An Empirical Analysis of Smart Connected Home Data

Joseph Bugeja(✉), Andreas Jacobsson, and Paul Davidsson

Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden
{joseph.bugeja,andreas.jacobsson,paul.davidsson}@mau.com

Abstract. The increasing presence of heterogeneous Internet of Things devices inside the home brings with it added convenience and value to the householders. At the same time, these devices tend to be Internet-connected and continuously monitor and collect data about the residents and their daily lifestyle activities. Such data can be of a sensitive nature, given that the house is the place where privacy is naturally expected. To gain insight into this state of affairs, we empirically investigate the privacy policies of 87 different categories of commercial smart home devices in terms of data being collected. This is done using a combination of manual and data mining techniques. The overall contribution of this work is a model that identifies and categorizes smart connected home data in terms of its collection mode, collection method, and collection phase. Our findings bring up several implications for smart connected home privacy, which include the need for better security controls to safeguard the privacy of the householders.

Keywords: Smart home · IoT · Data model · Privacy policies

1 Introduction

An increasing number of consumers install Internet of Things (IoT) devices in their homes. This benefits the householders offering an improved ability to control and automate relevant aspects of their house and daily home chores. According to Gartner the number of IoT devices is expected to exceed 20 billion by 2020[1]. IoT consumer devices found in private homes include learning thermostats, energy tracking switches, IP-based video doorbells, smart speakers, and more. These devices tend to use embedded sensors and the Internet to collect and exchange data with each other and their users, integrating the digital with the physical environment inside the home.

As these IoT devices become more widespread in homes so is the volume, granularity, and diversity of data collected by these devices. For instance, smart televisions, may be equipped with microphones and cameras allowing for interaction through voice commands and gestures; and a smart thermostat may capture the temperature, humidity, and activities inside the home to predict and adjust the room temperature automatically. While this brings added convenience and value to the householders, the home being the

[1] https://www.gartner.com/newsroom/id/3165317 [accessed May 06, 2018].

D. Georgakopoulos and L.-J. Zhang (Eds.): ICIOT 2018, LNCS 10972, pp. 134–149, 2018.
https://doi.org/10.1007/978-3-319-94370-1_10

natural place where private conservations are held, and where there is an implicit trust between the householders and their house, makes the investigation of data being collected by smart home devices fundamental in understanding privacy concerns. Furthermore, this is useful to realize what is at stake if a device is compromised [1].

Researchers have studied the privacy challenges of smart connected homes and proposed different mitigations to protect user privacy (e.g., [2]), but no study has looked in detail into the actual data collection practices of commercial smart home manufacturers. To different extents, this limits the applicability of previous studies when it comes to applying them to the complexity of real-world setups. Such setups tend to involve heterogeneous devices and data, ranging from low-power devices to high-performance nodes supporting multiple input and output data sources. The closest research work in this regard, concern mostly technical studies that have investigated data exfiltration, oftentimes indirectly. This is typically done by assessing and exploiting vulnerabilities of smart home devices through experiments. While experiments are likely to lead to more accurate results, the number of investigated devices tends to be relatively small (e.g., laboratory setup with 7 device types [3]), is focused on a subset of devices (e.g., web cameras [4]), and targeting specific data states (e.g., concerning stored data [5]). Our goal in this work is to identify all the main categories of data that are collected by smart home systems. This is needed to ground the conversation especially about smart home privacy with empirical evidence on data collected.

In conducting this study, we analyze the privacy policies concerning 87 different type of devices issued from 64 manufacturers of commercial smart home devices. Each manufacturer should have a privacy policy or a similar document detailing how data are captured by a device [6]. Policies are investigated using a hybrid approach combining manual analysis with data mining techniques. Overall, the main contributions of this work are: (i) identification of data types being collected by a smart connected home system; (ii) a categorization of smart connected home data types according to their data source; and (iii) a data model grouping the different data types according to their associated data collection mode, collection method, and collection phase. Our findings bring up several implications for smart home privacy, which include the need for better security controls to safeguard the privacy of the householders.

The rest of this paper is structured as follows. In Sect. 2, we describe data flows in a smart connected system in a generic way. Next, in Sect. 3, we formalize the specification of privacy policies. Following that, in Sect. 4, we review and group existing work related to privacy policy analysis according to the adopted approach. The utilized research design methodology is discussed in Sect. 5. Next, the policy analysis results, identified data types, application of the data categorization to the analyzed policies, and the data collection model are presented in Sect. 6. Finally, in Sect. 7, we discuss some implications of our findings, and conclude this paper.

