A Privacy-Orientated Distributed Data Storage Model for Smart Homes

Khutso Lebea and Wai Sze Leung(✉)

University of Johannesburg, Johannesburg, South Africa
{klebea,wsleung}@uj.ac.za

Abstract. With the wide adoption of smart home devices, users are concerned with what sensitive data these devices may be collecting and what that data may be used for. This paper proposes a way to reduce the level of dis-trust between the end-users and the companies that offer smart home hard-ware and/or personalised software services by allowing end-users to retain an optimal degree of control over their personal data (such as voice and video recordings) which is typically collected by the service provider and stored on the service provider's cloud platform.

Keywords: Privacy · Smart Devices · User Data Collection · IoT

1 Introduction

The Internet of Things (IoT) refers to physical devices that are connected to the Internet [1]. According to statista.com, the number of devices connected to the Internet in a typical United States home stands at an average of 10.37 devices in the year 2020. The list of devices includes but is not limited to mobile phones, computers, tablets, televisions and television smart boxes, video game consoles, smart speakers, smartwatches, and virtual reality devices [2]. These smart devices create, analyze, and store an abundance of user data in order to provide what the service providers deem the best possible end-user experience [3].

The creation of smart home devices that are capable of learning and processing information as close to human capacity is one of the core ambitions of Artificial Intelligence (AI) and Machine Learning (ML). To achieve this aspiration, AI and ML have had to make progress in numerous domains, with the aforementioned connected devices leveraging advancements such as object recognition, image processing, speech recognition, robotics, and natural language processing to learn and understand the environments they are designed to function in [4].

To train and perfect these AI concepts, one requires a considerable amount of data for training and testing purposes. Smart home industry giants such as Amazon, Google, and Apple all boast devices designed to make the lives of their end-users simpler and more efficient through smart assistance, a feat only achievable when the various AI

The original version of the chapter has been revised. A correction to this chapter can be found at https://doi.org/10.1007/978-3-031-35822-7_45

A. Moallem (Ed.): HCII 2023, LNCS 14045, pp. 184–193, 2023.
https://doi.org/10.1007/978-3-031-35822-7_13

technologies have the means to learn the habits, traits, likes, and dislikes of the user. This effectively equates to the need to collect data about the user for the smart devices to continuously improve in how they do their job [4].

The issue identified with the creation, processing, and collection of user data is mainly in *how* this data is collected, *who* has access to this data, and for what *purpose* is the data collected. For owners of smart home devices such as a Google Home product, an Amazon Alexa-enabled product, or an Apple Home- Kit device, there exists the privacy concern that said devices are listening and constantly creating, processing, and sending data back to the manufacturer for different reasons [5].

This research topic is a by-product of a research Master's focused on Context- Driven Authentication in which users' physical access patterns were extracted and applied in the decision-making powering an alternative to the existing two- factor authentication mechanism in place [6]. Surveying related literature, it is clear that many researchers have realized that the advancements in AI or any smart or learning system depend largely on the availability of a large number of input data, without which, would prove difficult to determine the future or success of AI projects [4, 6].

This paper proposes a multi-tier, privacy-orientated distributed data storage model, that will allow users who do not wish to share data with the service provider to do so for a fee. The approach is multi-tier, meaning users should be allowed to customize their sharing of information with service providers according to their privacy appetite. The multi-tier approach should increase the level of trust between end users and service providers, whilst increasing the adoption of smart home devices.

The rest of this paper is therefore structured as follows: Sect. 2 unpacks the objectives of this paper, followed by a discussion of the necessary background in Sect. 3 in the form of related work and a literature study. Section 4 then details the reasons behind user data collection while Sect. 5 discusses the legal approaches to solving user privacy issues. Finally, the proposed solution is presented in Sect. 6 before concluding the paper in Sect. 7.

2 Objective

The collection of user data by smart home companies is a problem for all users of smart home devices, regardless of the user's knowledge of Information Technology. This is because while data collected by a single device may be inconsequential, the combination of data collected by several devices can potentially expose patterns about the end-user [7].

The objective of the paper is thus to develop a framework that not only strikes an appropriate balance between allowing the end-user to retain control of the data being collected by smart devices, and the ability for the smart device to collect sufficient user data to function optimally; but serves as a viable solution to decreasing the level of distrust between end-users and smart home companies. To achieve this, the following section provides the necessary context by providing the background into the underlying problem.

