User-centred privacy inference detection for smart home devices

Alexia Dini Kounoudes and Georgia M. Kapitsaki
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
{adini-01,gkapi}@cs.ucy.ac.cy

Ioannis Katakis
Department of Computer Science
School of Sciences and Engineering
University of Nicosia
2417, Nicosia, Cyprus
katakis.i@unic.ac.cy

Marios Milis
SignalGenerix Ltd
Limassol, Cyprus
marios.milis@signalgenerix.com

***Abstract*—In the smart home, vast amounts of data are being collected via various interconnected devices. Although this assists in improving the quality of life at home, often the user is not aware of the details concerning data collection apart from the information available on the provider privacy policy. It is however important to put the user inside this loop of information, so that she is well informed on possible uses of the data and the potential risks that this may entail. Previous works have identified user activity inside the smart home and have pointed out privacy threats. In this work, we go one step further by offering data inference techniques and giving this information back to the user. We use a number of machine learning techniques to draw conclusions about the user routines or activities and we inform the user about our findings concerning data inferences through a dedicated web application. Our aim is toward user-centred privacy and is a proof of concept that can be reused by smart home and Internet of Things service providers in general in order to improve the services offered to the end-users. Our results indicate that a large number of data inferences are possible by using a combination of techniques.**

***Index Terms*—smart home; privacy; user-centred privacy; inference detection;**

I. Introduction

In the era of smart homes, sensitive data are recorded and transmitted to multiple service providers. In most cases, such data are used to provide high-quality, useful services to the citizens. There are situations however, where the same data can reveal sensitive information about the user if obtained by an unauthorised party. Unintended inferences have become the biggest threat to privacy, mainly as a result of the development of sophisticated machine learning techniques [1].

Devices ranging from a smart bulb to a smart thermostat collect data continuously in the smart home. The risk of undesired inferences drawn from personal data becomes higher as the number of smart devices used increases. For example, smart metering data can disclose when a person is at home, the frequency a user watches TV, cooks, sleeps or goes on holidays. Fitness tracker data can reveal how often a person exercises or his physical condition. This type of information is invaluable to third parties like employers, insurance companies, or marketing companies, who are very interested in gaining access to it [2]. The EU "General Data Protection Regulation" (GDPR) has arrived to give control to the users over their data and the protection of their privacy, therefore the need for purposeful tools that allow the users to be informed and understand the privacy risks of using a smart home device is imperative [3].

A key factor in the evolution of the smart home is the ability to detect and recognise the activities taking place in this environment. Machine learning techniques have been widely used to detect and predict events and activities taking place in a smart home as well as residents behaviour, aiming to provide a better service by analysing the collected user, environmental and context data [4]. Historical records of such events can be exploited to discover patterns in user activities, identify anomalous or suspicious states, or to perform health monitoring and house surveillance. Furthermore, when smart home data is combined with data from other sources like social networks, further privacy vulnerabilities become apparent, especially when context information is considered. For instance, the social network profile of a user can be used to predict the user's whereabouts at a specific time conforming to knowledge released by friends [4].

Most of the approaches found in literature utilise the smart home data in order to provide better services, such as elderly monitoring, improvement of smart home applications, healthcare or security [5], [6], without taking into consideration the protection of the user privacy when handling the user's data. The diversity of our work is that the smart home data is utilised in order to explore how the user privacy can be compromised in a smart home scenario using machine learning techniques and inform the users about any inferences that can be drawn from their data.

In this work we focus on a smart home scenario in our effort to investigate whether the exploitation of data generated by smart home devices can lead to inferences about the occupants routines or other sensitive information. We are introducing a data inference testing framework using the data generated by a smart water meter and motion sensors deployed in a house, and by employing a number of machine learning techniques we aim to test whether such inferences are indeed possible. The results of the process are utilised in the PrivacyEnhACT privacy tool that we briefly present, which aims to inform the user about possible privacy vulnerabilities stemming from her smart home devices data and privacy preferences.

The main **contributions** of this paper are: a) We present a data inference testing framework tailored to the smart home. b) Our approach is able to inform the user about potential unwanted inferences, allowing them to perform appropriate adjustments in order to prevent them. c) We provide a proof of concept web application that implements the presented approach - a preliminary user evaluation is presented. d) The data used in the experiments are publicly available in the Zenodo portalfor replication purposes. To our knowledge, this is the first work that enables the user to be involved in the process of privacy inference detection by initiating the process.