2 Smart Connected Home Data Flows

A smart connected home system is composed of a number of physical and virtual entities, and data flows occurring between them. At an abstract level, a smart home system is made of three entities:

1. *Smart device:* Hardware components such as Internet-connected household devices, networked appliances, or wearable technologies. These collect data about the householders and the environment and use it to communicate with other devices and users. An example of a smart device is Amazon Echo – an 'intelligent' smart speaker – that uses voice detection technology to detect and respond to tasks such as streaming music. Smart devices may feature integrated sensors, actuators, and processors.
2. *Service:* Software components such as mobile applications, web applications, and Application Program Interfaces (APIs). These use, retain or transfer data from smart devices to implement the householders' desired goals, e.g., enhanced security and safety. Smart home services are commonly deployed over a cloud infrastructure, gateway devices, or as native services inside a device.
3. *User:* Householders, service providers, network providers, and other stakeholders, that together create and enable the smart home ecosystem. The householders tend to be the main subject of data collection by the smart devices surrounding them.

Any smart connected home system involves the exchange of data to enable its function. This is done through a communication infrastructure, consisting of protocols, typically IP-based, and networking components such as router, bridges, and hubs. This data can range from non-personal data (e.g., room temperature), personal data (e.g., resident names), sensitive data (e.g., health conditions), and more.

Based on the work of Ziegeldorf et al. [7] but adapted for the smart connected home environment, we identify four main data flows occurring between the entities. We refer to these flows as: interaction, collection, dissemination, and presentation (see Fig. 1). In the interaction phase, the user, actively or passively, interacts with the smart device. Here, as an example, the householder might switch on an IP-enabled lightbulb. In the collection phase, the smart device, e.g., the connected-lightbulb, gathers the information, e.g., data about the bulb state (on/off), and delegates that to the corresponding service, e.g., embedded software in the connected-bulb. Services then process the data to provide desired function (e.g., turning on the room light) and may initiate further actions on their own. Once that is done, in the dissemination phase the service relays the information towards the data subject or a third-party. This is mediated through a smart device, which can be the same as the one used for interaction phase or a different one. Such device then implements an actuation response (e.g., adjusting the room light level) or provides a notification (e.g., mobile notification displaying the current room light level) to the user. We refer to the final stage, as the presentation phase.

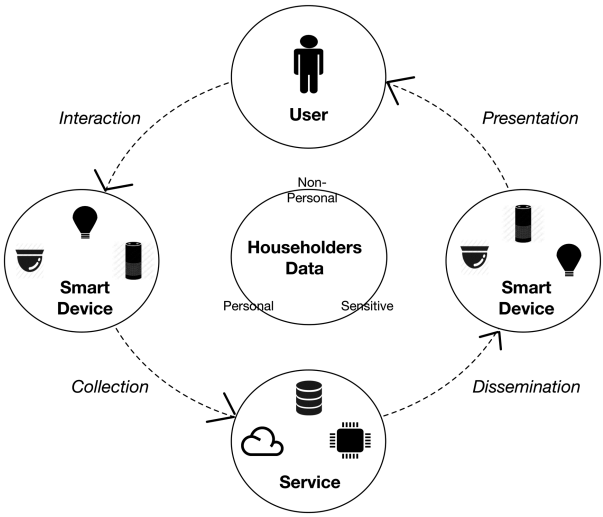

Fig. 1. A smart connected home system entities, data, and data flows. Data in a smart home system tends to initially flow from the user (*interaction*), typically the householder. Consequently, this is received (*collection*) by the smart home device, and then processed by a software service(s). In turn, this is used to send (*dissemination*) a notification to the user (*presentation*) or to change an environment parameter, e.g., temperature.

Although not enforced by law, companies, especially those operating in the US, are required to post notices of data practices, especially concerning their privacy management procedures [8]. In the US, regulators such as data protection authorities or the US Federal Trade Commission leverage companies' privacy policies to assess and enforce regulatory compliance [9]. In the past, privacy policies were mostly the target for web-based systems, however nowadays especially with new regulations (e.g., the European General Data Protection Regulation[2]) these are becoming key also for IoT-based applications, such as smart connected homes. This is especially to meet the individual's right to restrict data processing.

3 Privacy Policies

Privacy policies aim to answer important questions, such as: what data are collected, for what purposes is this data used, and with whom is the data shared with, amongst other things. Based on the privacy requirements specification proposed by Breaux et al. [10] but applying it to the smart connected home, at an abstract level, policy requirements can be formally specified as: <*S, Ac, A, D, P*> where:

- *S,* the policy scope,
- *Ac*, a set of actors among whom data are shared,

[2] https://www.eugdpr.org [accessed May 06, 2018].

