Is Your Home Becoming a Spy? A Data-Centered Analysis and Classification of Smart Connected Home Systems

Joseph Bugeja*
joseph.bugeja@mau.se
Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University
Malmö, Sweden

Andreas Jacobsson
andreas.jacobsson@mau.se
Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University
Malmö, Sweden

Paul Davidsson
paul.davidsson@mau.se
Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University
Malmö, Sweden

ABSTRACT

Smart connected home systems bring different privacy challenges to residents. The contribution of this paper is a novel privacy grounded classification of smart connected home systems that is focused on personal data exposure. This classification is built empirically through k-means cluster analysis from the technical specification of 81 commercial Internet of Things (IoT) systems as featured in PrivacyNotIncluded – an online database of consumer IoT systems. The attained classification helps us better understand the privacy implications and what is at stake with different smart connected home systems. Furthermore, we survey the entire spectrum of analyzed systems for their data collection capabilities. Systems were classified into four tiers: app-based accessors, watchers, location harvesters, and listeners, based on the sensing data the systems collect. Our findings indicate that being surveilled inside your home is a realistic threat, particularly, as the majority of the surveyed in-home IoT systems are installed with cameras, microphones, and location trackers. Finally, we identify research directions and suggest some best practices to mitigate the threat of in-house surveillance.

CCS CONCEPTS

• **Social and professional topics** → Privacy policies; **Surveillance**; • **Security and privacy** → *Human and societal aspects of security and privacy*.

KEYWORDS

IoT, smart home, home automation, privacy, unsupervised classification, survey, web mining

ACM Reference Format:
Joseph Bugeja, Andreas Jacobsson, and Paul Davidsson. 2020. Is Your Home Becoming a Spy? A Data-Centered Analysis and Classification of Smart Connected Home Systems. In *10th International Conference on the Internet of Things (IoT 2020), October 06–09, 2020, Malmö, Sweden.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3410992.3411012

IoT 2020, October 06–09, 2020, Malmö, Sweden

ACM ISBN 978-1-4503-8758-3/20/10…$15.00
https://doi.org/10.1145/3410992.3411012

1 INTRODUCTION

A smart connected home system is a residence equipped with connected devices providing the occupants the facility to monitor and control, possibly remotely, different aspects of the home typically using mobile apps [6]. As smart home technologies continue to evolve at a rapid pace, connected devices are becoming more sophisticated, providing benefits beyond simply comfort and convenience, such as energy efficiency, safety and security, and healthcare and fitness support. Despite the positive prospects for the spread of smart home technologies, the adoption of smart connected homes seems uncertain [39]. A major obstacle to adoption is the nature of the technology that gives rise to new concerns about privacy [27].

Smart connected home systems utilize diverse sensing technologies that collect various amounts of data, for example, data about habits, behaviors, and private health details, to provide personalized services to users. The collection and processing of personal and sensitive user data in combination with the increasing deployment of Internet-connected devices in the home expose residents to new privacy risks. Indeed, in 2019, Ren et al. [22] in their experimental analysis of 81 smart connected home systems, discover that all the examined systems expose data to eavesdroppers through plaintext traffic flows. This exposure can allow for eavesdroppers to learn a user's interactions with a device, opening the potential for profiling and other privacy-invasive techniques [22]. Moreover, by attacking a smart connected home system, potential intruders can get private and intimate details about households. The home represents the most personal and protected sphere of our lives, and thus we argue that it is imperative to gain a better understanding of the underlying system characteristics, in particular in terms of data that are being exposed when using smart connected home technologies. This analysis work can lead to more informed means for assessing privacy risks in smart connected home systems.

Similar to the risk assessment model by Strugress et al. [33], we posit that to comprehensively assess privacy risks within a smart connected home we can treat the home as a catalog of data collecting capabilities. Data collecting capabilities are features of interest that are captured directly by the connected device and/or its accompanying mobile apps. Specifically, we focus on cameras, microphones, and location, as the data collecting capabilities. These capabilities are arguably perceived as the most privacy-invasive technologies [10], carry the highest security risk and impact to users [2], and are guarded by most of the current privacy regulations such as the European Union's General Data Protection Regulation (GDPR) [37].

Noting the lack of empirical studies focusing on smart connected home system classification intended for privacy risk analysis, in this paper we aim to provide a deeper understanding of the issue by answering the following research questions:

- How can smart connected home systems be categorized according to their data collection capabilities?
- What are the data collection capabilities of the different smart connected home systems categories?

For developing the categorization, the technical specification of 81 commercial connected systems as featured in PrivacyNotIncluded[1] is programmatically mined, and k-means cluster analysis is used for forming the device categories. The goal of cluster analysis is to form groups of objects so that similar objects are grouped together, or even in nearby clusters, and objects in different groups become as dissimilar as possible [23].

In answering our research questions, we contribute to research on the privacy of smart connected homes by:

- Developing a classification of connected home systems based on their data collection capabilities;
- Conducting a quantitative survey of connected home systems in terms of their data collection capabilities; and
- Identifying research directions and best practices to minimize privacy risks of smart connected home systems.

