A Privacy-Centered System Model for Smart Connected Homes

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@mau.se, †andreas.jacobsson@mau.se, ‡paul.davidsson@mau.se

***Abstract*—Smart connected homes are integrated with heterogeneous Internet-connected devices interacting with the physical environment and human users. While they have become an established research area, there is no common understanding of what composes such a pervasive environment making it challenging to perform a scientific analysis of the domain. This is especially evident when it comes to discourse about privacy threats. Recognizing this, we aim to describe a generic smart connected home, including the data it deals with in a novel privacy-centered system model. Such is done using concepts borrowed from the theory of Contextual Integrity. Furthermore, we represent privacy threats formally using the proposed model. To illustrate the usage of the model, we apply it to the design of an ambient-assisted living use-case and demonstrate how it can be used for identifying and analyzing the privacy threats directed to smart connected homes.**

***Index Terms*—Internet of Things, system model, privacy, privacy threats, home data, smart home, smart living.**

I. Introduction

The smart connected home is a System of Systems (SoS) involving heterogeneous Internet of Things (IoT) devices encapsulating both information and physical spaces. Along with the conveniences and efficiencies brought by these Internet-connected devices, e.g., by smart thermostats, privacy threats in smart homes are resulting in a potential for abuse, misuse, and appropriation of user data.

Despite considerable theoretical and practical contributions from the scholarly and industry communities a standardized representation that accounts for the smart home entities and its data flows is missing. Such a system model is needed for the systematic identification of privacy threats affecting smart homes and as a precursor for privacy risk analysis [1].

To address this need, we present a privacy-centered system model for smart connected homes. This model captures the dynamics of a smart connected home, and the properties and requirements for modeling privacy. In doing so, we leverage Nissenbaum's theory of Contextual Integrity (CI) [2]. CI is based on the premise that privacy is defined as the appropriateness of information flows. Inappropriate information flows are those that violate context-specific informational norms, a subclass of general norms governing respective social contexts. The model usefulness is illustrated by applying it to an ambient-assisted living use-case, and its effectiveness is demonstrated by identifying and exemplifying potential privacy threats occurring in this system.

The remainder of this paper is organized as follows. In Section II we provide a review of related work. The privacy-centered system model is presented in Section III. Next, we identify how the proposed model can be used for privacy threat analysis in Section IV. Then, in Section V we formally describe an ambient-assisted living setup, including the identification of its privacy threats. Finally, in Section VI, conclusions and possible directions for future work are outlined.

II. Related work

Barth et al. [3] formalize some aspects of CI using Linear Temporal Logic. While this framework allows for the precise specification of privacy laws, it is mostly suited for compliance more than on the identification of privacy threats.

Ni et al. [4] introduce a group of Privacy-Aware Role Base Access Control (P-RBAC) models allowing for expressing and reasoning about privacy policies. While the models proposed are generic, they are concentrated on permission assignments and conflict detection.

Omoronyia et al. [5] presented a privacy framework that supports the selective disclosure of personal information in software applications. While this framework can be applied to smart connected homes it requires the explicit representation of states and transitions making it challenging to model a realistic smart home scenario.

Moshin et al. [6] propose a risk analysis framework to formally analyze risks using probabilistic model checking. While this framework includes a system model and is used for similar purposes to ours, it is focused on security threats.

Although this is only a brief account of the relevant contributions, we observe a shortage of privacy models tuned for smart connected homes. It is also evident that some models, while providing a solid theoretical foundation they are challenging to apply in use-cases involving commercial devices. Part of this problem is the lack of a standard model that describes a smart connected home. This is the research gap that we seek to address in this paper.

III. Privacy-centered system model

Based on our previous work, e.g., [7], about the smart connected home, and using the CI as an overarching framework, an IoT-based smart connected home setup, S, can formally be

described as a tuple (H, N, U, L, D, P) where H: house, N: nodes, U: users, L: links, D: data, and P: policy.