3 Background

In 1984, the American Association of house builders officially announced the first version of a smart home. The term "Smart Home" is not restricted to the home, and it has a broader meaning that encompasses any technological environment, including smart cities and smart factories [3, 8, 9].

A smart home is defined as a cyber-physical system built on the IoT, computers, and smart appliances, along with human interactions through communication networks and the Internet [8, 9]. These devices typically communicate with each other and the service provider servers over the Internet via Wi-Fi [10].

3.1 Internet of Things

The architecture of IoT is as layered as follows [1, 11]:

1. Physical devices and controllers
2. Connectivity
3. Edge computing
4. Data accumulation
5. Data abstraction
6. Application
7. Collaboration and processes

Developed by the IoT World Forum in October of 2014 [11], the above architectural layers provide a common framework for the deployment of an IoT solution, with data in such a setup typically bidirectional in nature [12]. This research focuses on the 4th layer where data collected by the physical layer is stored, analyzed, and processed, before being made available to the other layers [3].

Currently, there are two ways to store data collected by the physical layer. The data can be stored and processed, either locally, or on the cloud. When it comes to cloud storage, the data can be stored on a public cloud, a private cloud, or using a hybrid approach with cloud storage [5, 13].

Industry-leading organizations typically prefer storing data using the private cloud approach [5] as this grants the organization complete control over the data and it is not publicly accessible. Under such an arrangement, the organization can set up any security management and day-to-day operation internally. Where third parties access the data, this is generally done through a service level agreement (SLA) contract which typically stipulates what rights the third party has to the data.

One of IoT's security and privacy issues is that end-users are not always aware of the data that the physical layer devices collect [14]. Since the data collected is stored on the organization's private cloud, it is often not possible to see what data the organization is collecting, or for how long it will be kept. As a result, the role of user privacy in IoT has remained largely unexplored in IoT, more so in a smart home context [3].

IoT is a relatively new field in the Information technology space. Much re- search has been done on the privacy and security issues that come with IoT on all of the various layers. However, solutions that are concerned with the use and access of data stored in the data accumulation layer are scarce. In the data accumulation layer, research on privacy and security is conducted mainly to ensure that nodes only have access to the data when authenticated and authorized to do so [15]. Mainly the research done looks at security threats such as a denial- of-service attack or a man-in-the-middle attack [11, 12, 16]. These vulnerabilities have more to do with the architecture of the IoT solution, keeping the data secure within the system.

3.2 User Concerns

The collection of personal data combined with the increased number of Internet- connected devices exposes the user to privacy and security risks. Furthermore, the data collected is generally processed and analyzed by the service provider offsite and not locally within the device containing the data. This fear of privacy risk adds to the potential barrier to adopting smart home devices [3].

Due to the increase in devices and sensors collecting data, there is a need to find an increasingly accurate method of measuring information privacy concerns [11, 15]. While smart home devices are meant to make life easier, they do come at a price. Several reports have shown that the convenience of smart home devices comes at the expense of privacy and cyber security [17]. According to an ADT consumer privacy survey, about 93% of consumers with smart devices in their homes are concerned about how the service providers use and share user data. Respondents say that smart home companies need to take measures to protect their personal data [17, 18].

Consumer concerns go beyond the protection of their personal data and what these companies do with that data. Consumers are afraid of unauthorized data collection and the sharing of their data with third parties. These concerns are over and above the threat of hackers gaining access to these data vaults and using the data stored there for nefarious purposes [17].

Hackers aside, smart home devices are capable of collecting so much information and have so many capabilities. It becomes very difficult to know what information is being collected by the service provider and with whom the service provider shares that information. Furthermore, it can be difficult for the user to know when a device is collecting data or when that specific feature is turned off. In essence, the only time anyone can know for certain that the device is not collecting information is when the device is currently not powered on [17, 18].

The recent uptake in smart home devices has ushered in several new service providers and brands that come onto the market at relatively lower prices as compared to the mainstream smart home service providers. These brands typically hit lower price points for their hardware because they prioritize convenience above privacy, exposing the end-user to a heightened degree of risk [18]. A re- cent example of new smart home service providers taking shortcuts to penetrate the market is Wyze, a company that sells inexpensive smart home cameras. It has recently come to light that Wyze knew for several years that hackers could remotely access its camera feeds, but said nothing to customers [19].