The remainder of this paper is structured as follows. In Section 2, we summarize the data lifecycle phases of a smart connected home. Next, in Section 3, we provide related work on classifying IoT and smart connected home systems. In Section 4, we present the adopted research method. The experimental results are introduced in Section 5. Next, in Section 6, we discuss the usefulness of the proposed classification and suggest an equation that can be used to measure the data sensitivity of smart connected home systems. Some of the most urgent research topics and best practices are discussed in Section 7. Finally, in Section 8, conclusions are drawn and directions for future work are identified.

2 THE SMART CONNECTED HOME DATA LIFECYCLE

Despite various technical, usability, and form-factor differences, connected systems tend to share similar data lifecycle phases: data generation, data collection, data processing, and data disclosure [40].

Data subjects interact with a connected system at the data generation stage. Typically, this is done through a mobile app that is specific to the smart connected home system. The input, which may consist of personal data, is then received by the system and potentially stored. Collected data can be stored in different places such as in connected devices, gateways, or cloud servers. Next, data are analyzed at the data processing stage. Here, data may be processed at remote backends, such as cloud servers. Finally, data may be presented to the data subject, or are forwarded to third-parties for further analysis at the data disclosure phase. Each of these phases results in privacy threats to data subjects [7].

In this paper, we focus on the data collection phase. This is the first phase when data from a data subject are received by a smart connected home system. Moreover, this is considered the entry point when the resident's privacy can get compromised [28]. Privacy is commonly described as an inherent human right that enables individuals to have "personal autonomy, freedom of association, moments of reserve, solitude, intimacy, and independence" [8].

Data collection is affected by two main activities that could compromise the privacy of data subjects [32]: *surveillance* and *interrogation*. Surveillance consists of methods of watching, listening, and recording of a subject's activities. Interrogation describes methods that ask or elicit information from a subject. While interrogation occurs with the awareness of the subject, surveillance can be covert [32].

Here, we focus on the surveillance class. This is especially as IoT systems have been reported to monitor the behavior and activities of data subjects, their environment, including children and guests, in that environment [34]. Moreover, the aggregated data from various IoT systems can lead to total surveillance of users [38].

3 RELATED WORK

Several efforts have been made to develop a classification for IoT devices and smart connected home systems.

The International Telecommunication Union (ITU) [31] classifies devices into four categories according to type and functionality: data-carrying device, data-capturing device, sensing and actuating device, and general device. The ITU classification is generic and can be applied for classifying different IoT devices. Nonetheless, it fails to represent certain home devices, e.g., smart speakers, that embed sensors and can act as gateway devices.

Moawad et al. [4] classify IoT objects into four levels (Level 0 - Level 3) depending on the device's interaction abilities. Focusing instead on the device's features, Hoang [16] classify smart home appliances into eight different categories: content retrieval, content storage/usage, communication/messaging, remote surveillance, remote control, remote maintenance, instrument linkage, and networked game. Both [4] [16] are useful for conducting a comprehensive analysis of the home. Nonetheless, they are theoretical models and do not take into account the nature of commercial smart connected home systems.

Bormann et al. [5] in Request For Comments (RFC)-7228 classify devices into three distinct classes (C0, C1, C2) to differentiate between IoT devices based on their degree of resource constraints, notably RAM and ROM requirements. The European Union Agency for Network and Information Security [9] extended [5] adding high-capacity devices – devices, such as gateways, that are typically powered by the mains supply, and can implement strong security features. While [5] and [9] have precise specifications for differentiating between devices, they are mostly intended for capturing the hardware specifications and support for security enhancing mechanisms.

Alam et al. [29] offer a taxonomy of smart home monitoring devices, broadly classifying each as either a sensor, a physiological device, or a multimedia device. Bugeja et al. [6] presented a classification of smart connected home devices categorizing them into eight clusters depending on their functionality: energy and resource management, entertainment systems, human-machine

[1] https://foundation.mozilla.org/en/privacynotincluded [accessed Aug 14, 2020].

interface, health and wellness, household appliances and kitchen aids, networking and utilities, security and safety, and sensors. Recently, Kumar et al. [12] categorized IoT devices into eleven classes depending on their type: wearable, game console, home automation, storage, surveillance, work appliance, home voice assistant, vehicle, media/TV, home appliance, and generic IoT. The work of Kumar et al. [12] has been empirically derived and evaluated. Nonetheless, [12] [6] do not focus on the surveillance aspects of Internet-connected devices.

Focusing on privacy aspects is Hong's [17] classification of pervasive devices. This clusters devices into a three-tier pyramid in which each tier poses different privacy challenges based on the capabilities of the devices in that tier. Tier 1 represents highly personal devices such as tablets. Tier 2 represents devices that support basic interactivity such as TVs. Tier 3 represents constrained devices such as RFID-enabled ID cards. While [17] is useful for discoursing about privacy threats it does not consider the different data collection capabilities of devices to form the tiers.

