct sensitive data analysis on IoT firmware and empirical studies on IoT devices. According to the research plan, the leading institution, Zhejiang University, solely conducted the survey (and indeed, it was). During the whole analysis, although this institution does not have an IRB, we followed principles outlined in the Menlo Report [11] and the local regulations to protect the rights of human participants.

Detailed Steps. We took the following steps to perform the experiments ethically. (1) All participants were informed about the purpose of the study and consented to participate in the survey before filling out the questionnaire. *Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection*. For instance, our user study only recruits adults instead of children. (2) *We ensured that the questions were not connected to participants' identities when designing the questionnaire*. Meanwhile, *we respect participants' right to determine their own best interests*. For instance, we respect their right to keep their age and gender secret in the questionnaire. (3) When we perform forensic analysis on a device, we only detect the magic strings we predefined rather than any other information. Thus, *we did not collect or use any user information during our analysis*. (4)

TABLE II: Summary of the investigated devices. # represents the leading vendors of a certain type of devices (we mask the model names to prevent potentially malicious actors).

ID	Vendor	Model	Device Type
1	Amazon#	Fi...	Media
2 3	Arris#	AC... AC...	Network Node Network Node
4	China Mobile	CM...	Media
5	Dahua#	TP...	Surveillance
6 7	D-Link#	DC... DI...	Surveillance Network Node
8	Google#	Ch...	Media
9 10 11	Hikvision#	DS... DS... CS...	Surveillance Surveillance Surveillance
12 13	Hiwifi	1S... 3p...	Network Node Network Node
14 15	HP#	La... La...	Work Work
16 17 18	Huawei#	HG... EC... Hi...	Network Node Media Media
19 20	Phicomm	K2... K2...	Network Node Network Node
21	MaxHub	V5...	Work
22	Mercury	MW...	Network Node
23 24	Netgear	R8... WN...	Network Node Network Node
25	Roku#	EX...	Media
26	Schneider#	M2...	PLC
27 28 29 30	TP-Link#	TL... TL... TL... WR...	Network Node Network Node Network Node Network Node
31 32	Tuya	TY... SC...	Network Node Surveillance
33	Xiaomi	R3...	Network Node

We claim that *we do not expose any user data and metadata to others*. We ensured that the authors from other institutions were not engaged in any step of the work involving human subjects. These authors have access to only the aggregated results presented in the paper. Besides, we deleted all the metadata, such as the extracted IoT firmware, device logs, etc., generated in the analysis process. (5) *We reported our findings to the corresponding vendors and actively communicated with them to alleviate potential sensitive data leakage risks.* Besides, we masked the detailed information of the tested IoT firmware and devices to prevent potentially malicious actors.

IV. User Awareness

In this section, we first show the real-world re-using of IoT devices. Then, we investigate the three research questions raised in §I from a user's perspective. Finally, we characterize users' preferences that may increase the security risks of user-data leakage caused by improper disposal.

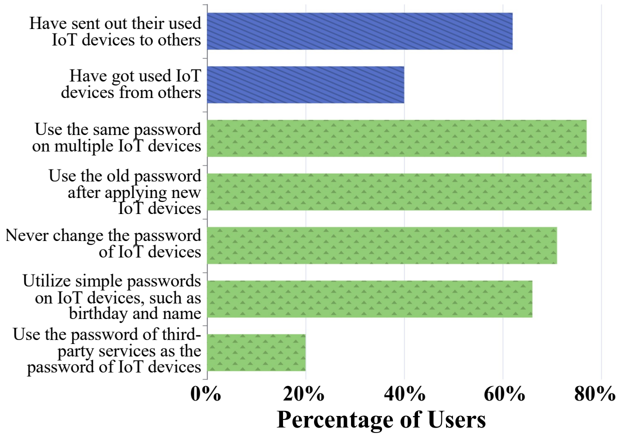

Fig. 4: Partial results of our user study, from which, we can learn that (1) IoT device re-using is common and (2) users' password management of IoT devices is very poor.

IoT device re-using. Our user study shows that IoT devices are widely spread in people's daily lives. About 97% of the surveyed users have ever used IoT devices, including smart cameras, routers, printers, etc. Besides, the re-using of IoT devices is frequent. For one thing, users replace IoT devices common. Over 91% of the users update IoT devices within five years. For another thing, most users are willing to re-using IoT devices. As shown in Figure 4, after a new deployment of an IoT device, 62% of the users choose to sell, discard, or lend the old one. Moreover, 40% of the users have ever bought or borrowed IoT devices from others. This result indicates that many used IoT devices are re-using among different people, which, however, means that an adversary can easily obtain the used IoT devices from a victim leading to new security risks of user-data leakage.

Which kinds of sensitive data do users believe reside in used IoT devices? This study shows that 80.2% of the surveyed people are worried about the leakage of their personal data in used IoT devices. Interestingly, however, many users do not really understand what sensitive information may store in a used IoT device. For example, 25.6% and 38.8% of the users have no idea of that home WiFi information and third-party accounts may be stored in a used IoT device, respectively. Such misunderstanding of sensitive data in a used IoT device may lead to improper data disposal with the used IoT device, resulting in data leakage risks.

Which methods do the surveyed users use to dispose of the sensitive data? Our study presents that 48.9% of the users erase their sensitive data by overwriting, soft resetting, or hard resetting when disposing of a used IoT device. The remaining 51.1% of the users do not erase their sensitive data in IoT devices before selling or sending it to others. Indeed, 42.9% of these users have no idea about how to erase sensitive data in a used IoT device. This result reveals that many users are unfamiliar with the data protection techniques provided by IoT vendors, which hinders the proper data disposal with used IoT devices.

Do the surveyed users believe that the disposal method they use is effective? For users who erase sensitive data through different methods, 86.9% of them believe that the used disposal methods can effectively erase the sensitive data in used IoT devices. For the remaining users, 22.9% of them are unaware of the potential data leakage risks. The reason may be that, about 56.3% of them believe that IoT vendors encrypt the sensitive data in IoT devices to protect customers' privacy and security. This result reveals that most users trust the data disposal and protection methods provided by IoT vendors, which, however, are not as effective as users expect (as will be discussed in §V-D and §VI-B).

Users' preference. In fact, to alleviate data leakage in used IoT devices, various IoT vendors require users to set suitable passwords for their IoT devices. However, our study indicates that users' password management of IoT devices is very poor (shown in Figure 4): (1) about 77% of the users set the same password for multiple IoT devices; (2) for 78% of the users, the old password is still applied in a newly deployed IoT device; (3) 71% of the users never change the password of their IoT devices; (4) 66% of the people use their sensitive information involving birthday, workplace, ID number, or its variance as the password of an IoT device; and (5) about 20% of the users leverage the password of a third-party service, including email, blogs, etc., as the password of an IoT device. This practice may increase the security risks of data leakages caused by improper data disposal. For example, once the password of a used IoT device is leaked, an attacker may leverage the password to launch a credential-stuffing attack to control more devices and third-party accounts of the victim.

Summary of findings. IoT device re-using is universal in users' daily lives, during which more than half of the users do not erase their sensitive data in used IoT devices. Only 13.1% of the remaining users have ever doubted the effectiveness of the used disposal methods. This user study indicates that users usually do not clearly understand what sensitive data may store in a used IoT device and the effectiveness of data disposal methods. On the one hand, users may underestimate the data leakage risks caused by IoT re-using. On the other hand, many users lack the technical knowledge to dispose of their sensitive data properly. These findings reveal an urgent need to investigate the data disposal of used IoT devices.