Despite such a concern, most service providers do allow their users to determine what data is collected by smart devices. The power to determine what data is collected is typically found in the permission settings of the smart device. With this, a user can deny permissions that they deem to be too intrusive. However, restricting the device's access to certain data may result in the device not working to its full potential. Another concern is that most such permission settings are set to some default that may or may not favour a more data-driven gathering profile. To counter this, the user is then expected to follow additional, sometimes cumbersome steps to deactivate permissions for a more liberal and relaxed data gathering profile [17].

3.3 Terms and Conditions

One of the consumer issues highlighted in the ADT survey is the fact that 40% of the respondents admitted that they do not feel knowledgeable about privacy, and this issue is made worse by the terms and conditions put up by the service providers. These documents, along with the transparency reports, tend to be lengthy, typically written using technical and legal jargon, which a normal person may not necessarily understand [20, 21]. One approach that can be viewed as a step in the right direction of addressing the legal information overload and making it more accessible to the typical user, has been employed by Apple, making use of pop-up prompts that inform the user to opt-in to a service, and explaining what the purpose of that service is and how it works.

The data collected by smart devices, the purpose of the data collection, the duration of the data storage, along with third parties with whom this data will be shared are usually buried in these terms and conditions. There are typically no alternative routes or agreements that can be made with the service provider unless the product is a business or enterprise version [17]. Companies like Apple, for example, have since identified that certain data collecting and sharing permissions should not be set by default. With their iOS 14.5 update, Apple's mobile operating system has introduced a pop-up that asks iOS users to opt-in to the tracking of their activities within each individual app [22].

To understand why smart home companies go to such lengths to collect user data, one should look at the reasons that drive data collection and data sharing among companies.

4 Why is Data Collected?

One of the reasons why smart home companies or anyone interested in building a learning system needs to collect data is simply to make the system better at what it is designed to do, the better the system is, the more potential to make money.

4.1 Making Money

In 2007, it was estimated that the global market for smart home services was about $38.50 billion. This forecast is set to increase to an estimated $125.07 billion by 2023. At this point, the global market penetration will be at an estimated 19.5% [11], with a large number of devices and appliances following a one-time purchase transactional revenue model.

Such a revenue model is however something of a misnomer - rather than only benefit from the proceeds of the sale of the device alone [23], vendors are also potentially capable of tapping into additional revenue models once the user connects their device to the Internet and is presented with a bevy of subscription services. Some vendors may, for example, opt to supplement the one-time purchase revenue with an advertisement-based revenue model [23, 24].

Advertisement-based revenue models typically include sharing user data with third parties. In this way, information that contributes to building up customer profiles of the end-users could be collected, opening the user up to targeted advertisements [24].

4.2 Improving the Service

In general, smart devices that are designed to learn about their end-users tend to improve their functionality the more they are used. It would therefore be beneficial for smart devices that are already deployed in a household to share what it already knows of the inhabitants' habits with newly-introduced smart home devices to provide an overall smart home experience that is consistent and capable of providing smart assistance [25].

Such a notion is not novel - in 2017, iRobot, a smart home company that manufactures smart cleaning robots admitted that they were considering sharing the floor plan data collected by their smart vacuums with companies such as Google and Amazon. The reason behind this, as said by the CEO of iRobot, was for these companies to develop and provide products and services that would be suited to the end-users' home [23].

Such sharing of users' information is a contentious topic - the problem lies in the fact that service providers are generally not very transparent with their customers concerning what they do with the customer's personal information. This problem is particularly problematic given that service providers appear to be operating in a lawless territory, with most cases bringing scandalous and non- ethical behaviors of the service providers to light while the law appears to have no way of dealing with it [20].

5 A Legal Approach

Another concern linked to the opaque nature of how smart devices share personal data is linked to elements of government spying through the use of in-home cameras and smart speakers [26]. Although the more established smart home service providers (such as Apple, Amazon, Facebook, and Google) have indicated that they disclose when and if governments demand customer data in their transparency reports, the process is not always satisfactory. Apple, for ex- ample, claims that because the data they turn over is anonymized, there is no need to disclose or report with whom the data is shared with [26]. The argument is that Since the data is stored on the cloud, law enforcement agencies, and government agencies are lawfully able to request the data from these service providers if they believe this data could assist in a criminal investigation [17].