Reviewing the existing work, we observe that classification models tend to be derived manually by analyzing the existing scholarly works (e.g., [29]), insightful observations (e.g., [17]), and pragmatic experience of the IoT domain (e.g., [31]). This makes the discussed classifications more dependent on subjectivity than when the clusters are formed through machine learning methods. For this reason, we build our classification automatically from the actual specifications of commercial smart connected home systems by leveraging unsupervised machine learning. Moreover, we build our classification to use it as a foundation for conducting privacy risk analysis.

4 DEVELOPING THE CLASSIFICATION

4.1 Data collection

There are different IoT product collections platforms. Two of these are smarthomedb.com (SmartHomeDB) and Mozilla's "*privacy not included" (PrivacyNotIncluded). SmartHomeDB is an extensive online platform describing the technical capabilities of different smart connected home systems. PrivacyNotIncluded is an online database focusing on consumer IoT systems. In our case, although it features less systems, we select PrivacyNotIncluded for devising our categorization for three main reasons. First, it includes apps, and these apps are fundamental for managing the home devices but can also constitute privacy risks. Second, it includes more recent devices than SmartHomeDB. Third, the specification of systems is based and derived primarily from privacy policies. Privacy policies are legal documents acting as a contract between a manufacturer or service provider and customers [1].

At the time of our study (as of January 2020) PrivacyNotIncluded consisted of commercial products selected from top sellers on Amazon Prime Day, including those featured in the Target Open House, and systems that were highly rated across a variety of consumer product websites such as Wirecutter, The Toy Insider, PC Magazine, Tech Radar, and Gear Brain [36].

In the dataset, systems offering different categories of functionality: energy and resource management (e.g., smart thermostats), entertainment systems (gaming consoles), health and wellness (fitness trackers), human-machine interface (digital personal assistants), household appliances and kitchen aids (robotic vacuum cleaners), and security and safety (video doorbells), are included.

All the items found in the dataset were used as input to the data processing stage.

4.2 Data processing

In the data processing stage, we rely on web mining to extract the IoT device capabilities. The extraction was done using Cascading Style Sheet (CSS) selectors. In the actual implementation, *scrapy*[2] – an application framework for crawling web sites and extracting structured data – was used as a library in the Python programming language.

Each system was consequently represented by a vector of size n, with n being a constant equal to the total number of surveyed capabilities ($n = 6$). Each of the capabilities; each of them being a nominal variable; was coded as a binary vector, with 1 indicating that the capability is supported, and 0 otherwise[3]. These represent the recording capabilities of the systems, represented as attributes, as shown in Table 1.

After the data extraction was completed, a total of 103 instances was collected. From this, 16 duplicate entries and 6 instances with unspecified attribute values were removed. This left a total of 81 unique instances. Moreover, additional features, e.g., encryption support, while they were collected, they were excluded for clustering purposes. This is as the classification is focused on the sensing data the systems collect.

The actual output was saved into a CSV file to facilitate the data analysis. Finally, a text column, category, was added to the output file. This column represented the functional category the corresponding smart connected home system belongs to.

4.3 Data analysis

The CSV file was loaded into RStudio[4]. Using RStudio we performed cluster analysis using k-means clustering. The k-means clustering algorithm is the most commonly used data mining technique to partition observations into k clusters [25]. Given that the data was binary encoded, Manhattan distance [11] was used as a distance measure.

To determine the optimal number of clusters we follow a two-step approach. First, we applied the gap statistic method [35] – a statistical testing method that compares evidence against the null hypothesis to determine the optimal number of clusters. For the implementation, we used the *amap* and *factoextra* libraries in RStudio. Next, we manually assess the clusters for their meaning according to the cluster center values for each data collection capability.

All the six features (capabilities) identified in Table 1 were used as input for clustering the systems. After the clustering was performed, we used our previous work [6] on smart connected home functional categories as a reference for populating the category column. The selection of the category name was done by searching the database used in [6] for a possible previously assigned category

[2] https://docs.scrapy.org/en/latest/intro/overview.html [accessed Aug 14, 2020].
[3] We mark a capability that is not defined or indicated as N/A in the dataset as 0. However, for these records we separately inspect their corresponding device specification and privacy policy to ensure that we have a valid value for the attribute.
[4] https://rstudio.com [accessed Aug 14, 2020].

Table 1: Characteristics of the data collection capabilities used as features for developing the classification of systems.

Type	Description	Product Examples
Camera	Video and/or audio through sensors embedded in the connected systems	WyzeCam [24]
(App) Camera Access	Video and/or audio through a mobile app	Ring Video Doorbell [30]
Microphone	Audio through sensors embedded in the connected systems	Apple Homepod [19]
(App) Microphone Access	Audio through a mobile app	Google Home [13]
Location	Geophysical location through sensors embedded in the connected systems	Facebook Portal [21]
(App) Location Access	Geophysical location through a mobile app	Amazon Smart Plug [3]

for the system, or otherwise done manually by comparing a system's functionality with systems sharing similar characteristics in [6].