V. Sensitive Data in IoT Devices

In this section, we investigate "**RQ1**: *which kinds of sensitive data reside in used IoT devices?*". This question helps us understand the potential leakage risk of the stored sensitive data without proper data disposal. As discussed in §III-B, due to scalable and ethical considerations, we translate this question to "*IoT devices collect which kinds of user data?*" Specifically, we leverage our sensitive data analysis system to detect user-data collection in Dataset-1. In the rest of this section, we first evaluate the accuracy of our sensitive data analysis system. Then, we elaborate on the analysis result. Furthermore, we elaborate on sensitive information residing in real devices in Dataset-2.

A. The Accuracy of Sensitive Data Analysis

The accuracy of the sensitive data analysis system influences the result of **RQ1**. Thus, before performing the large-scale analysis with the system, we manually examine its accuracy on 11 randomly chosen firmware images from Dataset-1. As shown in Table III, these images represent various device types and vendors. Our system reports 808 user-data collections in the tested firmware images. After manual analysis, we determine that 728 out of these reports are true positives. Besides, the system misses 137 user-data collections which we manually find in the tested firmware images. Thus, the precision and recall of the system are 90.10% and 84.16%, respectively. Compared to the SOTA sensitive information tracking systems, such as [17], our system achieves a similar precision while its recall is relatively lower than the SOTA (92.59%). We will discuss how to further improve the detection accuracy in §VII. However, considering the measurement purpose of this study, we believe this conservative system (that reports a lower bound on user-data collections) can still provide valuable insights on the following fundamental questions: *What types of sensitive information are collected by IoT firmware? How many sensitive items are collected?*

B. Types of Sensitive Data in IoT Devices

To determine the classifications of sensitive data, we first manually analyzed the 728 detected user-data collections of the 11 random samples. As shown in Table IV, we determined to categorize the collected sensitive data into four classes based on prior works (such as [17]) and our empirical analysis. Meanwhile, we explore the potential consequence of data leakage of each kind of sensitive data caused by improper disposal. Then, by leveraging the sensitive data analysis system, we perform a large-scale analysis on the 4,749 firmware images listed in Table I. We find 121,984 user-data collections in 3,611 firmware images. All user-data collections were classified manually by their code context, such as the variable name of the collected data.

(1) *Device management* information. Most devices have a device management account to configure this device and check the device status. Although such information seems bound to the specific device, it is still sensitive for the following reasons. First, according to our user study in §IV, users tend to use the same device management account and password across different devices. Thus, once an attacker obtains a device management account, he/she may control the user's other devices with the same account. Second, users also tend to use the password of third-party services as the password of IoT devices. Thus, the attacker may conduct a credential stuffing attack to access the user's third-party accounts (e.g., email accounts and bank accounts) and launch various subsequent attacks [3], [7], [29], [34], [35], [37].

(2) *Network setting* information. Most IoT devices need a home network to communicate with other devices and internet

TABLE III: The analysis result of random samples (we mask the firmware names to prevent potentially malicious actors).

Vendor	Device Type	Firmware	#Source	#Sink	#Collection	TP	FP	FN	Precison	Recall
fastcom	Router	FE...	467	815	456	427	29	86	93.64%	83.24%
TP-Link	AC Controller	TL...	227	329	216	190	26	24	87.96%	88.79%
Phicomm	Router	K2...	158	198	89	80	9	20	89.89%	80.00%
ROOter	Router	Ar...	45	61	20	14	6	5	70.00%	73.68%
MiCasaVerde	Gateway	ve...	41	45	4	1	3	1	25.00%	50.00%
TP-Link	Surveillance	TL...	6	18	6	5	1	1	83.33%	83.33%
trendnet	Router	te...	31	706	5	3	2	0	60.00%	100.00%
GL.iNet	Router	mt...	49	67	5	3	2	0	60.00%	100.00%
OpenWRT	Access Point	21...	7	2	2	2	0	0	100.00%	100.00%
Avalon	Miner	av...	40	62	3	1	2	0	33.33%	100.00%
8devices	SOM	li...	10	2	2	2	0	0	100.00%	100.00%
Total			**1081**	**2305**	**808**	**728**	**80**	**137**	**90.10%**	**84.16%**

TABLE IV: Summary of the sensitive data types in IoT devices.

Class	Type	Potential Risks
Device Management	Admin/User Account UUID Device Name	Attackers may conduct credential stuffing attacks and control user's devices.
Network Settings	WiFi SSID & Password IP Settings	Attackers may intrude into the home network environment and attack other devices.
Third-Party Account	DDNS Account PPPoE Account Email Account VPN Account FTP Account	Attackers may steal confidential emails, intrude into the VPN network environment, hijack DNS, upload malicious files, download sensitive files, and conduct phishing attacks.
User Portrait	Owned Devices Browsing History Surveillance Video Security Settings	Attackers may steal the user's sensitive data, blackmail the user, bypass network/physical security protections, and break into the house.

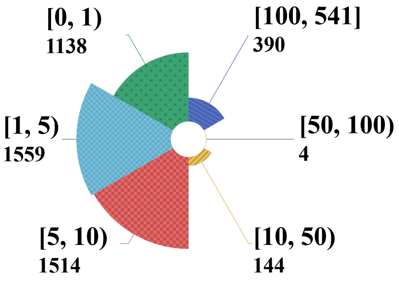

Fig. 5: The distribution of firmware images that store different numbers of sensitive items.

servers during the working process. Thus, IoT devices tend to store network setting information. For example, the home WiFi SSID and password are stored to connect to the home network automatically after a restart. Once an attacker obtains the WiFi SSID and password, he/she could be able to break into the user's home network. Consequently, the attacker can discover and attack other devices in the home network.

(3) *Third-party account* information. IoT devices may require third-party accounts to provide various services to users. For example, a smart camera may require a user's email address to inform the user when it detects a suspicious movement. Once an attacker obtains such third-party account information, he/she could be able to access these accounts. Furthermore, the attacker can launch many severe attacks, such as stealing confidential emails, hijacking DNS, and uploading malicious files. Moreover, since users tend to use the same username and password for different third-party services, the attacker can conduct credential-stuffing attacks to control more accounts and launch further attacks.

(4) *User portrait* information. Deeply integrated into daily lives, IoT devices may store user portrait information. For example, a surveillance device may store the audio or video of a user's daily life. Once an attacker obtains such user portrait information, he/she may sell the user's sensitive data or blackmail the user. Moreover, some IoT devices also store users' security settings, such as monitoring areas in the house. The attacker may break into the house without being monitored by leveraging such information.

C. The Distribution of Data Collection

This section investigates the distribution of the user-data collections discovered in Dataset-1. Specifically, we want to answer the following questions.

(1) How many sensitive items can a firmware image collect at most? A sensitive item is a piece of sensitive data. We investigate the number of sensitive items collected by each firmware image. As shown in Figure 5, we find that 2,052 (43.2%) of the analyzed firmware collect at least 5 sensitive items. Among them, 390 (8.2%) of the firmware collect more than 100 sensitive items. One IoT firmware image collects up to 541 sensitive items. This result reveals that collecting sensitive data is quite universal in IoT firmware images. Thus, it is important to conduct proper data disposal after the usage of a device.

(2) What is the most pervasive sensitive data type? Among all the sensitive data types listed in Table IV, we want to know the most pervasive one. Thus, we investigate the number of firmware images that collect each type of sensitive data. As shown in Figure 6, it is interesting that IoT firmware images collect more sensitive data than users expect (learned from §IV). For example, over 25.6% of the surveyed users believe that used IoT devices do not store WiFi information. However, we find that WiFi SSID and WiFi password are the most pervasive sensitive data types collected by IoT firmware (discovered in 958 and 933 firmware images, respectively). On the one hand, the misunderstanding of users may lead to improper data disposal with used IoT devices, further resulting in data leakage.