- H is the physical building which the residents inhabit.
- N is a set of physical components of the smart connected home space. Effectively, $N = \{C \cup M \cup B\}$ where C: connected devices, M: mobile devices, and B: backends.
- U is a set of human users interacting with the smart connected home space.
- L is a set of communication channels between the nodes and users.
- D is a set of data being collected and processed by the smart connected home setup.
- P is a set of rules describing how data are transmitted between the different entities.

A. House

House represents the physical building where the residents live and perform their daily activities, e.g., cooking, cleaning, sleeping, etc.

From a privacy modeling perspective, we can define $H = \{z_1, z_2, \dots z_n\}$, where $z_i \in H$. z_i represents a zone, i.e., a specific space (area), e.g., the living room, inside the home.

B. Nodes

- *Connected devices:* Hardware units, e.g., domestic appliances, that can sense, actuate, process data, and communicate. These devices differ by their purpose and capabilities but tend to be static and located in a fixed zone in H. A connected home device can implement several core capabilities $\subseteq \{connectivity, sensing, actuating, interaction, storage, processing, gateway, programming, remote{-}admin, \dots\}$. At a minimum, they incorporate one transducer (sensor or actuator) for interacting directly with the physical world; and involve at least one network interface (e.g., Zigbee) [8]. A set of connected devices is denoted as: $C = \{c_1, c_2, \dots c_n\}$, where $c_i \in C$.
- *Mobile devices:* Portable devices such as smartphones, tablets, and wearables that are often used to configure and manage the smart connected home system. These have similar capabilities as C but are in addition mobile. A set of mobile devices is denoted as: $M = \{m_1, m_2, \dots m_n\}$, where $m_i \in M$.
- *Backends:* Infrastructure, often managed by a third-party vendor, that stores and processes data on behalf of connected devices and users. We distinguish between two main backend types $\subset \{edge, cloud\}$, with edge (e.g., dedicated gateways/hubs) having the storage and processing performed inside H, and cloud outside of H. A set of backends is denoted as: $B = \{b_1, b_2, \dots b_n\}$, where $b_i \in B$.

C. Users

Users are the stakeholders that utilize, enable, or maintain the services provided by nodes. Excluding the data attacker, from a privacy standpoint, there are three role types in an IoT system: data subject, data controller, and data user [9].

Data subjects are the human individuals that can be personally identified through the data in question. Typically, this is the device owner but may also include other family members and guests that interact with the smart connected home [10] through M and sometimes directly via a physical interface offered by C. Data controllers, e.g., a cloud service provider, are responsible for B, and often are involved in collecting and processing data on behalf of data subjects. Data users, are frequently the data subject, but can also include, e.g., employees of the data controller, that access the released data.

D. Links

A smart connected home may use a variety of network protocols ranging from wired/wireless, short-range/long-range, and IP-based/non-IP based. We use the term *link* to represent a communication channel.

Links are associated to nodes and users, via a mapping function, to form a graph with $(N \cup U)$ representing vertices, and L representing edges.

Essentially, this graph indicates the possible paths over which information between subjects, senders, and recipients can be exchanged.

E. Data

Data represent the data attributes that are collected/processed by a smart connected home configuration during system setup and/or operation.

Formally, this can be represented as a set of tuples (d_i, d_s, d_p, d_t) where d_i: data item, d_s: data subject, d_p: processing purpose, and d_t: retention time.

d_i represents specific attributes that the system is collecting/processing. This ranges from specific granular data items such as the data subject's "weight" to more generic data categories such as "usage" information depending on the device type [10].

d_s represents the data subject. This can be a human user in particular to represent cases involving personal data. It can also include "system" representing nodes, such as connected devices, that may generate data automatically.

d_p indicates the purpose for collecting/processing d_i, e.g., "marketing," "user authentication," "billing," etc.

d_t describes the general condition for storing d_i with possible values $\in \{indefinite, purpose, date\}$ with: $indefinite$ indicating there is no time constraint for the deletion of the data; $purpose$ indicating that the data has to be deleted after the completion of the corresponding purpose, i.e., after d_p is attained; and $date$ indicating the actual date/time for the deletion of d_i.

F. Policy

Policy represents the rules (norms) pertaining as to how data are transferred between the different entities.