5 EXPERIMENTAL RESULTS

5.1 Classification of smart connected home systems

The gap statistic indicates that four clusters, k=4, is the optimal amount for clustering our dataset via k-means. This value, which is indicated by $fviz\_gap\_stat()$ function available in $factoextra$ library, as the first local maximum, was compared with other clusters values but it was found to be the most fitting and informative choice for our dataset. Large values of k (e.g., k=8) are overfit for our dataset, and the stats for smaller k (i.e., $k \leq 3$) are significantly lower. The scree plot for the gap statistics is shown in Figure 1. Since k-means clustering starts with k randomly selected centroids, we first set a seed for R's random number generator via $set.seed(123)$. We then performed the actual k-means clustering with k=4.

The categorization obtained consists of 4 classes (k=4): Cluster 1 with 32 systems (39.5%), Cluster 2 with 15 systems (18.5%), Cluster 3 with 26 systems (32%), and Cluster 4 with 8 systems (10%). The cluster features, alongside the number of systems implementing a given capability, are summarized in Table 2 and their labels; identified through introspection; are described hereunder:

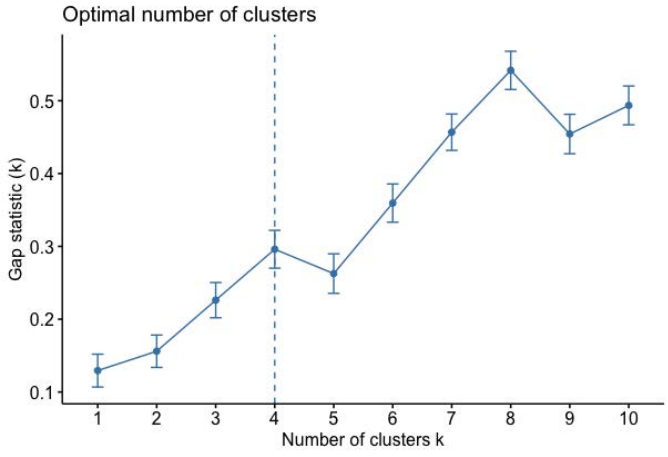

Figure 1: Scree plot for the optimal number of clusters according to gap statistics. The optimal number of clusters by this metric is four.

- **Cluster 1 – App-based accessors:** Systems that predominately utilize mobile apps to collect data from users. All the systems in this category collect audio and location data through mobile apps. Additionally, they may access video data from apps. This category includes mostly entertainment systems. An example is Amazon Fire TV Stick [18]. This is a media player that streams video, music, and games to the TV.
- **Cluster 2 – Watchers:** Systems that collect the behavior of users typically by using sensors embedded inside the connected device. All the systems in this category extract both video and audio data directly from the device's hardware. Some systems may also track the location of users and/or the device. Common systems in this cluster belong to the security and safety domain. An example is Nest Cam Indoor Security Camera [14] which supports 24/7 video recording, motion detection alerts, and noise detection via multiple microphones.
- **Cluster 3 – Location harvesters:** Systems that are focused on capturing precisely the location of users. All the systems in this category collect location data from the mobile app, and may also utilize built-in sensors on the device. Common systems in this class tend to be related to the health and wellness, household appliances and kitchen aids, and energy and resource management category. An example is Fitbit Ionic Watch [20]. This is a GPS smartwatch that is used for 24/7 heart-rate monitoring, health and wellness insights, payments, and more.
- **Cluster 4 – Listeners:** Systems that are primarily designed to capture audio data from users. Common systems in this class are related to the human-machine interface category. An example is Google Home [13], a voice assistant and smart speaker that can be used to control the home, play music, call friends, etc.

5.2 An analysis of smart connected home systems

The results of the quantitative survey, grouped according to where the sensors are installed, are summarized in Figure 2. Hybrid indicates systems that can read an attribute from either the device or mobile app, including both concurrently.

Sensors on-app is correlated to the app-based accessors and location harvesters. This is as 100% of the devices in these clusters

Table 2: Distribution of data collection capabilities across the different cluster means.

Capability	App-based accessors	Watchers	Location harvesters	Listeners
Camera	0	15	0	0
Microphone	14	15	5	5
Location	14	9	11	1
(App) Camera Access	25	11	12	2
(App) Microphone Access	32	12	0	3
(App) Location Access	32	9	26	0
Camera Device & App Access	0	11	0	0
Microphone Device & App Access	14	12	0	2
Location Device & App Access	14	5	11	0

utilize apps for reading the audio and/or location. Sensors on-device is highly correlated to the watchers cluster as 100% of the devices in this cluster use the actual hardware for recording audio/video data, and at 63% to the listeners cluster. Hybrid is mostly correlated to the watchers category as apps are used on average by 77% of the devices in that cluster.

Overall, location is the most collected attribute. This is collected by 89% of the surveyed systems. It is read at 93% from the mobile app. On the contrary, camera is the least popular sensor, being used by 67% of the surveyed systems.

The majority of connected systems rely on apps for collecting data. In the case of audio data, the differences between apps and devices is not as high. Specifically, it is collected by 14% more apps than devices.