TABLE V: Sensitive information residing in IoT devices. ✓ = we can extract the sensitive data; ✗ = we fail to extract the sensitive data; ○ = we can extract the sensitive data while developers attempt to protect it.

Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Admin User name		✓	✗			✓	✗			✓	✓	✗	✗	✗		✓			✓	✓		✗	✗	✗			✓	✗	✓	✗			✗
Admin Password		✓	✗			✓	✗			✗	✗	○	○	✗		✗			○	○		✗	✗	✗		✗	✗	✗	✗	✗			✗
WiFi SSID	✗	✓	✓	✗	✗	✓	✓	✗			✗	✓	✓	✓		✓	✗	✗	✓	✓	✓	✓	✓	✓	✗		✓	✓	✓	✓	✓	✗	✓
WiFi Password	✗	○	✓	✗	✗	✓	✓	✗			✗	✓	✓	✓		○	✗	✗	✓	✓	✓	✓	✓	✓	✗		✓	✓	✓	✓	✓	✗	✓
FTP User name		✓				✓					✓																						
FTP Password		✓				✓					✗																						
DDNS User name		✓	✓			✓				✓									✓	✓			✓				✓	✓	✓	✓			✓
DDNS Password		✗	○			✓				✗									✓	✓			○				✓	○	✓	○			✓
Email Address	✗	✓	✓			✓		✓		✓											✓		✓		✓								
Email Password	✗	✗	○			✓		✗		✗											✗		○		✗								
VPN User name		✓										✓	✓	✓										✓			✓						
VPN Password		✓										✓	✓	✗										✓			✓						
Other Device			✓							✓		✓	✓	✓					✓	✓		✓					✓	✓	✓	✓			✓
Others	✓	✓	✓	✓				✓		✓		✓	✓				✓	✓							✓		✓		✓				

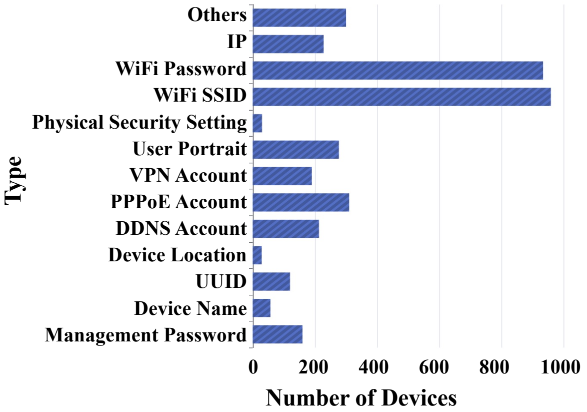

Fig. 6: The distribution of firmware images that store different types of sensitive data.

On the other hand, this result is reasonable because most IoT devices require home WiFi information to communicate with other devices and internet servers. However, as the most pervasive sensitive data type, home WiFi information has never been reported by prior studies [17], [20]. This finding reveals that the data disposal of used IoT devices needs more attention.

We also investigate user-data collections over the years and user-data collections in different geographical regions. The results reveal that user-data collections in IoT have been universal worldwide in recent years. More details on the distribution of the user-data collections are deferred to Appendix §A-D.

D. Device Analysis

To investigate the user-data collection in real-world IoT devices, besides firmware analysis, we also conduct empirical studies on IoT devices in Dataset-2. Specifically, as described in §III-C, we set up each device using magic strings and detect sensitive items by forensic analysis. Table V shows that various kinds of sensitive information reside in the 33 IoT devices in Dataset-2. Totally, we find 185 sensitive items. On average, one device contains 5.61 sensitive items. The result again shows that sensitive data is universal in used IoT devices. Consistent with the results reported in §V-C, each sensitive data type discovered by firmware analysis is also discovered by device analysis.

Whether IoT vendors provide proper data protection for users? 118 (63.8%) out of the 185 sensitive items are stored in plain text. For 12 sensitive items discovered through web interfaces, we observe that IoT vendors have attempted to protect them by setting the attribute *type* of an *input* tag to "*password*". In this case, the sensitive item in the *input* tag is shown as a sequence of "*". Unfortunately, we find that such protection is inadequate—one can easily bypass it by changing the value of *type* to "*text*".

Case Study of the `MaxHub` Smart Screen. We find a sensitive item through the application interfaces of a work device—the smart screen produced by `MaxHub` (whose ID is 21 in Table II). Specifically, when one sends an empty message to the *7434* network port of the device, this application interface will respond with a message containing a WiFi SSID and the corresponding password. According to further empirical analysis, we find that this application interface is used by a system application of the smart screen device to get WiFi information. However, except for connecting a network through this application interface, we believe that there exist other more secure methods to achieve this goal. For example, the application can obtain WiFi information through software APIs. This result reveals that it is important to regulate the operations that handle sensitive data in IoT devices.

The above result reveals that IoT vendors do not provide adequate data protection. This result is also opposite to users' expectations, i.e., user-data may leak in various unexpected ways.

Summary of findings. Our findings show that sensitive data is universal and diversified in IoT devices. Users' expectations of sensitive data stored in IoT devices bias from reality—some of the most pervasive sensitive information (e.g., WiFI SSID and password) is stored in used IoT devices without many users' awareness. Moreover, we also observe that some IoT developers attempt to protect user data by the attribute settings of access interfaces. However, the protections are oftentimes insufficient, which can be easily bypassed. Although the distribution of user-data collection varies in different vendors, release times, and locations, considering that the sensitive data residing in used IoT devices has severe security and privacy impact, our findings reveal the importance of proper data disposal of used IoT devices.

VI. Understanding Data Disposal Methods

In this section, we aim to answer the following questions proposed in §I from the real-world. **RQ2**: *which methods can be used to dispose of sensitive data?* and **RQ3**: *is the disposal method effective in erasing the sensitive data?* Specifically, we first elaborate on the data disposal methods provided by the devices in Dataset-2 (§VI-A). Then, we investigate the effectiveness of disposal methods by forensic analysis (§VI-B).

A. Data Disposal Methods

We manually investigate the methods that can dispose of the devices in Dataset-2 and categorize these methods. (1) One may overwrite or remove sensitive data through a user interface. For example, one can log into the configuration page of a smart router to overwrite/remove the WiFi password. (2) One may perform a soft resetting by clicking the "reset to factory defaults" button on the configuration page of an IoT device. (3) One may perform a hard resetting by pressing the RESET button on the device. (4) One may perform a firmware upgrade by clicking the "upgrade firmware" button on the configuration page of an IoT device. (5) One may log in to the terminal of an IoT device and overwrite/remove the files that store user data. The last method requires technical skills and is thus too complex for typical users to conduct. Therefore, we choose to investigate the first four user-friendly methods. Next, we elaborate on the methods provided by the devices in Dataset-2 and investigate the effectiveness of these methods.

B. Effectiveness of Disposal Methods

Available access interfaces. When conducting forensic analysis on a device, the first step is to detect as many available access interfaces of this device as possible. This task can be completed by existing tools, such as `nmap` [12]. Finally, we find 213 access interfaces in total and each device provides more than 6 interfaces on average. Then, based on our empirical experience, the access interfaces of an IoT device can be categorized into the following three types according to their utilities: (1) *user interface*, (2) *debug interface*, and (3) *application interface*.

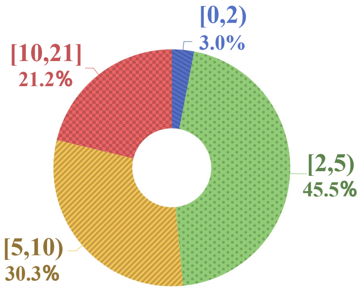

Fig. 7: The distribution of IoT devices with different numbers of access interfaces.