To address such concerns, several governments have created regulations and policies that are aimed at assisting smart home service providers in reassuring their customers that their data is not at risk. These policies include The EU General Data Protection (GDPR), the California Consumer Privacy Act (CCPA), and in South Africa, the Protection of Personal Information Act (POPIA) [20, 21]. Such initiatives have, however, not necessarily lived up to their expectations as some of the minimal laws contain ambiguous provisions that tend to blur and complicate data protection issues [17].

It has been estimated that the law is typically about five years behind developing technologies [20]. One reason legislation is not the solution to technological problems is that it is difficult and often impossible for technology industry leaders, not to mention the government and lawmakers, to predict new technologies before they emerge. This leaves the law to constantly play a game of catch-up after the fact [20]. In South Africa, POPIA assented in parliament on the 19th of November 2013 but only commenced on the first of July 2020, with the expectation of compliance one-year later [21]. Given the slow turn of the law, it is clear that technological problems cannot be solved by legislation alone. Rather, technological solutions to technological problems should be a more realistic approach [20].

6 A Technical Approach

With smart home devices as established as they currently are, any solution aimed at identifying an optimal middle ground between privacy concerns and functionality should build on the use of existing technology and methods that are already established. One aspect that should be considered is the fact that later iterations of smart home devices are getting smarter, with increased capacity for local storage (local data abstraction according to the IoT architecture) [1]. This allows for service providers such as Amazon to shift the processing of user data to its local environment. In this way, the smart device is potentially capable of processing commands and building up the user's profile without ever having to connect to the Internet [27]. Such an approach will however lead to an increase in manufacturing costs as such offline-capable devices must have the hardware specifications that ensure it is capable of performing the computations independently.

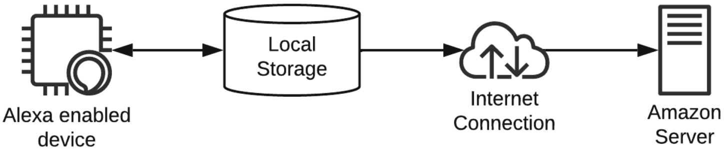

Fig. 1. Architecture of the proposed framework

Figure 1 illustrates the proposed framework. Using Amazon as the example, the framework should ideally work for different smart home service providers and their platforms. Ideally, all devices will have access to a local storage facility, where all the collected data will be stored. Each service provider's devices should be linked to their own service provider shared storage facility. In other words, all devices from a particular service provider will have access to the same storage facility. This decision is based on the fact that not all service providers need the same data for their devices, and in instances where the data may be the same, the processing could be different.

Within each local storage facility, there should be permissions that are in the control of the end-user. In this way, end-users should have the ability to grant access to service providers, determine the level of access, and duration of that access, as well as the power to revoke access to personal data. The storage facility should ideally restrict any party from copying or modifying the data stored for data accuracy reasons.

Devices in the same household should be able to communicate over a local network and share data if needed and if the end-user permits such action to be taken. The smart device will still be connected to other devices but not directly to the Internet. Thus, any data that is to be transferred to the service provider will have to first be authorized by the end-user.

Different users tend to have different thresholds and appetites for various things, and privacy is no exception. This research categorizes end-users into three different categories, with room for other categories to exist in between, depending on the need for granularity. The three categories can be described as follows:

Overly Concerned with all User Data: Users should be allowed to share little to no information with the service provider. In this scenario, all data should be created, processed, and stored locally. Because service providers make money from the data they collect, the user should be comfortable moving from an advertisement-based secondary revenue model to a monthly service subscription or a higher one-time purchase price. The user also needs to acknowledge that some of the functions of the device may not be fully operational due to the limitations of sharing little or no data with the service provider.

Moderately Concerned with Some User Data: Users should be allowed to use parts of the software that they are comfortable with while leaving out and not agreeing to parts they feel are too invasive. In this category, the user should be allowed to choose what information they are comfortable sharing with the service provider. Based on the number of data points the two parties are in agreement with, a monthly subscription for the omitted data could be arranged.

Not Concerned with User Data at All: For security and general awareness concerns, the user should at least be made aware of what information is being collected from them and for what reason. This information should be put in layman's terms and be easy to understand.

Regardless of the category of user, they should not be bombarded with all the terms and conditions when they install an app or buy a new smart home device. The setup should be quick and easy, with each level or feature of the software educating the user on what that feature requires from the user and what granting the feature access to the requirements means for the end-user.