Formally, this can be represented as a set of tuples (lg_i, dp_i, s, r, c) where lg_i: link group identifier, dp_i: data

permissions, s: sender $\in (N \cup U)$, r: recipient $\in (N \cup U)$, and c: condition for transmission specifying when the data are transferred to the recipient(s). To c we also include null ($\emptyset$) condition signifying that the transmission link is always active.

lg_i represents a link group identifier, $lg_i \in L$. This can refer to a group e.g., $l_{HAN}|\ l_{HAN} = \{doorbell, speaker, lock\}$ or otherwise a single logical channel e.g., $doorbell\_speaker$. This especially allows for scalability of the model.

dp_i is a set of pairs (di, op) specifying authorized operations $\subseteq \{read, write, \ldots\}$ that can be performed on di by r.

For c, we adopt Ni et al. [4] notation where c consists of a conjunction ($\wedge$) of context variables as exemplified below:

- Time, domain=$\{$*9:00–17:00*$\}$; it represents different temporal periods.
- Date, domain=$\{$*20190101–2020101*$\}$; it represents a range of dates.
- Location, domain=$\{$*zone* $\in H$*, city, state/country*$\}$; it represents the place where the sender has to be located.

Other examples could be for instance: TransmissionFrequency, OwnerInform, OwnerConsent, etc. All links have a corresponding rule that is represented in the policy.

IV. Privacy threat identification

For identifying privacy threats we leverage Ziegeldorf et al. [1] enumeration of IoT privacy threats. However, we formalize the threats and contextualize them using the proposed system model introduced in Section III. Also, we present a set of primitives to help express the privacy threats:

- $\Phi^{d_s}_{d_i} \implies$ returns `true` if d_i is an explicit identifier, e.g., name, biometric identifier, or a quasi-identifier, e.g., birth date and gender, that can be used to identify d_s.
- $\theta^{d_s}_{d_i} \implies$ returns `true` if d_i has location information that can be used to identify d_s or his/her household.
- $\gamma^n \implies$ returns `true` if a system identifier, e.g., MAC address, or a quasi-identifier, e.g., power and time, are transmitted by n allowing for its identification.
- $perms^u_{d_i} \implies$ returns the set of authorized operations, op, permitted on d_i for u.
- $read^{l_i}_{d_i} \implies d_i$ is received over l_i. This operation results in the recipient learning about d_i.
- $write^{l_i}_{d_i} \implies d_i$ is sent over l_i. This operation results in storage of d_i.

Identification: This characterizes the threat of linking a persistent identifier, e.g., name, address, and age, with a data subject and thus revealing the identity of the individual.

Formally, $identify \stackrel{\text{def}}{=} \forall p \in P, \exists\ (dp_i <> \emptyset) \wedge (\Phi^{d_s}_{dp_i.d_i} =$ `true`).

Localization and tracking: This allows for the recording of a person's location and thus track movement.

Formally, $tracking \stackrel{\text{def}}{=} \forall p \in P, \exists\ (dp_i <> \emptyset) \wedge (\theta^{d_s}_{dp_i.d_i} =$ `true`).

Profiling: This represents the threat of collecting and correlating information about individual activities to generate new information from the original data.

Formally, $profiling \stackrel{\text{def}}{=} \exists\ (l \in L) \wedge (\Phi^{d_s}_{d_i} =$ `true`$) \wedge write^l_{d_i}$.

Linkage: This consists in linking different separated systems such that the combination of data sources reveals information that the subject did not disclose or intended to.

Formally, $linkage \stackrel{\text{def}}{=} \forall l_i, l_e \in L \wedge (l_i <> l_e) \wedge (read^{l_i}_{d_{i_1}} \wedge read^{l_e}_{d_{i_2}} \implies read^{l_e}_{d_{new}}) \wedge (read \notin perms^u_{d_{new}})$.

Privacy-violating interaction and presentation: This refers to exposing personally identifiable information to individuals who are not supposed to have access to it.

Formally, $interaction \stackrel{\text{def}}{=} \exists\ (l \in L) \wedge (u \in U) \wedge read^l_{d_i} \wedge (read \notin perms^u_{d_i})$.