- A, a set of actions that are performed on the data,
- D, a set of data elements on which actions are performed,
- P, a set of purposes for which data may be acted upon.

In our case, $S \in \{device, website, service, all\}$, indicating that a privacy policy may cover the smart device, webpage a user may interact with, services such as mobile applications, and all of these. Ac indicates data recipients, typically the service providers. $A \in \{interaction, collection, dissemination, presentation\}$, corresponding to the data flows defined earlier in Sect. 2. D represents data about the smart home environment. P specifies the data collection and usage reasons as provided by the vendor.

For the scope of this paper, we are mainly interested in finding D where $S = \{all\}$, and $A = \{collection\}$. This is the first logical phase where data from the user is received by a smart home device. In a sense, this is arguably the entry point whereby which the householders' privacy can get compromised.

4 Related Work

Despite efforts to make privacy policies machine-readable, e.g., with P3P, or formalize them, e.g., using EPAL, policies are oftentimes written in natural language [11]. Given this, we observe three distinct approaches concerning the extraction of data types and semantics from privacy policies: Natural Language Processing (NLP), machine learning (ML), and hybrid approaches.

4.1 NLP Approaches

Alohaly et al. [12] used NLP techniques to extract data types from privacy policies. This was done by first locating text fragments that were relevant to the data collection practice. Then, noun phrases were extracted from the retrieved sections keeping only those that were present in the Information Type Lexicon [13]. These represented actual data types. The Information Type Lexicon is a dictionary based on privacy policy annotations. This was created through manual, human annotations, and with the help of an entity extractor based on part-of-speech (POS) tagging. The entity extractor succeeds at finding 92% annotations (from 15 policies). In our work, we use some of the data elements and categories present in the Information Type Lexicon, as guidance for retrieving and grouping potential data items. Nonetheless, its main limitation is that it has a saturation limit of 31–78%. Therefore, its overall effectiveness is limited for previously unseen policies or domains.

Bhatia et al. [14] used constituency parse trees from POS tagged sentences to automate the extraction of hyponyms found in privacy policies. Hyponyms are specific phrases that are sub-ordinate to another more general phrase, e.g., a GPS location is a kind of real-time location. The authors manually identified a set of hyponyms among 15 privacy policies and then formalized the patterns using a tree regular expression language (Tregex). An important conclusion of this study is that only 17% of the data types found in the dataset appeared in WordNet (a popular lexical database). Alas, this study involved policies that were not connected to the IoT domain.

Costante et al. [15] leveraged Information Extraction (IE) techniques to extract a website's data collection practices from its privacy policy. The proposed approach leverages the semantics of the text leading to consistently accurate results. However, the extraction rules are manually created for specific scenarios. This thereby limits their reusability, e.g., for identifying data items collected by a smart home system.

4.2 ML Approaches

Costante et al. [16] proposed a supervised learning approach to determine which data practice categories (e.g., data sharing) are covered in a privacy policy. The privacy categories are extracted from privacy regulations, while text classification and machine learning techniques are used to verify the categories that are covered by a policy. This study is however focusing on the evaluation of privacy policies for their completeness. This is different from our study, where we are interested in the extraction of data types.

Liu et al. [17] explores the problem of aligning or grouping segments of policies based on the privacy issues they address. This is done by using clustering and hidden Markov model based alignment methods. In this study a corpus of 1,010 collected policies was used. One conclusion of this work is that unsupervised methods reach far better agreement with the consensus of crowd workers than originally estimated. Despite this, the applicability of this study for our purposes is restricted especially as it is not directly the scope of the cited study to extract data types.

4.3 Hybrid Approaches

Zimmeck and Bellovin [18] introduced an architecture for analyzing privacy policies using rule and ML classification with crowdsourcing. Policies were classified into different categories as derived from privacy legislation: collection, encryption, ad tracking, limited retention, profiling, and ad disclosure. Here, features (similar to data types) are extracted using bigrams represented as regular expressions. While the feature extraction part is interesting for our study, the main focus of the referenced work is on classifying policies for privacy factors. In our work, we employ a similar data extraction method but also consider unigrams as potential data types.