To understand the distribution of the access interfaces of IoT devices, we first investigate the distribution of devices with different numbers of access interfaces. As we can see from Figure 7, IoT devices have many available access interfaces. Specifically, more than 51.5% (17) of the devices provide at least 5 access interfaces. One device provides up to 21 access interfaces. The results reveal that access interfaces are common in IoT devices, providing the precondition for forensic analysis on IoT devices.

We then investigate the distribution of different interfaces and present the most pervasive interfaces in Figure 8. It is interesting to see that 76 (35.7%) out of the 213 discovered interfaces are `Unknown`, which indicates that existing tools cannot identify the type of these interfaces. This result reveals that IoT vendors tend to develop self-defined interfaces to accomplish various utilities on their products. We also discover many interfaces that are widely used in traditional PCs. For example, we find 43 HTTP interfaces (ranking 2), demonstrating that IoT devices often provide web services to users. Besides, we also find 8 `ssh` interfaces (ranking 4) and 4 `telnet` interfaces (ranking 8). Previous studies, such as [30], have highlighted that it is dangerous to open such interfaces since one may attack a device through them. However, our findings reveal that vendors and users still leave such dangerous interfaces open and suffer such security risks. We present the detection results of several dangerous interfaces in Table VI.

Results. Through the detected access interfaces, we can conduct forensic analysis to discover sensitive data in the devices after performing different disposal methods. Table VII shows the result of forensic analysis. Overall, we find 23 ineffective disposal methods in 9 devices. Specifically, data overwriting and firmware upgrades of all these 9 devices are ineffective. Besides, the soft resetting and hard resetting of 1 and 5 devices are ineffective, respectively. The results reveal that the disposal methods, including data overwriting, soft/hard resetting, and firmware upgrades, oftentimes cannot effectively erase users' sensitive information in a used IoT device. This is because IoT vendors do not really erase sensitive data residing in a storage unit of a device. Instead, they only mark the state of the storage unit from valid to invalid. Thus, the users may face the risk of high-volume data leakage since disposal methods cannot prevent attackers from obtaining user data from used

TABLE VI: Dangerous Access Interface of IoT devices. ✓ = we can access the interface directly. ○ = we can bypass the protection of the interface. ✗ = we fail to access the interface since it is protected.

Access Interface	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
TTL			✓	✗	✓	✗	✓		✗	○	✗	✓	✗	○	✓	✓	✓		✓			✗	✓		✗				✓			
SSH/Telnet	○		○				○		○			○	○	✗			○		○				○									
ADB																								✓			✓					✓

TABLE VII: The effectiveness of disposal methods. ✓ = we can obtain sensitive information after disposal.

Access Interface	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Overwriting		✓									✓	✓			✓			✓	✓								✓		✓				
Soft resetting																			✓														
Hard resetting											✓				✓			✓	✓										✓				
Firmware upgrade		✓									✓	✓			✓			✓	✓							✓	✓		✓				

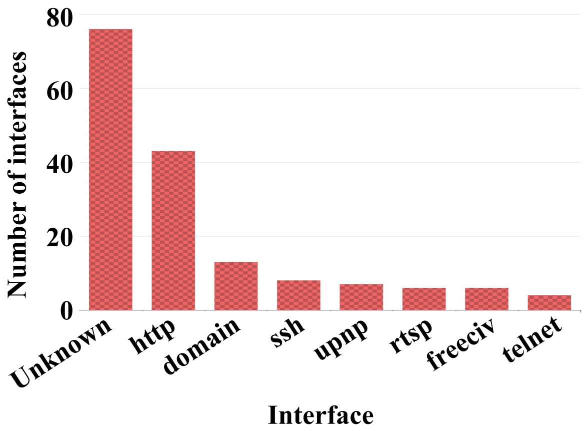

Fig. 8: The most common interfaces on IoT devices.

TABLE VIII: The sensitive data obtained after different data disposal methods; ✓ = we can extract the sensitive data after disposal; ✗ = we fail to extract the sensitive data after disposal.

Disposal Method	Overwriting		Soft Resetting	Hard Resetting	
	27	29	20	12	19
Admin User name	✓	✓	✓	✗	✓
Admin Password	✗	✗	✓	✗	✓
WiFi SSID	✓	✓	✓	✓	✓
WiFi Password	✓	✓	✓	✓	✓
FTP User name					
FTP Password					
DDNS User name	✓	✓	✓		✓
DDNS Password	✓	✓	✓		✓
Email Address					
Email Password					
VPN User name	✓			✓	
VPN Password	✓			✓	
Other Device	✓	✓	✓	✓	✓
Others	✓			✓	

IoT devices.

Table VIII shows several cases of ineffective disposal methods. Next, we elaborate on several insightful findings from these cases. (1) The effectiveness of different disposal methods varies even on the same device. For instance, after hard resetting, we can obtain all the sensitive data residing in a `Phicomm` device (whose ID is 19). In contrast, its soft resetting can effectively erase sensitive data. This finding indicates that the lack of consistent implementation standards is an essential reason for ineffective disposal. (2) Soft resetting seems more effective than other disposal methods. We only find one device adopts ineffective soft resetting. This result reveals that users should at least conduct soft resetting before reselling/discarding their devices. (3) We fail to recover sensitive data for some devices because the sensitive data in these devices is encrypted. For example, one device from Hiwifi and two from TP-Link (whose IDs are 12, 27, and 29) encrypt the admin password. Therefore their sensitive data cannot be recovered. Our firmware scanner currently does not support data decryption, which is a fundamental challenge in cryptography. However, we observe that data encryption is not universal in IoT devices. For example, most (84.0%) sensitive data in these three devices and all sensitive data in the other two devices are stored in plain text, which is opposite to users' expectations.

C. Further Analysis on Firmware Images

To better understand the status quo of the risk of ineffective data disposal, we conduct an in-depth investigation of the results of the forensic analysis. Specifically, we investigate the implementation of the data disposal of the tested devices by analyzing their firmware. We observe two characteristics of the devices that are influenced by ineffective data disposal. First, these devices use wear-leveling write filesystems, such as jffs2 and ubifs. Second, they perform data disposal by deleting user-data files (such as "rm -r /userdata") rather than conducting flash-level erasing. Based on these observations, we manually analyze more IoT firmware to infer whether they also face the risk of ineffective data disposal.

Specifically, we randomly sample 500 firmware images from Dataset-1. For each firmware, we examine these two questions: (1) whether it uses wear-leveling write filesystems and (2) whether it performs data disposal by deleting user-data files. For the first question, we can check the boot file (such as `/sbin/firstboot`) of each firmware to see whether it mounts a jffs2 or ubifs filesystem. For the second question, we can check the reset file (such as `/etc/rc.button/reset`) to see whether a firmware image only deletes user-data files for resetting.

Result. We find that 107 out of the 500 investigated firmware images use wear-leveling write filesystems. 66 out of these 107 images perform data disposal by deleting user-data files.

Thus, 13.2% of the investigated firmware images face the risk of ineffective data disposal. These results reveal that the data disposal oftentimes cannot effectively erase user data in used IoT devices. Therefore, even "aware" users who perform data disposal before reselling/discarding their used devices still face the risk of data leakage.

Case study of ineffective data disposal. Among the 66 firmware images that face the risk of ineffective data disposal, we find that 30 images use the same third-party program `jffs2reset` [10] to perform data disposal. This open-sourced program is widely used in IoT firmware. However, it does not effectively erase user data since it only deletes user-data files with the "rm" command. This case shows that IoT vendors tend to leverage third-party program when developing firmware images. However, the third-party program may lead to security and privacy risks.