II. Related work

The patterns of water consumption can be used to infer house occupancy, vacation periods, and even which rooms are being used and when, and occupants activities like using the toilet, bathing, etc. [7]. Data collected from motion sensors may contain contextual information like date, time of day or location, that can be exploited in order to infer single or multi-user presence in a house [8]. Machine learning models can be applied to motion sensor data to infer which particular rooms are occupied in a house or even the occupants identities, like in the work of Yang [9]. Fahim proposed a framework that analyses the electricity consumption of a house to unveil household characteristics relating to the economical or social status of the family, the number of occupants, or the home appliances being used [10]. McDonald used electricity data from smart meters to analyse the trends between different customers, obtaining information such as personal details about families, like the time they wake up and have breakfast, when they go to work, if someone stays at home during the day, etc [11]. Most of these works do not aim to utilise the detection of the inferences towards the user benefit, but for the amelioration of various services.

The data privacy vulnerabilities that are created from the linking of data from different smart devices are presented in the work of Zheng by validating how this linkage can reveal personal information about the owners [4]. Kroger provides an analysis on how sensitive information can be captured from data generated by smart devices, and be used to make additional inferences about habits, preferences, and so on [12]. The same author presents his findings in relation to the extraction of inferences from human speech and other sounds collected by smart speakers or other smart devices that record audio [13], and accelerometers [14]. Bilgin presents his findings regarding the likelihood of the improper use of smart meter energy consumption data in order to expose the routines and habits of people [15].

Chamarajnagar explores if location information and sensor temporal sampling can be used to compromise a smart home user's privacy [16]. A privacy model is proposed and preliminary analysis shows that it is possible to compromise the privacy of a user through inferences drawn from their data. The diverseness in our work is that the user is engaged through the provision of a a tool used to perform a privacy check on the smart home devices data and be informed about the privacy risks stemming from them. This dimension makes our work capable of satisfying the need for purposeful tools that give control to the users over their data, as per the GDPR.

Occupancy monitoring is an area that has attracted a lot of research interest, aiming to provide insights for improving energy or water consumption in buildings, cut down expenditure and enhance performance. However, in the smart home scenario, if the results of occupancy monitoring are misused then the privacy of the occupants can be compromised. Siraj developed a framework that can prevent occupancy detection in a smart home through the use of adversarial machine learning techniques [17]. The framework aims to enhance the protection of the users' privacy, offering customised user privacy preferences. Singh exploits heterogenerous sensors and machine learning techniques such as Linear Discriminant Analysis (LDA) and Random Forest, in order to estimate the number of persons in a room [18], while Yang shows that by using smart meter and motion sensor data, occupancy related inferences are possible, like the number of occupants or even their identities [9]. While in our work we aim to draw similar conclusions from the data, we take the procedure to the next level by making the process user-centred in our effort to make the users aware of the privacy risks originating from their data.

Activity detection in smart homes is another area that has received a lot of attention, mostly for Ambient Assisted Living applications, where the user data is being used to provide a better service. Yet, most of the works do not acknowledge the privacy risks this process entails for the user. Wilson proposes a methodology to infer the activity profiles of households using smart meter data [19]. In the work of Yassine [20], the authors exploit smart meter data in order to discover activity patterns using machine learning techniques for healthcare applications, aiming to provide assistance to the elderly living alone when anomalous activities are detected.

The insights obtained in some of the above works form a privacy threat, calling for solutions like the one we propose, where the users are being notified about the privacy vulnerabilities of their data and can take actions towards protecting their privacy. The main advantage of our approach is that we detect inferences from smart home data in order to make the users aware that such threats exist so that they can decide whether they wish to continue using the services or adjust their privacy preferences in a way that the data created cannot be used for activities that could compromise their personal sphere.

III. Methodology

In this section we describe the methodology we used in order to examine and analyse the data in the smart home scenario, using data from a smart water meter and three

Arduino-based motion sensors. Even though we apply this methodology on a specific setting, the described approach can be replicated in other scenarios with similar sensor availability, for example smart watches or fitness bands that create and collect personal user data.