The Usable Privacy Project [19] uses natural language processing, machine learning, privacy preference modeling, crowdsourcing, and formal methods to semi-automatically annotate privacy policies. This project has annotated over 7,000 privacy policies using machine learning classifiers. Additionally, 115 privacy policies (the OPP-115 Privacy Policy Corpus) were manually annotated by law students. Given the amount of policies reviewed, achieved results, and online availability of this project, we utilize the obtained data categorization here as guidance for creating a data categorization for the smart connected home.

4.4 Main Observations

In reviewing the existing work, we draw three main observations. First, we observe that the majority of the existing work involves to different extents manual analysis as part

of the preprocessing stage of policies. This facilitates tackling problems that remain hard to solve with automated methods by leveraging human intelligence especially to resolve ambiguities in parsing policies. Typically, this utilizes crowdsourcing, and commonly uses the Amazon Mechanical Turk platform[3]. Second, we note limitations, especially in the pure NLP approaches, when it comes to reusing the data types, obtained lexicon, and data extraction rules to other domains that were not initially considered. Third, we observe that the most cited, current, and rather generic approaches tend to use a hybrid approach leveraging manual and semi-automated methods. However, it is interesting to note that in the hybrid but also in the ML approaches the extraction of data types is usually not the main focus of the underlying study.

For these reasons, we adopt a hybrid approach combining manual with data mining to identify and categorize collected data types in a smart connected home system. However, different from previous studies we focus on manufacturers of smart connected home systems, and leverage both unigrams and bigrams to extract these.

5 Research Methodology

In this section we describe the research approach that was adopted to identify and categorize the data types found in smart connected home privacy policies.

5.1 Data Collection

As a precursor for identifying privacy policies, an IoT product collection platform was consulted first to identify smart home devices and their corresponding manufacturer. In this regard, there are two main databases, namely, *iotlist.co* (IoTList) and *smarthomedb.com* (SmartHomeDB) that can be used.

IoTList is an online directory listing IoT devices available on the market. SmartHomeDB is an open community-supported database focusing specifically on smart connected home devices. In our case, we relied on SmartHomeDB, as in comparison to IoTList, this database targets consumer devices intended for the smart home, and also has a product ranking system. The ranking system was used to select the top reviewed devices.

At the time of our evaluation (as of January 2018) SmartHomeDB identified a total of 87 different categories of devices (e.g., lighting system, scale, voice command device, etc.). For each category, the most reviewed device within that was selected. In case, a category contained multiple devices that had the same ranking the most specific device was selected. Example, if two devices within the '*Sensor: Door*' category had the same number of reviews but one featured multiple sensor types, and one functioned as a standalone, then the latter was chosen. This was done to avoid having technologies already covered by other devices affecting the data categorization.

[3] https://www.mturk.com [accessed May 06, 2018].

After the 87 device types were logged, a Python script that interfaced with *Google Search API*[4] was created to retrieve and download the privacy policy of each device. Here, the manufacturer name appended to the specific product name were used as search queries. After the corresponding policy was downloaded it was automatically converted from HTML or PDF format to text in preparation for the data preprocessing stage.

5.2 Data Preprocessing

After the policies were downloaded, the resultant corpus was first manually inspected. Here, the policies were investigated and sorted according to their scope. In our case, we were interested in policies that covered the entire spectrum of possible input data sources, i.e., policies where $S = \{all\}$ (cf. Sect. 3).

At this stage, policies were also inspected to ensure that they identified the collection phase. In case, policies were only not available in English then these were translated using *Google Translate*. Furthermore, if a manufacturer produced more than one device and that was covered by the same policy (or a supplement in the same policy) then only one version of the policy was kept. Since policies covered other actions besides the collection phase, we manually removed other sections from the policies not pertaining to that.

Following the manual preprocessing stage, the policy was further preprocessed in memory using *R*. Here, numbers, punctuation marks, extraneous white spaces, and stop words (using *tm* package) were removed from the corpus as these did not contribute to data types. Additionally, text was converted to lower case, and stemmed using Porter stemming algorithm (using *SnowballC* package).

5.3 Data Type Identification and Categorization

To identify possible data types we transformed the preprocessed policy documents (the corpus) into a term-document matrix. Essentially, a term document matrix is a two-dimensional matrix whose rows represent the terms and columns represent the documents. In our case, the terms represent possible data types and the documents represent privacy policies.

As a method for locating terms in privacy policies, we employed a n-gram tokenizer (using *RWeka* package) to find both unigrams and bigrams. Unigrams are needed to detect data types, such as “gps”, and bigrams are needed to find other instances, e.g., “ip address” or data categories, e.g., “contact information”.