Summary of findings. Our finding reveals that the effectiveness of data disposal, including data overwriting, soft/hard resetting, and firmware upgrades, are oftentimes ineffective. Specifically, due to the inherent characteristics of the storage unit of a device, one can still obtain sensitive data after the data disposal. Moreover, most IoT vendors do not provide data encryption as users expect. In summary, one can obtain sensitive data in IoT devices through various easy-to-conduct methods (such as the forensic analysis performed in this paper). Our finding raises an alarming issue that sensitive data in IoT devices faces severe risks, calling for future efforts to protect sensitive data in IoT devices.

D. Implications

Our investigation shows that when disposing of highly sensitive data stored in IoT devices, existing data disposal is oftentimes inadequate and ineffective. Thus, sensitive data in used IoT devices faces severe leakage risks. These findings call for IoT vendors, users, policy makers, and the research community to ponder the data disposal of used IoT devices.

Implications for IoT vendors. We have reported all the findings to the corresponding IoT vendors and received responses from four vendors. They have confirmed the potential data leakage risk caused by improper data disposal. Interestingly, these vendors show different attitudes toward this problem. For one thing, three vendors indicate that this is an important issue and will take action to alleviate the potential leakage of user data. However, one vendor points out that they will not provide extra protection for the related IoT devices because (1) the devices have expired and (2) they have provided a more secure user configuration by a specifically designed app. Thus, this vendor believes users should share some responsibility for data leakage if they still use other (insecure) configuration methods.

In addition, we also provide multiple potentially useful suggestions for IoT vendors. For example, we suggest vendors improve the device resetting methods (e.g., by erasing the flash storage) to ensure that data disposal can really erase the sensitive data in a device. We also suggest vendors encrypt sensitive data before storing it in an IoT device. Finally, we suggest that the data leakage of expired devices also deserves attention as long as the devices are still recycled in the market. Besides, IoT vendors should actively implement adequate security mechanisms. More details on the best practice information are deferred to Appendix §A-E.

Implications for users. We observe that users may misunderstand the security risk of data leakage before reselling/discarding an IoT device. We hope that our study would serve as an alarm for users to protect their sensitive data when using an IoT device. Moreover, we provide the following comments for users to prevent data leakage. (1) It is important to erase sensitive data before selling, renting, or dropping used IoT devices. Additionally, it is necessary to ensure that the used disposal method can effectively erase sensitive data. For example, one may leverage our system to evaluate whether a data disposal method is effective. (2) If one cannot ensure the effectiveness of a disposal method, one may erase data in a device by multiple methods multiple times. (3) It is important to use different passwords for different accounts. Meanwhile, it would be helpful to use a new password when replacing an old device with a new one. Specifically, we obtained an acknowledgment from a worldwide leading industry control company after we reported our findings to them. We continually work with the company to examine the IoT devices they adopt and help them identify sensitive data leakage risks.

Implications for policy makers. Our finding reveals that users are suffering from the risk of data leakage when reselling/discarding their devices. Thus, we hope our findings would encourage policy makers to propose better regulations to protect users' data. First, it is important to regulate which kind of and how sensitive data can be stored in IoT devices. For example, IoT vendors should encrypt sensitive information before storing them in the device. Second, it is necessary to guarantee accountability when re-using IoT devices. For example, a recycler should ensure that users' sensitive data is erased before being resold.

Implications for the research community. Our work uncovers a critical risk of data leakage in the IoT ecosystem caused by improper data disposal. Additionally, due to the common re-using of IoT devices, this security risk tends to be pervasive in people's daily lives. We hope our study would pave the way to conduct further comprehensive studies to improve data disposal with used IoT devices. For example, our finding shows that the wear-leveling write strategy (a widely used write strategy in IoT devices) still stores "deleted" sensitive data after data disposal. This is the root cause of ineffective data disposal. Suppose researchers can design more secure write strategies for flash file systems used in IoT devices. The data disposal with IoT devices will become more effective.

VII. Discussion

In this work, we take the first step to understand data disposal along with IoT devices and obtain multiple enlightening findings. However, there still exist several limitations in our work. Next, we discuss the limitations and potential directions for future work.

More accurate analysis of user-data collections in IoT firmware images. As presented in §III-B, we implement a sensitive data analysis system to discover user-data collections in IoT firmware images. However, it may miss user-data collections. For example, if a firmware image collects user data by leveraging a pair of closed-source APIs, our system cannot report this collection. Besides, the system also has FPs when mistakenly identifies a wrong SAPI. We plan to leverage NLP techniques in machine learning research to improve the system. Specifically, one can train an NPL model to classify the SAPIs in the firmware. Then, by leveraging the identified SAPIs, our system may discover user-data collections more accurately. Despite the above limitations, our system already shows the ability to discover user-data collections with reasonable precision and recall. By leveraging this system, we provide the first systematical view of user-data collections with IoT devices and uncover a new risk of data leakage in the IoT ecosystem.

Manual analysis. We perform the necessary manual analysis in this paper for several goals. First, we prepare a list of widely-used library SAPIs and a list of known common APIs in §III-B to support our sensitive data analysis system. It could be unaffordably time-consuming to predefine all these APIs in our firmware dataset manually. Fortunately, by leveraging our two-layer API inferring method, we only need to prepare a small number of APIs. Specifically, our system achieves reasonable precision and recall with only 40 predefined APIs. Second, we conduct an empirical study on IoT devices in §III-C to investigate user-data disposal methods.

We believe designing and developing more automatic methods to reduce manual effort is an interesting future direction. For example, the SAPIs reported by our system can be used to train an AI model to identify SAPIs automatically. Besides, one may develop an automatic system to configure, dispose of, and discover user data in IoT devices.

Large-scale and long-term study. To investigate the geographic diversity and trend of user-data collection in IoT firmware, we manually collect hundreds of firmware images' release times and sales districts. This dataset enables us to have a preliminary understanding. However, further focused research on large-scale and long-term datasets is required to understand these fundamental questions better.

VIII. Related Work

User perceptions of IoT sensitive data. Many researchers have investigated user awareness and understanding about the data leakage problem of IoT devices in their daily life [21], [24], [25], [26], [31], [38]. These studies have shown that during the daily use of IoT devices, many users are distrustful of IoT devices due to privacy and security concerns [21], [31]. Nevertheless, many users are still willing to accept the risks in favor of the convenience offered by IoT devices. In addition, they tend to feel limited responsibility for mitigating risks due to constrained options or lack of knowledge to conduct protections [26]. Our findings corroborate some of the prior observations (e.g., most users are concerned about data leakage). However, our work mainly aims to investigate the user awareness of sensitive data in IoT devices and their understanding of data disposal when re-using IoT devices. For example, we investigate how users depose of used IoT devices and do they trust the data disposal methods provided by IoT vendors.

IoT privacy analysis. Researchers have conducted various works to study privacy leakage risks in the IoT ecosystem [17], [18], [20], [22], [23], [28], [32], [33], [36], [40], [41]. Some of these works focus on the privacy collection problem in the companion APPs of IoT devices [17], [22], [23], [32]. For example, Celik et al. [17] proposed SAINT to perform static taint analysis on IoT applications. By leveraging SAINT, they evaluated 230 SmartThings market APPs and found 138 (60%) APPs include sensitive data flows. Other prior works focus on the privacy leakage problem of IoT network traffic [18], [20], [33], [36], [41]. For example, Chu et al. [18] found that due to the lack of encryption/authentication, personal data are unprotected when smart toys communicate with cloud services. By contrast, to the best of our knowledge, our study is the first work investigating the data disposal of used IoT devices. Our study enables us to reveal the serious risk of data leakage while re-using IoT devices.