7 Conclusion

This paper looked at the processing and storage of user data created by smart home devices, highlighting that advancements in IoT, AI, and ML have relied on the availability of data for improvements while noting the issues related to user data collection by service providers. Related work in this field was examined, leading to the conclusion that there is a lack of literature in the field of IoT with an emphasis on smart homes and user data storage.

An alternative approach to storing user data, which follows a distributed data storage model, is proposed. This approach would make it possible to bridge the gap between upholding user privacy and the need for access to personal data to provide the convenience of smart home devices.

References

1. Atlam, H.F., Wills, G.B.: Technical aspects of blockchain and IoT. In: Advances in Computers, pp. 1–39. Elsevier (2019). https://doi.org/10.1016/bs.adcom.2018.10.006
2. Average number of connected devices residents have access to in U.S. households in 2020. https://tinyurl.com/mpfte866
3. IoT Architecture: the Pathway from Physical Signals to Business Decisions. https://tinyurl.com/2p93wjhc
4. Adadi, A.: A survey on data-efficient algorithms in big data era. J. Big Data **8**(1), 1–54 (2021). https://doi.org/10.1186/s40537-021-00419-9
5. Rosado, D.G., Gomez, R., Mellado, D., Fernandez-M.E.: Security analysis in the migration to cloud environments. Future Internet **4**, 469–487 (2012). https://doi.org/10.1109/EST.2013.13
6. Lebea, K., Leung, W.S.: A model for context-driven authentication in physical access control environments. In: Kim, K.J., Kim, H.-Y. (eds.) Information Science and Applications. LNEE, vol. 621, pp. 319–328. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1465-4_33
7. Smart home privacy: How to avoid data paparazzi. https://tinyurl.com/3t9tkccr
8. Wang, P., Ye, F., Chen, X.: Smart devices information extraction in home Wi- Fi networks. Internet Technol. Lett. 42 (2018). https://doi.org/10.1002/itl2.42
9. Wang, P., Chen, X., Ye, F., Sun, Z.: A smart automated signature extraction scheme for mobile phone number in human-centered smart home systems. EEE Access **6**, 30483–30490 (2018). https://doi.org/10.1109/ACCESS.2018.2841878
10. Hashizume, K., Rosado, D.G., Fernández-Medina, E., Fernandez, E.B.: An analysis of security issues for cloud computing. J. Internet Serv. Appl. **4**, 1–13 (2013)
11. Bertino, E.: Data security and privacy in the IoT. In: EDBT (2016)
12. Tabassum, M., Kosinski, T., Lipford, H.R.: “I don’t own the data”: end user perceptions of smart home device data practices and risks. In: Fifteenth Symposium on Usable Privacy and Security, pp. 435–450. SOUPS (2019). https://doi.org/10.1109/ISSA.2016.7802925
13. Choosing a Smart Home Hub?—Why Cloud vs Local Matters. https://tinyurl.com/sex5rpcm
14. Guhr, N., Werth, O., Blacha, P.P.H., Breitner, M.H.: Privacy concerns in the smart home context. SN Appl. Sci. **2**, 1–12 (2020)
15. Smart Home and Data Protection: Between Convenience and Security. https://tinyurl.com/399fkt6u
16. Suresh, S., Sruthi, P.V.: A review on smart home technology. In: 2015 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1–3. IEEE (2015)
17. What are the benefits of home automation?. https://tinyurl.com/tbazmxem
18. ADT Survey Reveals Strong Consumer Expectations for Smart Home Privacy Protections. https://tinyurl.com/5yf3hrxx
19. Wyze knew for years that hackers could remotely access its cameras, but didn’t tell anyone. https://tinyurl.com/t5nmk7b8
20. A Losing Game: The Law is Struggling to Keep Up With Technology. https://tinyurl.com/ynz4ce3n
21. Information Regulator in South Africa. https://tinyurl.com/36h7hp2v
22. Apple Says its Updated, Opt-In Prompts for User Data Tracking on iOS will Come into Effect. Next Week. https://tinyurl.com/4kdjyzhh

23. What you should know about smart home data collection. https://tinyurl.com/2tdjh4xn
24. Revenue Model Types in Software Business: Examples and Model Choice. https://tinyurl.com/fm882rbp.
25. The case for building a data-sharing culture in your company. https://tinyurl.com/mwa4kste
26. Many smart home device makers still won't say if they give your data to the government. https://tinyurl.com/336s9sa5
27. New Amazon Echo devices will have local voice processing, giving users more privacy. https://tinyurl.com/mwtfp3fc