Consequently, the resultant list was saved and scanned manually for data types and possible categories. In doing so, and in grouping the different elements, we were guided by the “Information Type Lexicon” [13] and the “Usable Privacy Project” [19].

[4] https://github.com/abenassi/Google-Search-API [accessed May 06, 2018].

6 Results

In this section, we present the results obtained after executing the research methodology described in the previous section.

6.1 Privacy Policy Analysis

From the collected dataset, the number of reviews varied significantly with mean ($\mu = 2{,}439$) and standard deviation ($\sigma = 6{,}098$). Effectively, the most reviewed device (34,372 reviews) was a media player, and the least reviewed devices consisted mostly of sensor devices.

There were multiple 12 manufacturers producing multiple device types ranging from 2 up to 4 most reviewed products. In total, covering 87 different devices types, there were 64 manufacturers, out of which 3 had no privacy policy, and one had a policy that was available only in Chinese.

Different policies differed in terms of their scope. In Table 1, we summarize the different types of policies available from different smart connected home manufacturers, including the number of documents covering each scope, and the number of device types covered by each policy scope.

Table 1. Distribution of smart connected home privacy policies.

Privacy policy scope	Corpus size	Number of device types
Website	27	34
Service	4	4
Website and service	7	7
Website, service, and device	23	39

All policies contained information about data collected, however since we were interested in the entire smart connected home environment then we focused the rest of this work on the policies that covered all components (i.e., the website, service, and device). In total, this included 39 different device types manufactured by 23 different manufacturers.

6.2 Smart Connected Home Data Categorization

The results of our analysis indicate that there are two main entities that are subject to data collection by smart home stakeholders:

User. This represents the householders who are the end-users of the smart home system. However, this also may include visitors, guests, or other physical entities who are interacting directly or indirectly with the smart home system. Data categories in this dimension include:

- *Contact information.* Information that can be used to communicate or correspond with users. Examples of data types here are: email, phone number, address, contact list, and friend user IDs. Typically, this information is captured during system setup.
- *Personal and account details.* Data that can be used to identify and authenticate a user. This includes names, login identifiers, social networking services account information, passwords, security codes, profile pictures, job titles, gender, birth date, and body metrics data (e.g., weight). All the surveyed device types require this type of information in order to use the services being offered by the smart home system.
- *User activity data.* Data that corresponds to a measurement of user physical activities or interactions. Examples found in this category include: interaction data, e.g., voice commands, browser data, webpages accessed, service used, features accessed, activity time and duration, and inputted search query terms. The type of user activity data captured and utilized is dependent of the type of device.
- *Configuration settings.* Customization or personalization information related to the smart home system. Examples of data types here include: device and service specific settings, operating schedules, contact preferences, cookie settings, and language preferences. Commonly, this data are stored on the mobile applications but may also be stored in the device and cloud infrastructure.
- *Location information.* Position data related to the current geographic location of the user. In a smart home system this is provided either directly by the user or is derived automatically, e.g., from the IP address, or through location-based technologies (e.g., by capturing a smartphone GPS' signal). Such data may be used, e.g., to provide location-based services, e.g., weather forecasts.
- *Financial information.* Payment related information required to purchase additional services/products, e.g., extra cloud storage space. Data types here, include credit card data such as card numbers, expiration dates, and related security codes. Commonly, this information is captured during the smart home setup phase.
- *User-generated content.* Data submitted voluntarily by the user, typically as part of a survey or in relation to technical issues. Examples of data types here include: feedback, opinions, reviews, comments, uploaded files, interests, demographic data, and other information furnished by the user e.g., for due diligence process, technical support, and marketing research. Typically, this data are collected over the web interface.
- *Offline data.* Data that may be used to identity the user but is captured indirectly, e.g., by visiting a service provider premises or by talking to service personnel. Examples of data types captured here include: CCTV footage, phone conversations, and offline interactions.

Device. IoT devices that are present inside the home environment (e.g., IP camera, smart TV, network-enabled washing machine), and end-user devices such as smartphones, tablets, and smart watches that can be used to monitor and control the smart home system. Data categories in this dimension include:

- *Device information.* Technical information describing a hardware device. Data types here include: device identifiers (e.g., IP address, MAC address, and IMEI number), performance information, network/connectivity-related information (e.g., Wi-Fi

status and Bluetooth data), firmware and software versions, sensor status, battery charge level, diagnostic information, and consumption data (e.g., energy consumed).
- *Environmental data.* Technical data describing the smart home environment including its surroundings. Typically, data types here feature parameters that are captured by sensors. Examples of these found in the reviewed policies include: motion, humidity, ambient light, CO2 concentration, and rain level.