IX. Conclusion

In this paper, we perform the first systematical study of the user-data disposal of used IoT devices. We first conduct a user study to understand user awareness of existing data disposal methods, from which we find that for most users, the lack of awareness and technical skills hinder them from properly disposing of their used IoT devices. Then, we conduct sensitive data analysis on 4,749 firmware images and 33 devices to discover user-data collections in IoT firmware images and validate the effectiveness of data disposal methods. The results show that while there are way more sensitive data than users expect, current data protections of used IoT devices are inadequate. Moreover, not consistent with user expectations again, the disposal methods of used IoT devices are often ineffective. In summary, without proper disposal, one can obtain various sensitive data, such as user portraits, third-party accounts, etc., from a used IoT device by various easy-to-conduct methods. Our findings uncover a serious (but without sufficient attention) risk of high-volume data leakages caused by improper data disposal of used IoT devices. We propose multiple suggestions for both users and IoT vendors. We believe our work can serve as a critical enabler for improving IoT security and privacy.

X. Acknowledgment

This work was partly supported by National Key R&D Program of China (2020YFB2010900), NSFC under No. U1936215, the State Key Laboratory of Computer Architecture (ICT, CAS) under Grant No. CARCHA202001, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

REFERENCES

[1] Amazon, Mar 2022. https://www.amazon.com.

[2] binwalk, Mar 2022. https://github.com/ReFirmLabs/binwalk.

[3] Check Point research reveals how hackers can intrude networks via fax machines, Mar 2022. https://cisomag.eccouncil.org/check-point-research-reveals-how-hackers-can-intrude-networks-via-fax-machines/.

[4] Craigslist, Mar 2022. https://auburn.craigslist.org/.

[5] eBay, Mar 2022. https://www.ebay.com/.

[6] The Global E-waste Monitor 2020, Mar 2022. https://ewastemonitor.info/gem-2020/.

[7] Hackers can steal your identity and bank details from a coffee machine, Mar 2022. https://cisomag.eccouncil.org/hackers-can-steal-your-identity-and-bank-details-from-a-coffee-machine/.

[8] Internet of Things (IoT) active device connections installed base worldwide from 2015 to 2025, Mar 2022. https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide/.

[9] jefferson. JFFS2 filesystem extraction tool, Mar 2022. https://github.com/sviehb/jefferson.

[10] jffs2reset, Mar 2022. https://git.openwrt.org/?p=project/fstools.git;a=blob;f=jffs2reset.c.

[11] The Menlo Report, Mar 2022. https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf.

[12] nmap, Mar 2022. https://github.com/nmap/nmap.

[13] Privacy Commissioner Urges IoT Manufacturers to Enhance the Transparency of Their Privacy Protection Measures, Mar 2022. https://www.pcpd.org.hk/english/news_events/media_statements/press_20170124.html.

[14] Switched on to value: Powering business change, Mar 2022. https://wrap.org.uk/sites/default/files/2021-03/WRAP-switched-on-to-value-powering-business-change.pdf.

[15] Total Green Recycling, Mar 2022. https://www.totalgreenrecycling.com.au/about-us/.

[16] Ado Adamou Abba Ari, Olga Kengni Ngangmo, Chafiq Titouna, Ousmane Thiare, Alidou Mohamadou, Abdelhak Mourad Gueroui, et al. Enabling privacy and security in cloud of things: architecture, applications, security & privacy challenges. Applied Computing and Informatics, 2019.

[17] Z Berkay Celik, Leonardo Babun, Amit Kumar Sikder, Hidayet Aksu, Gang Tan, Patrick McDaniel, and A Selcuk Uluagac. Sensitive information tracking in commodity iot. In 27th USENIX Security Symposium (USENIX Security 18), pages 1687–1704, 2018.

[18] Gordon Chu, Noah Apthorpe, and Nick Feamster. Security and privacy analyses of internet of things children's toys. IEEE Internet of Things Journal, 6(1):978–985, 2018.

[19] Avirup Dasgupta, Asif Qumer Gill, and Farookh Hussain. Privacy of iot-enabled smart home systems. In Internet of Things (IoT) for Automated and Smart Applications. IntechOpen, 2019.

[20] Shuaike Dong, Zhou Li, Di Tang, Jiongyi Chen, Menghan Sun, and Kehuan Zhang. Your smart home can't keep a secret: Towards automated fingerprinting of iot traffic. In Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, pages 47–59, 2020.

[21] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into iot device purchase behavior. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

[22] Earlence Fernandes, Jaeyeon Jung, and Atul Prakash. Security analysis of emerging smart home applications. In 2016 IEEE symposium on security and privacy (SP), pages 636–654. IEEE, 2016.

[23] Earlence Fernandes, Justin Paupore, Amir Rahmati, Daniel Simionato, Mauro Conti, and Atul Prakash. Flowfence: Practical data protection for emerging iot application frameworks. In 25th USENIX security symposium (USENIX Security 16), pages 531–548, 2016.

[24] Julie Haney, Yasemin Acar, and Susanne Furman. " it's the company, the government, you and i": User perceptions of responsibility for smart home privacy and security. In 30th USENIX Security Symposium (USENIX Security 21), 2021.

[25] Julie M Haney, Susanne M Furman, and Yasemin Acar. Smart home security and privacy mitigations: Consumer perceptions, practices, and challenges. In International Conference on Human-Computer Interaction, pages 393–411. Springer, 2020.

[26] Julie M Haney, Susanne M Furman, and Yasemin Acar. User perceptions of smart home privacy and security. 2020.

[27] Sergey Hardock, Ilia Petrov, Robert Gottstein, and Alejandro Buchmann. From in-place updates to in-place appends: Revisiting out-of-place updates on flash. In Proceedings of the 2017 ACM International Conference on Management of Data, pages 1571–1586, 2017.

[28] Grant Ho, Derek Leung, Pratyush Mishra, Ashkan Hosseini, Dawn Song, and David Wagner. Smart locks: Lessons for securing commodity internet of things devices. In Proceedings of the 11th ACM on Asia conference on computer and communications security, pages 461–472, 2016.

[29] Hang Hu, Steve T.K. Jan, Yang Wang, and Gang Wang. Assessing browser-level defense against idn-based phishing. In 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, August 2021.

[30] Deepak Kumar, Kelly Shen, Benton Case, Deepali Garg, Galina Alperovich, Dmitry Kuznetsov, Rajarshi Gupta, and Zakir Durumeric. All things considered: an analysis of iot devices on home networks. In 28th USENIX Security Symposium (USENIX Security 19), pages 1169–1185, 2019.

[31] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW):1–31, 2018.

[32] Jingjing Ren, Daniel J Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In Proceedings of the Internet Measurement Conference, pages 267–279, 2019.

[33] Diego Rivera, Antonio García, María Luisa Martín-Ruiz, Bernardo Alarcos, Juan Ramón Velasco, and Ana Gómez Oliva. Secure communications and protected data for a internet of things smart toy platform. IEEE Internet of Things Journal, 6(2):3785–3795, 2019.

[34] Kaiwen Shen, Chuhan Wang, Minglei Guo, Xiaofeng Zheng, Chaoyi Lu, Baojun Liu, Yuxuan Zhao, Shuang Hao, Haixin Duan, Qingfeng Pan, and Min Yang. Weak links in authentication chains: A large-scale analysis of email sender spoofing attacks. In 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, August 2021.

[35] Michael A. Specter, Sunoo Park, and Matthew Green. Keyforge: Non-attributable email from forward-forgeable signatures. In 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, August 2021.

[36] Milijana Surbatovich, Jassim Aljuraidan, Lujo Bauer, Anupam Das, and Limin Jia. Some recipes can do more than spoil your appetite: Analyzing the security and privacy risks of ifttt recipes. In Proceedings of the 26th International Conference on World Wide Web, pages 1501–1510, 2017.