Inventory attacks: This represents the unauthorized collection of information about the existence and characteristics of personal things.

Formally, $inventory \stackrel{\text{def}}{=} \exists\ (n \in N) \wedge (\gamma^n =$ `true`$)$.

Lifecycle transitions: This refers to when nodes disclose private information during changes of control spheres in their lifecycle.

Formally, $lifecycle \stackrel{\text{def}}{=} \exists\ (l \in L) \wedge (u \in U) \wedge write^l_{d_i} \wedge (write \notin perms^u_{d_i})$.

V. Ambient-assisted living use-case

To illustrate the usage of the proposed model, we apply it to the setup illustrated in Figure 1. Using S=(H, N, U, L, D, P) we describe the smart home setup:

- House, $H = \{house\}$
- Nodes, $N = \{doorbell, lock, speaker, manufacturer, smartphone\}$
  $C(speaker).capabilities = \{gateway, storage, processing, interaction\}$
  $B(manufacturer) = cloud$
- Users, $U = \{owner\}$
- Links, $L = \{doorbell\_speaker, lock\_speaker, speaker\_cloud, cloud\_smartphone, owner\_smartphone, owner\_speaker\}$
- Data, $D = \{(lock_status, system, $ *lock open/close*$, indefinite)$, $(cmd, system,$ *application command*$, purpose)$, $(video, visitor,$ *video of guest(s)*$, purpose)$, $(audio, owner,$ *voice interaction*$, purpose)\}$
- Policy, $P =$
  $\{(doorbell\_speaker, \{(video, \{read\})\}, doorbell, speaker, \emptyset)$,
  $(lock\_speaker, \{(lock\_status, \{read\})\}, lock, speaker, \emptyset)$,
  $(speaker\_cloud, \{(audio, \{read\})\}, speaker, manufacturer, Time = \{8:00-24:00\} \wedge Location = \{house\})$,
  $(cloud\_smartphone, \{(cmd, \{read\})\}, smartphone, manufacturer, \emptyset)$,
  $(owner\_smartphone, \{(cmd, \{read\})\}, owner, smartphone, \emptyset)$,
  $(owner\_speaker, \{(audio, \{read\})\}, owner, speaker, \emptyset)\}$

In Table I, a summary of the identified privacy threats including reasons for their occurrence is provided.

VI. Conclusions

IoT technologies deployed inside the home challenge the long-held notion that the home is a private, protected, and intimate place. To help in the systematic identification and formal modeling of privacy threats we developed a novel privacy-centered system model based on the theory of CI. Overall, this also contributes towards adding more transparency about risks emerging out of other IoT systems.

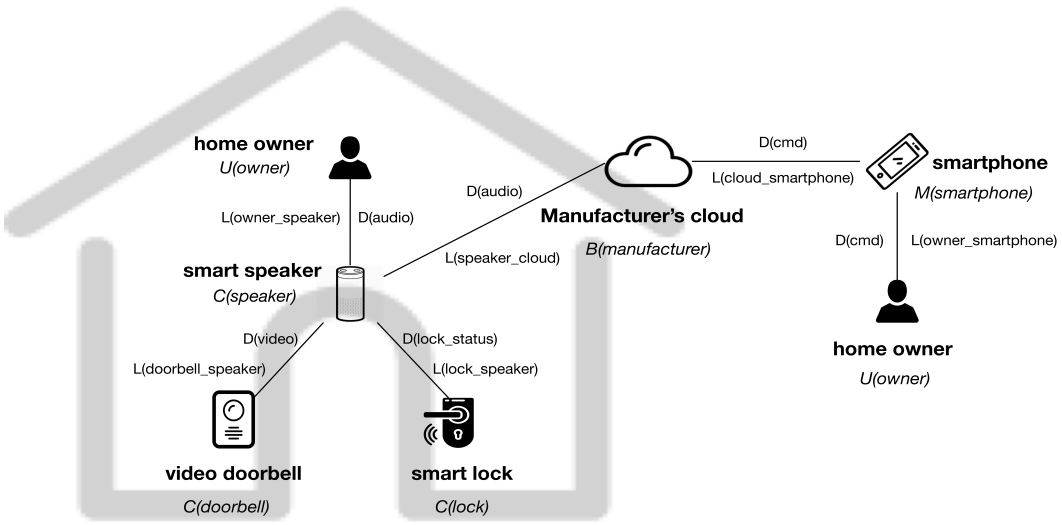

Fig. 1. Smart connected home setup consisting of three connected devices – smart speaker, video doorbell, and smartlock; mobile device – smartphone; and a cloud endpoint. This configuration allows the homeowner the possibility to unlock a door using his or her voice as input and remotely through a smartphone.