11% percent of the systems do not access the location; 28% do not access the microphone; and 33% do not access the camera.

Using our previous work [6] on smart connected home functional categories as a reference taxonomy of devices, we observe that the analyzed 81 systems belong to 6 categories as follows: energy and entertainment (3), entertainment systems (36), health and wellness (20), household appliances and kitchen aids (2), human-machine interface (4), and security and safety (16).

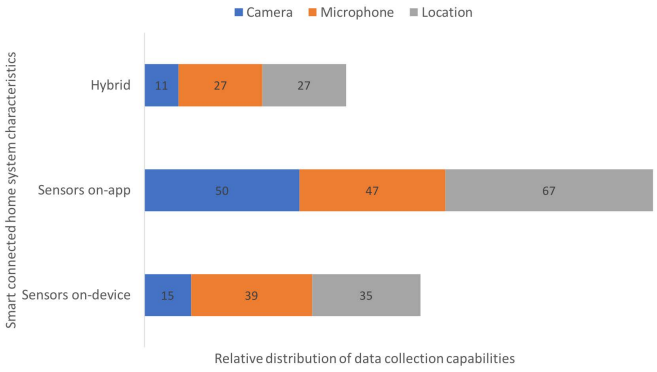

Figure 2: The availability of different sensor types across the entire distribution of surveyed systems (n=81). Values inside cells represent the number of systems supporting a particular data collecting capability. Hybrid indicates systems that can simultaneously read a data collecting capability from the device and mobile app. The majority of smart connected home systems rely on apps for collecting data, with location being the most collected attribute.

For hybrid systems; this is the riskier of the three categories depicted in Figure 2 because of the wider attack surface; we observe that within those, systems that access: cameras are at 91% in the security and safety domain; microphones are at 48% in entertainment systems followed by 37% in the security and safety domain; and location are at 32% in entertainment systems followed by an equal amount of 29% in both health and wellness, and security and safety systems. A total of 6% of the systems have access to all the data collecting capabilities, with 80% of these belonging to the security and safety domain.

By leveraging the computed classification of smart connected home systems in Section 5.1 we can gain additional insights on the investigated systems as displayed in Figure 3. In Figure 3, we display the ratio of device clusters present in each smart connected home functional category. As an example, we observe that the characteristics of listeners are dominant in the entertainment systems and human-machine interface category.

6 DISCUSSION

In this section, we compare our proposed classification of smart connected home systems with the related work showcasing the importance of the developed classification and analysis work. Then, we suggest a sensitivity metric as a potential indicator of data exposure of smart connected home systems. Finally, we identify limitations and potential extensions of our study.

Usefulness of the developed classification. To understand the relevance and usefulness of our proposed classification in Table 3 we compare it to the existing work introduced in Section 3. While the number of classes (n=4) attained is similar to the majority of the related work, our classification is specifically designed for privacy risk analysis and its main advantage is that it is automatically constructed. Having the classification developed in an automated way is better than manual classification construction; e.g., avoids time-consuming manual processing later and can be rebuilt at any point with updated data; for example as the database of devices grows or if new attributes of devices are encoded. Moreover, it reduces possible biases and subjectivity introduced when selecting and labeling data. The human expert can still be part of the classification construction by providing interpretation, evaluate

Table 3: Contribution of our classification of smart connected home systems in relation to the existing research work.

Research work	Classification criteria	In-scope systems	Focus	Development	No. of classes
ITU [31]	Device type and functionality	IoT devices	Generic	Manual	4
Moawad et al. [4]	Interaction capabilities	IoT devices	Generic	Manual	4
Hoang et al. [16]	Device capabilities	Smart home devices	Security	Manual	8
Bormann et al. [5]	Processing capabilities	Constrained IoT devices	Security	Manual	3
ENISA [9]	Hardware specifications	Smart home devices	Security	Manual	4
Alam et al. [29]	Data collection capabilities	Smart home monitoring devices	Generic	Manual	3
Bugeja et al. [6]	Functional categories	Smart connected home devices	Generic	Manual	8
Kumar et al. [12]	Functional categories	Smart home devices	Security	Manual	11
Hong [17]	Device capabilities and relationship to the device	IoT devices	Privacy	Manual	3
This paper	Data collection capabilities	Smart connected home systems	Privacy	Automatic	4

the system, or, in future work, interactive feedback on groupings. Finally, since our classification criteria are the data collection capabilities, this can account for general-purpose and multi-functional devices. Accounting for the nature of these devices is an aspect that is often missed by researchers. Overall, the classification with the accompanying analysis can be used to pinpoint trends in the development of commercial smart homes but as well serve as a foundation for identifying opportunities of generalizations and common solutions for the smart connected home when it comes to privacy risk analysis and risk mitigation.