[37] William J. Tolley, Beau Kujath, Mohammad Taha Khan, Narseo Vallina-Rodriguez, and Jedidiah R. Crandall. Blind in/on-path attacks and applications to vpns. In 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, August 2021.

[38] Blase Ur, Jaeyeon Jung, and Stuart Schechter. Intruders versus intrusiveness: teens' and parents' perspectives on home-entryway surveillance. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 129–139, 2014.

[39] Xiaojie Wang, Zhaolong Ning, MengChu Zhou, Xiping Hu, Lei Wang, Bin Hu, Ricky YK Kwok, and Yi Guo. A privacy-preserving message forwarding framework for opportunistic cloud of things. IEEE Internet of Things Journal, 5(6):5281–5295, 2018.

[40] Haitao Xu, Fengyuan Xu, and Bo Chen. Internet protocol cameras with no password protection: An empirical investigation. In International Conference on Passive and Active Network Measurement, pages 47–59. Springer, 2018.

[41] Hyunwoo Yu, Jaemin Lim, Kiyeon Kim, and Suk-Bok Lee. Pinto: enabling video privacy for commodity iot cameras. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pages 1089–1101, 2018.

[42] Li Zhang, Wenming Wang, Yuan Tan, Xunhu Sun, Zhi Sun, and Yang Yang. Design and implementation of a fine-grained nand flash programmer. In 2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, pages 257–261. IEEE, 2012.

[43] Ke Zhou, Shaofu Hu, Ping Huang, and Yuhong Zhao. Lx-ssd: Enhancing the lifespan of nand flash-based memory via recycling invalid pages. In Proc. 33rd Int. Conf. Massive Storage Syst. Technol.(MSST), pages 1–13, 2017.

[44] Wei Zhou, Yan Jia, Yao Yao, Lipeng Zhu, Le Guan, Yuhang Mao, Peng Liu, and Yuqing Zhang. Discovering and understanding the

security hazards in the interactions between iot devices, mobile apps, and clouds on smart home platforms. In 28th USENIX Security Symposium (USENIX Security 19), pages 1133–1150, 2019.

Appendix A

A. Distribution of the Surveyed Users

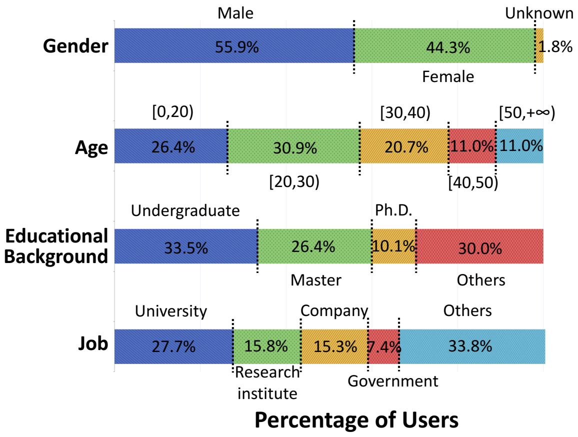

Fig. 9: The distribution of the surveyed users.

In this paper, we conduct a user study (described in §III-A) to understand the user awareness of the data disposal methods and the security risks caused by improper disposal. Our study attempts to understand different user awareness from various perspectives. Thus, the surveyed participants cover 321 participants with different ages, professions, educational backgrounds, genders, and regions to ensure a balanced assessment.

We show the distribution of the surveyed users in Figure 9, in which we mainly present the following four aspects that may highly influence our user study. (1) The distribution of the genders of the surveyed users is almost balanced. Specifically, 55.9% and 44.3% of the surveyed users are male and female, respectively (the genders of the remaining users are unknown). (2) Considering that users of different ages may have different experiences and understandings of data disposal with used IoT devices, we survey various users of different ages. However, IoT devices are currently more prevalent in the youth, and many aged users are not familiar with IoT devices such as smart cameras, printers, etc. Therefore, this user study includes more younger people. (3) We also survey users with different educational backgrounds, which is used to understand how people dispose of used IoT devices related to their educational backgrounds. For instance, do users with higher educational backgrounds dispose of used IoT devices more cautiously? However, our user study shows that most of them may neglect the provided disposal methods regardless of their educational backgrounds. (4) Finally, we present the distribution of participants with various jobs. Our analysis demonstrates that how users dispose of the used devices is barely related to their jobs.

```
1 # /usr/lib/lua/xiaoqiang/module/XQDDNS.lua
2 function editDdns(username, password, domain)
3     local uci = require(”luci.model.uci”).cursor()
4     uci:set(”ddns”, ”server”, ”username”, username)
5     uci:set(”ddns”, ”server”, ”password”, password)
6     uci:set(”ddns”, ”server”, ”domain”, domain)
```

Fig. 10: An example of function encapsulation.

```
1 # /usr/lib/lua/luci/controller/admin/wlextend.lua
2 function wirTrial()
3   local ssid = luci.http.formvalue(”wir_ssid”)
4   ssid = (string.gsub(ssid, ”\\\””, ”\””))
5   local authmode = luci.http.formvalue(”safeSelect”)
6   if authmode == ”WPA2PSK” then
7     encryp = luci.http.formvalue(”encSelect”)
8   apcli.config_apcli(ssid,authmode,encryp);
```

Fig. 11: An example of closed-source sink API usage.

B. A Motivating Example of Sensitive Data Analysis

Figure 2 shows an example of sensitive data in a router and the corresponding data collection code in the firmware. In this example, suppose a user uses the DDNS service in the router, the router stores the user's DDNS service information in a configuration file (shown in Figure 2a). Note that the content of this file is generated during the usage of the router. Thus, detecting the DDNS service information without obtaining this router is impractical. Fortunately, the collection code (shown in Figure 2b) of the DDNS service information resides in the router's firmware. Thus, we can infer the store behavior shown in Figure 2a by detecting the collection behavior shown in Figure 2b.

C. Two-layer API Inferring

First, many vendor-defined APIs are implemented by encapsulating library APIs. We can infer them based on the library APIs they encapsulate. Specifically, we prepare a list of widely-used library SAPIs (shown in Table IX). For each API f in this list, we also record its parameter p that reflects the source/sink data. Then, we collect the calls of each f in the firmware and perform data-flow analysis for each call to examine whether p comes from the caller's parameter. If so, we infer that the caller is a SAPI.

For example, as shown in Figure 10, "editDdns" is a vendor-defined API. This API is implemented by encapsulating a library sink API "uci:set". Since the sink data "username" in line 4 is data-dependent on the parameters in line 1, we infer that "editDdns" is a sink API.

Second, we observe that vendors tend to develop vendor-defined APIs for sink/source usage rather than other usages of the source/sink data. Thus, we can infer SAPIs by analyzing the usage of source/sink data. For example, as shown in Figure 11, we already know "ssid" in line 3 is a source data obtained by a predefined source API. In this case, we want to identify the corresponding sink API. "ssid" is used in lines 4 and 8. "gsub" in line 4 is a predefined known common API (shown in Table IX) and "apcli.config_apcli" is an unknown API. Thus, we infer that "apcli.config_apcli" is a vendor-defined sink API.

TABLE IX: The list of predefined APIs.