6.3 Smart Connected Home Devices and Data

As an application of the previously defined data categorization to the smart connected home, in this section we review each of the 39 devices identified in Sect. 6.1 in terms of their collected data. The results of this mapping is shown in Table 2.

Table 2. Matrix showing different device types alongside their collected data types: CI, contact information; DI, device information; PAD, personal and account details; UAD, user activity data; CS, configuration settings; LI, location information; FI, financial information; ED, environmental data; UGC, user-generated content; OD, offline data.

Device type	CI	DI	PAD	UAD	CS	LI	FI	ED	UGC	OD
Music player, gateway/hub	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Door bell, audio speaker, TV, irrigation controller	✓	✓	✓	✓	✓	✓	✓	✓	✓	–
Scale, plug, light switch, light bulb, wireless signal extender	✓	✓	✓	✓	✓	✓	✓	–	✓	–
Switch, power outlet, oven, clothes dryer	✓	✓	✓	✓	✓	✓	–	✓	–	–
Vacuum cleaner, floor mopper, floor scrubber, gutter cleaner	✓	✓	✓	✓	✓	✓	✓	–	–	–
Tracker	✓	✓	✓	✓	✓	✓	–	✓	✓	–
Blood pressure monitor, temperature sensor	✓	✓	✓	✓	✓	✓	–	✓	–	–
Remote control, light strip, cooker	✓	✓	✓	✓	–	✓	✓	✓	✓	✓
Air quality sensor, rain sensor, CO2 sensor, wind speed sensor	✓	✓	✓	✓	–	✓	✓	–	✓	–
Siren	✓	✓	✓	✓	–	✓	✓	✓	✓	–
Cloud camera	✓	✓	✓	✓	–	–	✓	✓	✓	–
Shower head water meter	✓	✓	✓	–	✓	✓	✓	✓	✓	–
Bar code scanner	✓	✓	✓	✓	✓	✓	✓	✓	–	–
Thermostat, Smoke detector	✓	✓	✓	✓	✓	✓	–	✓	–	–
Cloud camera	✓	✓	✓	✓	–	–	✓	✓	✓	–
Door lock	✓	✓	✓	✓	–	✓	✓	–	–	–
Accelerometer sensor	–	–	✓	✓	–	✓	✓	✓	✓	–

"✓" = data type is captured; "—" = data type is not specified to be collected.

Here, it can be noted that the devices types that collect all data categories are a music player, and a gateway/hub device. On the contrary, the device types that collect the least

amount of data are a door lock and accelerometer sensor. All investigated device types require some personal and account details from users.

6.4 Smart Connected Home Data Collection Model

It can be observed that the different data categories are collected either explicitly, i.e., the user manually provides the information directly to the system, or implicitly by the system. In the latter case, the data are collected automatically without involving the end-user's explicit awareness and consent. The data source for the explicit collection mode tends to be the user, in particular the householders, whereas the data source for the implicit collection mode tends to be the smart home device. In Table 3, we map each data category to its corresponding collection mode, and identify methods and phases during which this data are captured.

Table 3. Smart connected home data collection model.

Data category	Collection mode	Collection method	Collection phase
Contact information	Explicit	Website form, service	System setup
Device information	Implicit, explicit	Smart home device, end-user device	System operation, sync process
Personal and account details	Explicit	Website form, service	System setup
User activity data	Implicit, explicit	Sensors, service	System operation, sync process
Configuration settings	Explicit	Website form, cookies, service	System setup
Location information	Implicit, explicit	Smart home device, end-user device	System operation
Financial information	Explicit	Website form, service	Purchase process
Environmental data	Implicit	Sensors	System operation
User-generated content	Explicit	Surveys, feedback form, support ticket	System operation, troubleshooting process
Offline data	Implicit, explicit	Paper, digital	Offline

When it comes to the data provided explicitly by the user this correlates with the interaction phase of a smart connected system. Here, the user is typically setting up or registering a smart home device for the first time or customizing it to cater for new conditions. This type of data are typically provided through a mobile application furnished by the smart home device manufacturer or service provider. In some cases, such input is provided through a website or service managed by the service provider or third-parties (e.g., PayPal). Commonly, this type of data can be opted-out although doing so may sometimes hinder the smart home system performance or stability.