Overview. The methodology consists of four steps. The first step is to examine and clean the sensor data for analysis. We have identified two approaches for organising and analysing the data, which are discussed in part A of this section. In the second step, we perform clustering analysis on the data, and in the third step we exploit the clustering results focusing on their interpretation, in order to give a meaning to the clustered data and identify what inferences can be drawn. For this reason, we train a decision tree model to generate a set of rules, which can be used combined with the insights obtained from visual observations of the data. In the last step we use the clusters defined in Step 2 as new features in our training data and predict to which cluster new data points are assigned to. Based on this, we process new data aiming to detect if a number of inferences can be drawn from them. Then we inform the users about the identified inferences which they can tackle possibly by altering their privacy preferences or settings, as these affect the data collection [21]. The methodology was applied in the following cases/experiments:

Acknowledgment

This work was partially funded by the Research Promotion Foundation (RPF) of Cyprus RESTART 2016-2020 TAMIT (Novel Trustworthy Automatic Metering IoT Solution for Smart Cities Applications) project (grant agreement number ENTERPRISES/0618/0027). The authors would like to thank the Waterboard of Limassol, a partner in the TAMIT project, for the contribution towards the initial development of the research.

References

[1] S. Wachter and B. Mittelstadt, "A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai," *Colum. Bus. L. Rev.*, p. 494, 2019.

[2] S. R. Peppet, "Regulating the internet of things: first steps toward managing discrimination, privacy, security and consent," *Tex. L. Rev.*, vol. 93, p. 85, 2014.

[3] A. D. Kounoudes and G. M. Kapitsaki, "A mapping of iot user-centric privacy preserving approaches to the gdpr," *Internet of Things*, vol. 11, p. 100179, 2020.

[4] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart internet of things systems: a consideration from a privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[5] J. Wang, N. Spicher, J. M. Warnecke, M. Haghi, J. Schwartze, and T. M. Deserno, "Unobtrusive health monitoring in private spaces: The smart home," *Sensors*, vol. 21, no. 3, p. 864, 2021.

[6] S. T. M. Bourobou and Y. Yoo, "User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm," *Sensors*, vol. 15, no. 5, pp. 11 953–11 971, 2015.

[7] G. Eibl and D. Engel, "Influence of data granularity on smart meter privacy," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 930–939, 2014.

[8] E. Oriwoh and M. Conrad, "Presence detection from smart home motion sensor datasets: a model," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. Springer, 2016, pp. 1249–1255.

[9] L. Yang, K. Ting, and M. B. Srivastava, "Inferring occupancy from opportunistically available sensor data," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2014, pp. 60–68.

[10] M. Fahim and A. Sillitti, "Analyzing load profiles of energy consumption to infer household characteristics using smart meters," *Energies*, vol. 12, no. 5, p. 773, 2019.

[11] B. McDonald, P. Pudney, and J. Rong, "Pattern recognition and segmentation of smart meter data," *ANZIAM Journal*, vol. 54, pp. M105–M150, 2012.

[12] J. Kröger, "Unexpected inferences from sensor data: a hidden privacy threat in the internet of things," in *IFIP International Internet of Things Conference*. Springer, 2018, pp. 147–159.

[13] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy implications of voice and speech analysis–information disclosure by inference," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 242–258.

[14] J. L. Kröger, P. Raschke, and T. R. Bhuiyan, "Privacy implications of accelerometer data: a review of possible inferences," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 81–87.

[15] Z. Bilgin, E. Tomur, M. A. Ersoy, and E. U. Soykan, "Statistical appliance inference in the smart grid by machine learning," in *2019 IEEE 30th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*. IEEE, 2019, pp. 1–7.

[16] R. Chamarajnagar and A. Ashok, "Privacy invasion through smarthome iot sensing," in *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2019, pp. 1–9.

[17] A. Siraj *et al.*, "Avoiding occupancy detection from smart meter using adversarial machine learning," *arXiv preprint arXiv:2010.12640*, 2020.

[18] A. P. Singh, V. Jain, S. Chaudhari, F. A. Kraemer, S. Werner, and V. Garg, "Machine learning-based occupancy estimation using multivariate sensor nodes," in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–6.

[19] C. Wilson, S. Lina, V. Stankovic, J. Liao, M. Coleman, R. Hauxwell-Baldwin, T. Kane, S. Firth, and T. Hassan, "Identifying the time profile of everyday activities in the home using smart meter data," 2015.

[20] A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13 131–13 141, 2017.

[21] A. Dini Kounoudes, G. M. Kapitsaki, and M. Milis, "Towards considering user privacy preferences in smart water management," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019, pp. 209–212.