API	Type	API	Type
LuciHttp.formvalue	Source	lower	Common
http.formvalue	Source	lshift	Common
http.formvaluetabl	Source	match	Common
cursor.set	Sink	mkdir	Common
cursor.section	Sink	movelist	Common
cursor.set_list	Sink	or	Common
cursor.tset	Sink	pairs	Common
build_url	Common	parse	Common
checkwebauth	Common	pcdata	Common
confirm	Common	poll	Common
contains	Common	printf	Common
date	Common	read	Common
debug	Common	redirect	Common
decode	Common	remove	Common
dsp	Common	render	Common
entry	Common	rshift	Common
export	Common	search	Common
find	Common	session_retrieve	Common
formvalue	Common	sub	Common
get_all	Common	tonumber	Common
getiwinfo	Common	ubus	Common
get_wifidev	Common	ubus_state_to_http	Common
gsub	Common	upper	Common
if	Common	urlencode	Common
imatch	Common	validator	Common
install	Common	write_json	Common
ipairs	Common		

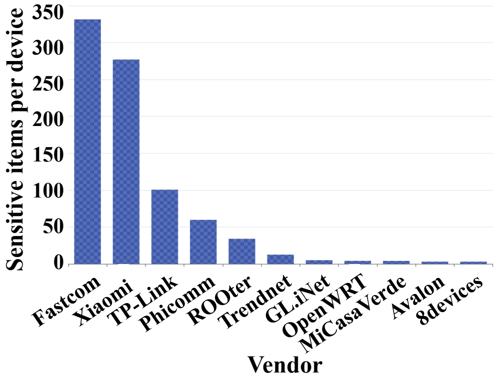

Fig. 12: The distribution of sensitive items in each device from different vendors.

D. More Investigation on the Distribution of Sensitive Information

Collections of user portraits. We found over 300 collections of user portraits, including browsing history, security settings, and pictures. For instance, 10 IP cameras collect the physical security settings, such as whether the user enables motion detection. Besides, we also find user portraits in forensic analysis on real devices. For instance, we found that a set-top box stores the profile photo of a third-party account. Specifically, this device stores a URL that is linked to the photo. Anyone who obtains this URL can access it without any authentication.

How user data collections distribute among different IoT vendors? Figure 12 shows that user data collections vary significantly in different IoT vendors. `Fastcom`, `Xiaomi`, and `TP-Link` are the top three vendors that store the most sensitive data. Specifically, on average, each device from `Fastcom`, `Xiaomi`, and `TP-Link` collects 331, 277, and 100 sensitive items, respectively.

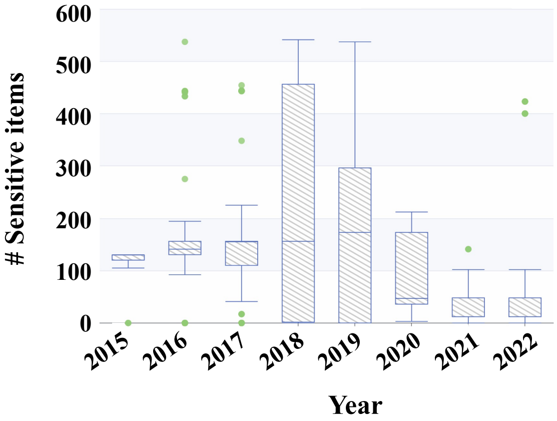

Fig. 13: User data collections over the years.

How user data collections distribute in different kinds of devices? We also investigated the distribution of sensitive items in different kinds of devices. The results demonstrate that network equipment devices tend to store more sensitive data than other types of devices. Specifically, the firmware of network node devices collects 36.89 sensitive items on average. By contrast, the firmware of other devices collects 16.73 sensitive items on average. Our empirical study on real devices also reveals the same results. For example, each network node device average contains 2.12 sensitive items related to third-party accounts. By contrast, each surveillance device, media device, and work device contains 1.33, 1.00, and 1.67 sensitive items related to third-party accounts on average, respectively.

What is the trend of user data collection in IoT firmware over time? Next, we investigate the progression of user data collection over time. First, an integral preparation is determining the release time of each firmware. We achieve this goal by collecting the release time reported on the website of each firmware manually. However, many vendors do not publish the release time of their firmware on the website. Besides, this job is quite time-consuming. Thus, we spend one day collecting the release time of the firmware images in Dataset-1. Finally, we obtain the release time of 427 firmware images and group them by their release year.

Figure 13 shows the user data collection progressed over the years for these images. Overall, user data collections show a sign of increase before 2019 and decrease after that. The explanation for the increment is that IoT devices tend to provide users with more utilities. Thus, over time, they collect more user data, such as third-party accounts. However, user data collections show a sign of a decrease after 2019, which conflicts with the explanation. The reason is that more and more IoT devices adopt cloud services. As reported in previous works, IoT devices upload a large amount of user data to cloud servers [17]. Thus, more and more user data are stored remotely in the cloud servers instead of locally in the devices.

Although data collection by firmware has decreased in recent years, we believe data disposal of IoT devices is still necessary.

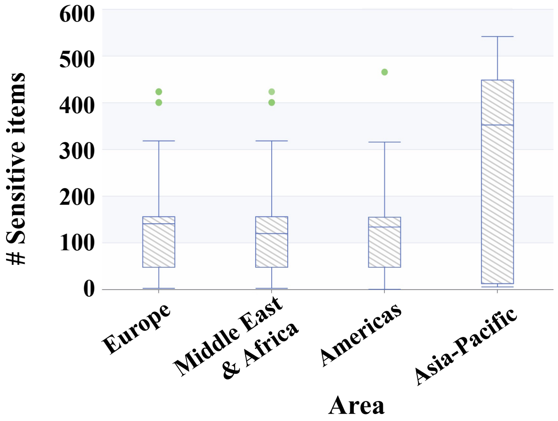

Fig. 14: User data collections distribute in different geographical regions.

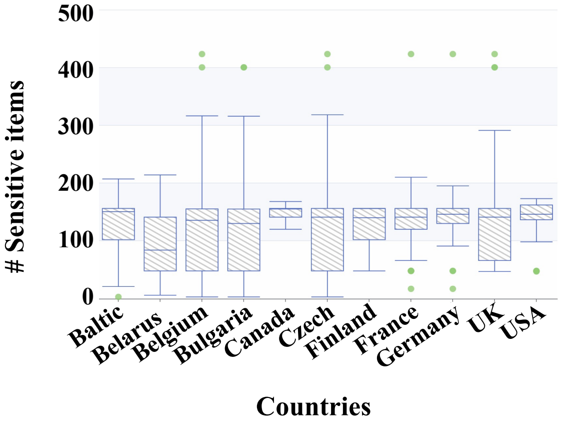

Fig. 15: User data collections in different countries.

First, the latest firmware images released in recent years still collect various sensitive items. The leakage of these items can result in severe consequences, as discussed in §V-B. Second, with the rapid development of IoT, device re-using has become more and more common nowadays. Improperly disposing of used devices can leak a large amount of sensitive data.

How are user data collections distributed in different geographical regions? Similar to release time, we also collect the sales district of the firmware images in Dataset-1 from the websites. Finally, we obtain the sales district of 567 firmware images and group them by their sales district. As shown in Figure 14, For Europe, Middle East & Africa, and the Americas, no significant difference can be statistically observed. However, firmware images from the Asia-Pacific area collect more user data than firmware from other areas. We also conduct a more thorough analysis of user data collections distributed in different countries. The results reveal that user data collection is universal in different regions. Furthermore, we group the 567 firmware images by their sales countries. As shown in Figure 15, the median number of user data collections in most countries is more than 100. The results again reveal that user data collection is universal in different regions.

E. Best Practice Information

We provide best practice information based on our observations to help IoT vendors assure data privacy. (1) Data encryption. We observed that several devices encrypt the sensitive data. Thus, we failed to recover sensitive data from them. For example, one device from `Hiwifi` and two devices from `TP-Link` (whose IDs are 12, 27, and 29) encrypt the admin password. (2) Firmware encryption. We found that a PLC device encrypts its firmware. Thus, we cannot recover sensitive data from it. (3) Flash-level erasing. We observed that several devices delete sensitive files by flash-level erasing. For example, a `Phicomm` router performs flash-level erasing for soft resetting. Thus, it can effectively erase sensitive data.