Measuring personal data sensitivity. Smart connected home systems make the residents prone to new privacy threats. Nonetheless, it is challenging to quantitatively measure the potential data exposure of a smart connected home system. Except for a few studies, e.g., Liu et al. [26] who calculate the privacy score of information shared on social networks, there are no standard metrics that evaluate the data sensitivity of a smart connected home system. Even though it is not the scope of this paper, we argue for the usefulness of a sensitivity score [or index or metric] that quantifies how sensitive a smart connected home system is with respect to the data it collects, processes, and distributes. Such a score could be defined in several ways and to illustrate this idea we present one potential candidate.

Let S denote the smart connected home system. Let $D = \{d_1, ..., d_n\}$ denote the set of data types that are collected by sensors used in S, for example, $\{image, audio, position\}$. The sensitivity of S is defined by the equation:

$$sensitivity(S) = \sum_{i=1}^{n} \sum_{j=1}^{k} weight(d_i, p_j) \times context(p_j)$$

where $weight(d_i, p_j)$ is a value assigned by a data subject indicating the relative importance of data type ($d_i \in D$), with respect to each privacy parameter p_j ($1 \leq k$), and $context(p_j) \in \{0, 1\}$ depending on whether p_j is relevant in the current context or not. Some examples of p_j are: data type sensitivity, location sensitivity, and data accessibility. In our case, for demonstration purposes and basing it on our classification, we adopt possible weights for the parameters as follows: data type sensitivity (4=image; 3=audio; 2=position), location sensitivity (4=living room; 2=kitchen; 1=basement), and data accessibility (3=device and app; 2=app; 1=device).

For illustrating the usage of the suggested sensitivity equation, we utilize a random stratified sampling strategy based on the category field. This is done to pick an unbiased sample with all the different categories as represented in Figure 3. Moreover, a stratified sampling is preferred to simple random sampling as the used data size has a small sample size. Stratified sampling was conducted using the *rsample* library in RStudio. We keep the sample size equal to the number of available categories in the database (i.e., n=6); as

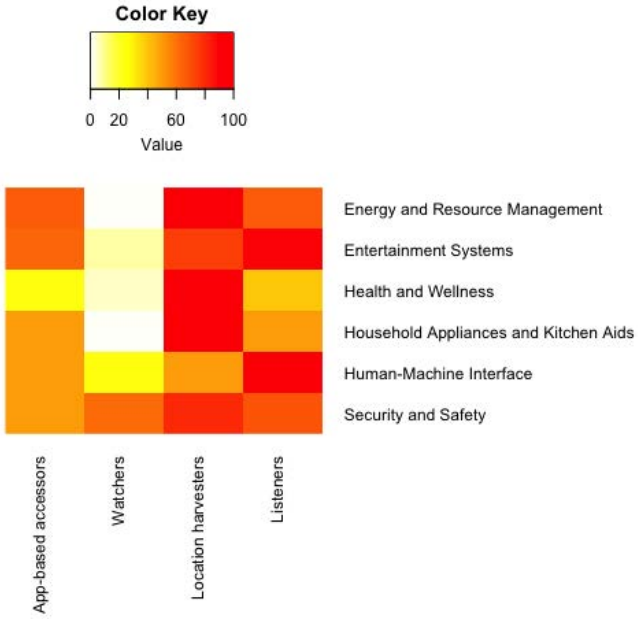

Figure 3: Heatmap representing the ratio of system clusters and smart connected home functional categories. Darker colors indicate a greater presence of a system cluster, and lighter colors otherwise. As an example, there are more watchers in the security and safety category than the rest of the system clusters.

Table 4: Random stratified sample of smart connected home systems including their data collection characteristics and sensitivity value.

System	Category	Cluster	System location	Data type	Data accessibility	Sensitivity
SmartThings Outlet	Energy and resource Management	App-based accessors	Kitchen	image, audio, position	app	21
Apple TV	Entertainment systems	Listeners	Living room	audio, position	device	15
Petnet SmartFeeder	Health and wellness	Location harvesters	Kitchen	image, audio, position	app	21
Oittm Smart Aroma Diffuser	Household appliances and kitchen aids	App-based accessors	Kitchen	image, audio, position	app	21
Apple Homepod	Human-machine interface	Location harvesters	Living room	audio	device	8
Nest Cam Indoor Security Camera	Security and safety	Watchers	Basement	image, audio, position	device and app	21

of today, this value also coincides with the typical amount of smart connected home systems available in a smart home. In Table 4, we indicate the chosen systems alongside their corresponding cluster, location where the system is installed, collected data types, channels over which the data types can be accessed, and the calculated sensitivity score. For simplicity purposes, we assume that all the features are relevant in the current context.

By computing the sensitivity values as indicated in Table 4, we observe 4 systems – SmartThings Outlet, Petnet SmartFeeder, Oittm Smart Aroma Diffuser, and Nest Cam Indoor Security Camera, that have the highest sensitivity score (21). Each of these collect all the sensitive data types. However, if it was the case that Nest Cam Indoor Security Camera was installed at a different location than the basement, then this device would have had the highest sensitivity score. This is as it would have had continuous access to data being generated from a more private zone. On the contrary, the device with the least sensitivity score (8) is Apple Homepod. This is because this device has access only to audio data, and access is limited to a dedicated device. Therefore, the potential data exposure of Apple Homepod is the lowest of the six analyzed devices.