For future work, it would be useful to evaluate the completeness of the proposed model through empirical studies, possibly involving a combination of case studies and exploratory descriptive studies. Second, it would be beneficial to express the system model, especially the policy, using a formal specification language such as Z3, S4P, or Promela. This would allow for the verification and analysis of desired functionality. Finally, it would be valuable to develop a threat model with vulnerabilities and attacker capabilities as an extension to the privacy-centered model for risk analysis of the smart connected home.

ACKNOWLEDGMENT

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] J. H. Ziegeldorf, O. G. Morchon, and K. Wehrle, "Privacy in the internet of things: threats and challenges," *Security and Communication Networks*, vol. 7, no. 12, pp. 2728–2742, 2014.

[2] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[3] Barth et al., "Privacy and contextual integrity: Framework and applications," in *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 2006, pp. 184–198.

[4] Q. Ni et al., "Privacy-aware role-based access control," *ACM Transactions on Information and System Security (TISSEC)*, vol. 13, no. 3, p. 24, 2010.

[5] I. Omoronyia et al., "Engineering adaptive privacy: on the role of privacy awareness requirements," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 632–641.

[6] M. Mohsin, M. U. Sardar, O. Hasan, and Z. Anwar, "IoTRiskAnalyzer: a probabilistic model checking based framework for formal risk analytics of the Internet of Things," *IEEE Access*, vol. 5, pp. 5494–5505, 2017.

[7] J. Bugeja, P. Davidsson, and A. Jacobsson, "Functional classification and quantitative analysis of smart connected home devices," in *2018 Global Internet of Things Summit (GIoTS)*. IEEE, 2018, pp. 1–6.

[8] M. Fagan, K. Megas, K. Scarfone, and M. Smith, "Core cybersecurity features baseline for securable iot devices: A starting point for iot device manufacturers," NIST, Tech. Rep., 2019.

[9] C. Li and B. Palanisamy, "Privacy in internet of things: from principles to technologies," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 488–505, 2018.

[10] N. Apthorpe et al., "Discovering smart home internet of things privacy norms using contextual integrity," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, p. 59, 2018.

TABLE I

SUMMARY OF THE IOT PRIVACY THREATS AND THEIR POTENTIAL OCCURRENCE IN THE USE-CASE PRESENTED IN SECTION V. FOR EACH THREAT, A REASON IS PROVIDED JUSTIFYING WHY THE POTENTIAL THREAT MAY EXIST BASED ON THE REFERENCED FORMULA. THE SYMBOL: ● INDICATES THAT THE THREAT IS PRESENT; ◐ INDICATES THAT THE THREAT IS A POTENTIAL FUTURE THREAT; AND ○ INDICATES THAT THE THREAT DOES NOT EXIST.

Privacy threat	**Occurrence**	**Reason**	**Formula**
Identification	●	$p = owner\_speaker$, $dpi.di = audio$, $d_s = owner$	identify
Localization and tracking	○	No location data are collected	tracking
Profiling	●	Audio data are sent over $l = speaker\_cloud$	profiling
Linkage	●	$l_i = doorbell\_speaker$, $l_e = lock\_speaker$, $d_i = video$, $d_e = lock\_status$. Possibly, the speaker may learn the time a user is at home	linkage
Privacy-violating interaction and presentation	○	Only the owner has access to the smart home system	interaction
Inventory attacks	○	No device fingerprint data are revealed	inventory
Lifecycle transitions	◐	No write operations are in P, but speaker has storage capabilities	lifecycle