To collect data automatically, a smart home system, tends to employ different technologies depending on the system and input channels being used. For instance, when interacting with a system via the web interface, cookies are typically used as a

mechanism to gather information about the householders' preferences. On the other hand, it is also common to collect information from third-party applications (e.g., Facebook) or from specific APIs. Additionally, smart home devices may employ sensors that automatically collect environmental data. While some data categories, e.g., in the case of cookies, can be opted-out by the user, there are data categories, in particular environmental data that cannot be opted-out.

7 Discussion and Conclusions

The growth and heterogeneity of IoT devices present in the smart connected home raises the importance of an analysis of data being collected by these devices. This is needed to explore what is at stake if a device is compromised, and as a precursor for conducting a privacy impact assessment.

In our study, we have analyzed privacy policies of 64 manufacturers, out of which we identified 23 policies that cover all the smart home components. We observe that no scholarly work has looked specifically into data collection stage of smart home systems. The exception here are a few security vulnerability studies that assess data being disclosed (sometimes indirectly) through lab experiments. While such experiments reveal actual data being collected, rather than declared types (as in the case of policies), they are dependent on the adopted test cases, utilized software/hardware tools, and the physical location of the assessor (or penetration tester). In reality, for many reasons, e.g., costs, time, and expertise required, it is often the case that only certain software components, device types, and data states are targeted. One example is examining built-in webservers embedded in IoT cameras for unauthorized transmission of data. Therefore, this limits the effectiveness of experiments when used as the sole method for identifying data being collected by smart home devices.

Privacy policies have been studied considerably by many researchers over the years as we have mentioned when reviewing the related work. However, we observe that even the most cited publications focus on the investigation of commercial entities operating typically in the traditional web-based systems domain and not on IoT systems such as smart connected homes. While we used some existing work (e.g., [19]), for guidance, we expanded on this domain with data categories that are suitable for investigating IoT-based systems, in particular smart connected homes. For instance, we added the 'Environmental data' to capture information measured by sensors, and 'User activity data' to encapsulate not only online activities, such as browsing websites of service provider, but other interaction channels as well, e.g., voice. Voice input is commonly used to interact with smart home devices.

We have identified 10 different data categories of smart home data, that correspond to data being generated by the user, typically the householder, and data being generated by the device. These categories were empirically derived by analyzing privacy policies through a hybrid approach combing manual analysis with data mining. Our findings reveal that certain data types, in particular 'Device information' and 'Environmental data', are passively and potentially continuously, being collected by smart home devices. A consequence of this is that the system may be automatically monitoring, building

detailed user profiles, and automatically inferring user activities without the householders' awareness. Such activities may include sensitive ones, ranging from indicating whether the residents are away to medical diagnosis of an individual. Data collection may be done even when actions are not initiated or given explicit consent by the user. This is especially as IoT-based systems may leverage data mining or artificial intelligence techniques that automatically learn, adjust their services, perform actions, and automatically reach conclusions based on the collected data [20].

Achieving privacy is an inherent trade-off in IoT systems, because IoT devices cannot provide their services and add value without collecting data. However, since such devices tend to be Internet-connected, have a tendency to be 'always-on', and are present in houses where privacy is naturally expected, this stresses the need for smart home developers to create better data security controls to safeguard the privacy of householders. Likewise, this raises the importance to have better methods for informing the householders of potential risks when purchasing and operating smart home technologies.

Smart home stakeholders may collect certain categories of data, related to both personal and non-personal data. Personal data is essentially data that can be used to identify an individual person; whereas non-personal data do not have the capability (on their own) to identify an individual person. Typically, personal data are collected by smart devices, e.g., to provide adapted responses to a user's current need with the fewest explicit information provided by the user [21]. In this study, we have noted that all surveyed manufacturers collect instances of personal data. This may include body metrics (and other physiological data) that are arguably the most sensitive data type. While one may assume a secure and trustworthy entity processing data according to its privacy policy, entities may be targeted by malicious threat agents, such as hackers. A consequence of this is that private information may be lost, sold to third-parties, and used inappropriately for commercial or malicious purposes [21].

As part of future work, we intend to expand this study to investigate the privacy practices of smart home service providers. One way of doing this is by extracting such information from privacy policies. Another avenue we seek to investigate is to complement this study with a lab experiment. In particular, it would be interesting to investigate some prominent gateways devices. Especially, this is as these tend to be the components; as is also evident in this study; that have access to the most data types, and thus an important point where security should be bolstered. Finally, we plan to develop controls to allow householders to be notified about surrounding IoT devices collecting personal information, and to control these collection practices.

Acknowledgments. This work has been carried out within the research profile "Internet of Things and People", funded by the Knowledge Foundation and Malmö University in collaboration with 10 industrial partners.