Limitations and extensions. Our research does not come without limitations. First, is the number of systems analyzed. While this may limit the generalizations and conclusions that can be reached, given its possibly non-representative nature, the proposed classification is a natural fit for the data analyzed also from an expert perspective, being interpretable. Nonetheless, we cannot guarantee for objective validity of the clusters, and thus although promising we encourage future research to further investigate and enhance the attained classification. Second, the number of sensor data collection methods was limited due to the database used. Should other data collection methods be used, and as the data grow in dimensions and size, other alternative partitioning methods may have to be considered. For instance, *Partitioning Around Medoids* and *Clustering LARge Applications* as alternatives to k-means [15]. Here, it would be also beneficial comparing the accuracy of clustered data with k-means with that of other clustering algorithms, and potentially with other datasets. This will aid in assessing the solution's stability, reliability, and validity. Finally, we suggested a metric for grading the data sensitivity of a smart connected home. Nonetheless, this metric only yields an approximate indication of personal data exposure and it needs further analysis and validation. In practice, other parameters, e.g., data accuracy, retention time, and trust in a manufacturer, could be included as privacy parameters for measuring privacy risk.

7 RESEARCH DIRECTIONS AND BEST PRACTICES

This section discusses some areas where further investigation is required and identifies some best practices that can be adopted by smart connected home system manufacturers and service providers.

Controlling data collection. Most of the analyzed systems, especially in the app-based accessors class, make it rather easy for data subjects to control their personal data exposure through mobile apps. However, such control might not be available for systems in particular in the watchers category. This is as these systems may require the user to physically disable a component, e.g., the microphone to curtail tracking. As a best practice, data subjects should be provided the possibility to opt-in for data collection, and thus data will not be collected unless the user decides so, or otherwise have the option to opt-out of such data collection. The facility to opt-in/opt-out should be available also for any identifiable data to be stored.

Transparency about data collection. Given the potential risks especially to surveillance introduced when adopting smart connected home systems, data subjects should be aware of data practices adopted by manufacturers or service providers. Awareness allows data subjects to reconsider what they are comfortable about sharing, and take action if desired. As a best practice, details of data that are being collected about them, the purpose or benefit gained for such collection, the location of data collection, and whether that data are retained or shared with other parties, should be known to data subjects. This can be achieved potentially through the use of software notifications, having details listed on a product website, or printed directly on product packaging.

Ethical uses of collected data. Additional information can possibly be inferred from the data collected by smart connected home systems. For instance, machine learning algorithms may automatically infer a user's sleep patterns and home occupancy, based on a user's location or movements within the home. This is particularly

prevalent with systems in the watchers class. Moreover, data curators may collect and assimilate different data types from users to create profiles that contain information about user's preferences, aptitudes, behaviours, and inferences about more sensitive attributes, such as race, health conditions, and financial status. This can lead to targeting and dissimilar treatment on the basis of these traits. As a best practice, a manufacturer should indicate what inferences can be made with the data and likewise the sensors that are used for making those inferences. Moreover, regulations should be in place to overall ensure the fair and ethical uses of collected data.

8 CONCLUSIONS

The growth and heterogeneity of smart connected home systems raise the importance of a classification that groups systems into categories indicative of their data exposure. This is an important precursor for the assessment of privacy risks targeting smart home residents.

Focusing on data exposure, we developed a novel four-tiered classification of smart connected home systems clustering systems into: app-based accessors, watchers, location harvesters, and listeners. We did this by clustering the data collection capabilities of 81 smart connected home systems. Also, we analyzed their distribution of data collection capabilities. Here, we observed that the vast majority of sensory data are collected through mobile apps. Overall, the presented classification and analysis can help us better understand the privacy implications and what is at stake with different smart connected home systems.

For future work, it would be useful to analyze a broader range of systems as a means to validate the presented classification. Another avenue is to explore other data types that the device sensors can collect to render a more comprehensive analysis of privacy risk. Finally, it would be beneficial to use the classification together with a set of quantitative privacy metrics as a basis for conducting a formal privacy risk analysis of the home.

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] Abhilasha Ravichander *et al.* 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. *arXiv preprint arXiv:1911.00841* (2019).
[2] Bako Ali and Ali Ismail Awad. 2018. Cyber and physical security vulnerability assessment for IoT-based smart homes. *Sensors* 18, 3 (2018), 817.
[3] Amazon. 2020. Amazon.com: Amazon Smart Plug, works with Alexa – A Certified for Humans Device: Amazon Devices. https://is.gd/ESh3fS
[4] Assaad Moawad *et al.* 2012. Introducing conviviality as a new paradigm for interactions among IT objects. In *Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments*, Vol. 907. CEUR-WS. org, 3–8.
[5] Carsten Bormann, Mehmet Ersue, and Ari Keranen. 2014. RFC 7228: Terminology for Constrained-Node Networks. *IETF Request For Comments* (2014).
[6] Joseph Bugeja, Paul Davidsson, and Andreas Jacobsson. 2018. Functional Classification and Quantitative Analysis of Smart Connected Home Devices. In *2018 Global Internet of Things Summit (GIoTS)*. IEEE, 1–6.
[7] Joseph Bugeja and Andreas Jacobsson. 2020. On the Design of a Privacy-Centered Data Lifecycle for Smart Living Spaces. In *Privacy and Identity Management. Data for Better Living: AI and Privacy. Privacy and Identity 2019. IFIP Advances in Information and Communication Technology, vol 576*, Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker (Eds.). Springer.
[8] Ann Cavoukian. 2011. Privacy by design in law, policy and practice. *A white paper for regulators, decision-makers and policy-makers* (2011).
[9] Cédric Lévy-Bencheton *et al.* 2015. Security and resilience of smart home environments. *European union agency for network and information security* (2015).
[10] Christian Debes *et al.* 2016. Monitoring activities of daily living in smart homes: Understanding human behavior. *IEEE Signal Processing Magazine* 33, 2 (2016), 81–94.
[11] Susan Craw. 2010. *Manhattan Distance*. Springer US, Boston, MA, 639–639. https://doi.org/10.1007/978-0-387-30164-8_506
[12] Deepak Kumar *et al.* 2019. All things considered: an analysis of IoT devices on home networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1169–1185.
[13] Google. 2020. Google Home – Smart högtalare och assistent för hemmet – Google Store. https://store.google.com/product/google_home
[14] Google. 2020. Nest Cam Indoor. https://store.google.com/product/nest_cam
[15] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.
[16] Nguyen Phong Hoang and Davar Pishva. 2015. A TOR-based anonymous communication approach to secure smart home appliances. In *2015 17th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 517–525.
[17] Jason Hong. 2017. The privacy landscape of pervasive computing. *IEEE Pervasive Computing* 16, 3 (2017), 40–48.
[18] Amazon Inc. 2020. Fire TV Family - Amazon Devices. https://is.gd/gx6pCl
[19] Apple Inc. 2020. HomePod - Apple. https://www.apple.com/homepod
[20] FitBit Inc. 2020. FitBit Ionic Watch. https://www.fitbit.com/se/ionic
[21] Facebook Inc. 2020. Smart Video Calling with Alexa Built-in | Portal from Facebook. https://portal.facebook.com
[22] Jingjing Ren *et al.* 2019. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference*. 267–279.
[23] Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.
[24] Wyze Labs. 2020. Wyze | Making Great Technology Accessible | Smart Home Devices. https://wyze.com
[25] Hosub Lee and Alfred Kobsa. 2016. Understanding user privacy in Internet of Things environments. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*. IEEE, 407–412.
[26] Kun Liu and Evimaria Terzi. 2010. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 1 (2010), 1–30.
[27] Paul Benjamin Lowry, Tamara Dinev, and Robert Willison. 2017. Why security and privacy research lies at the centre of the information systems (IS) artefact: Proposing a bold research agenda. *European Journal of Information Systems* 26, 6 (2017), 546–563.
[28] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.
[29] Muhammad Raisul Alam *et al.* 2012. A review of smart homes—Past, present, and future. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)* 42, 6 (2012), 1190–1203.
[30] Ring. 2020. Video Doorbell | Smart Doorbell Camera Wireless or Wired | Ring. https://shop.ring.com/products/video-doorbell
[31] ITU Telecommunication Standardization Sector. 2012. Recommendation ITU-T Y. 2060: Overview of the Internet of things. *Series Y: Global information infrastructure, internet protocol aspects and next-generation networks-Frameworks and functional architecture models*. (2012), 2060–201206.
[32] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. L. Rev.* 154 (2005), 477.
[33] J. Sturgess, J. R. C. Nurse, and J. Zhao. 2018. A capability-oriented approach to assessing privacy risk in smart home ecosystems. In *Living in the Internet of Things: Cybersecurity of the IoT - 2018*. 1–8.
[34] R. Thorburn, A. Margheri, and F. Paci. 2019. Towards an integrated privacy protection framework for IoT: Contextualising regulatory requirements with industry best practices. In *Living in the Internet of Things (IoT 2019)*. 1–6.
[35] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423.
[36] Janice Tsai. 2019. How We Evaluated the Products in Mozilla's *privacy not included Buyer's Guide (2018). https://is.gd/5K93GE
[37] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).
[38] Rolf H Weber. 2015. Internet of things: Privacy issues revisited. *Computer Law & Security Review* 31, 5 (2015), 618–627.
[39] Heetae Yang, Hwansoo Lee, and Hangjung Zo. 2017. User acceptance of smart home services: an extension of the theory of planned behavior. *Industrial Management & Data Systems* (2017).
[40] Jan Henrik *et al.* Ziegeldorf. 2014. Privacy in the Internet of Things: threats and challenges. *Security and Communication Networks* 7, 12 (2014), 2728–2742.