References

1. Bugeja, J., Jacobsson, A., Davidsson, P.: On privacy and security challenges in smart connected homes. In: Proceedings of the IEEE Intelligence and Security Informatics Conference (EISIC), pp. 172–175 (2016)
2. Ahlam, A., Laila, B., Slimane, B.: An overview of privacy preserving techniques in smart home wireless sensor networks. In: Proceedings of the IEEE 10th International Conference on Intelligent Systems Theories and Applications (SITA), pp. 1–4 (2015)
3. Apthorpe, N., Reisman, D., Sundaresan, S., Narayanan, A., Feamster, N.: Spying on the Smart Home: Privacy Attacks and Defenses on Encrypted IoT Traffic (2017). arXiv preprint arXiv: 1702.03681
4. Seralathan, Y., Oh, T.T., Jadhav, S., Myers, J., Jeong, J.P., Kim, Y.H., Kim, J.N.: IoT security vulnerability: a case study of a web camera. In: Proceedings of the IEEE 20th International Conference on Advanced Communications Technology (ICACT), pp. 172–177 (2018)
5. Boztas, A., Riethoven, A.R.J., Roeloffs, M.: Smart TV forensics: digital traces on televisions. Digital Invest. **12**, S72–S80 (2015)
6. Anscombe, T.: IoT and Privacy By Design in the Smart Home. https://www.welivesecurity.com/wp-content/uploads/2018/02/ESET_MWC2018_IoT_SmartHome.pdf. Accessed 06 May 2017
7. Ziegeldorf, J.H., Morchon, O.G., Wehrle, K.: Privacy in the Internet of Things: threats and challenges. Secur. Commun. Netw. **7**(12), 2728–2742 (2014)
8. Massey, A.K., Eisenstein, J., Anton, A.I., Swire, P.P.: Automated text mining for requirements analysis of policy documents. In: Proceedings of the Requirements Engineering Conference (RE) (2013)
9. Schaub, F., Balebako, R., Cranor, L.F.: Designing effective privacy notices and controls. IEEE Internet Comput. **21**(3), 70–77 (2017)
10. Breaux, T.D., Hibshi, H., Rao, A.: Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. Requirements Eng. **19**, 281–307 (2013)
11. Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S.M., Reidenberg, J.: Automated analysis of privacy requirements for mobile apps. In: Proceedings of the Network and Distributed System Security (NDSS) Symposium (2017)
12. Alohaly, M., Takabi, H.: Better privacy indicators: a new approach to quantification of privacy policies. In: Proceedings of the WPI SOUPS (2016)
13. Bhatia, J., Breaux, T.D.: Towards an information type lexicon for privacy policies. In: Proceedings of the IEEE Eighth International Workshop on Requirements Engineering and Law (RELAW) (2015)
14. Bhatia, J., Evans, M.C., Wadkar, S., Breaux, T.D.: Automated extraction of regulated information types using hyponymy relations. In: Proceedings of the IEEE 8th International Requirements Engineering Conference Workshops (REW), pp. 19–25 (2016)
15. Costante, E., Hartog, den, J., Petkovic, M.: What websites know about you⋆ privacy policy analysis using information extraction. In: Data Privacy Management and Autonomous Spontaneous Security, pp. 146–159 (2013)
16. Costante, E., Sun, Y., Petkovic, M., Hartog, den, J.: A machine learning solution to assess privacy policy completeness. In: Proceedings of the ACM Workshop on Privacy in the Electronic Society (2012)
17. Liu, F., Ramanath, R., Sadeh, N., Smith, N.A.: A step towards usable privacy policy: automatic alignment of privacy statements. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING) (2014)

18. Zimmeck, S., Bellovin, S.M.: Privee: an architecture for automatically analyzing web privacy policies. In: Proceedings of the USENIX Security Symposium (2014)
19. Sadeh, N., Acquisti, A., Breaux, T.D., Cranor, L.F., McDonald, A.M., Reidenberg, J.R., Smith, N.A., Liu, F., Russell, N.C., Schaub, F., Wilson, S.: The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. Carnegie Mellon University (2013)
20. Zhang, L.-J., Li, C.: Internet of Things Solutions. Services Transactions on Internet of Things (STIOT) **1**, 1–22 (2017)
21. Habegger, B., Hasan, O., Brunie, L., Bennani, N., Kosch, H., Damiani, E.: Personalization vs. privacy in big data analysis. Int. J. Big Data (IJBD) **1**, 25–35 (2014)