Towards Privacy-Preserving IoT Data Publishing

Mohammad Malekzadeh, Hamed Haddadi
Queen Mary University of London
m.malekzadeh/hamed.haddadi@qmul.ac.uk

Abstract

The abundance and availability of personal information online and from ambient sensors has lead to an increasing rate in occurrences of privacy breaches and potential security harms to individuals. Yet, access to these information is critical for the success of many beneficial applications such as population health studies or urban planning. During this research paper we design an integrated sensing framework in order to better handle the incoming continuous data stream (time-series) from sensors embedded in different kind of data sources. We provide a trusted zone for managing access to personal time-series data to provide utilities and gains while protecting the individuals' privacy.

1. Introduction

Smart devices around us are becoming more and more interconnected and thousands of measurements captured by sensors every minute, resulting in a wide variety of newly generated data. From smart meters in homes to smart shirts for athletes, to smart beds for elderlies, almost every device in our everyday life has the ability to produce data. Among all the types of data, data from sensors is the most widespread and is referred to as time-series data. Many entities want to access this so-called Internet of Things data over time for conducting data mining, and users also benefit from sharing their data with them.

Some of these applications focus on data collected for personal purpose (e.g., health monitoring), and others are developed for the collection of sensing data at a community-wide level in order to contribute to population studies(e.g. urban planning). This imperceptible data collection can also lead to major personal privacy concerns. Since detailed person-specific data in its original form often contains sensitive information about individuals, anonymously publishing such data without any revelation of sensitive information to third parties is a challenging task. Moreover, privacy concerns and issues arise when outsourcing this data to the cloud, because data subject losses the control over data.

There have been a number of attempts at addressing this need, but these are generally fraught with the shortcomings of centralisation and/or implicitly overexpose personal data to third parties. Recently, Databox [1] has aimed to solve this challenge by enforcing accountability and control by design at the users' end. The Databox serves as a platform upon which the processing of personal data can be done locally, in the context of containerised Databox apps, and only emit the necessary results to a third party. In this research paper, we describe a framework (as a part of Databox) for publishing IoT time-series data in an aggressive environment, so that the published data remains practically useful while individual privacy is preserved.

2. Research Questions and Challenges

In the near future, IoT allows for omnipresent data gathering and user tracking, but when these advantageous features designed improperly, can lead to privacy breaches that are limited the success of IoT vision [3]. The wide variety of connected devices that form the IoT, different types of data recorded by their sensors, and multiplicity of communication protocols, leading to inherent data security and privacy risks. The main questions is how desired statistics of personal time-series data can be released to an untrusted third party without compromising the individual's privacy. More precisely, how to transform time-series in such a way that after releasing, individual's sensitive information cannot be inferred from transformed time-series.

Analysis of time-series accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend, seasonal variation, cyclical or Irregular component). Prior approaches to privacy were devised for the static cases, and when they are extended to the dynamic and interactive cases of time-series data, scalability challenges arise. On the other hand, time-series are highly rich in information about users' behavior, particularly when continuously collected. This challenge makes it difficult to protect the privacy of one sensitive behaviour in isolation of the others and it limits the mechanisms that can be used. Finally, a principal task in privacy-preserving data publishing is the definition of proper measures that can truly assess the privacy-utility trade-off. Most of proposed measures focus on protecting the identity of a participant without considering the information content of the corresponding data. An acceptable solution for time-series publishing needs to accurately measure the amount of original information that is contained in the transformed data.

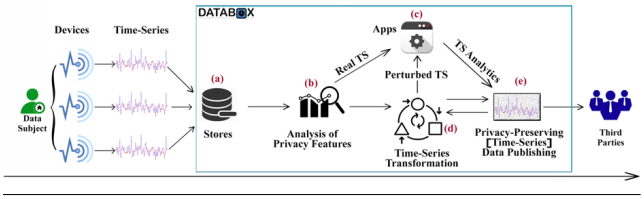

Figure 1. An architecture for privacy-preserving time-series data publishing

3. Sensing Framework

The pervasiveness of smartphones as well as the variety of their onboard sensors have enabled the automated acquisition of large scale data. Operating systems such as Android and iOS provide application programming interfaces (APIs) to access these sensors. SensingKit[1] is a new mobile sensing framework, compatible with iOS and Android devices, which enables capturing motion, location, and proximity data and transmitting them to a server. Sensors can be dynamically configured and data can be extracted in both CSV and JSON format [2].

We will take advantage of SensingKit architecture for making some application for smart devices and configure them in a way that each one reports its sensing data in a continuous manner and proper format to the Databox. Then, we will empower it with a robust component that integrates and stores these time-series data in a containerised trusted store for future processing (Figure 1 (a)). Stored time-series will be managed by a Time-Series Data Base (TSDB) in a centralised storage resides in Databox.

4. Initial Approach and Future Work

Currently, there are two major approaches to time-series data publishing for analysing purpose. One solution is sharing the data with trusted third parties without any alteration based on non-disclosure agreements. Another one is transformation (e.g. adding noise) of data before publishing them to the untrusted third parties. However, there are drawbacks to both of them: difficulties in trusting others and the trade-off between information loss and user privacy. Additionally, since in the both approach data typically resides in a shared environment such as cloud servers, users will know neither the exact location of their data nor the other sources of the data collectively stored with theirs.

Considering these problems, we propose a privacy by design solution for privacy-preserving IoT data publishing through the Databox [1]. Awareness of threats on collected time-series that can jeopardise user privacy, will help users to choose whether to provide apps their raw data or applying some transformation before granting access to them (Figure 1 (b)).

A Databox app (Figure 1 (c)) is an app specially designed to run on a Databox. Data subject can download a Databox app proposed by a third party and run it on their own device. Third parties' apps installed on Databox are able to request for several data from different sources and perform desired analytics on the large quantity of data. For example, report the relation between home temperature and user heart rate (e.g. one data is collected from a temperature sensor and another from a wearable watch). Note that, such analytics are only available via a centralised TSDB per each user. Results of these analytics are summarised and only reported to the user of Databox and never leave it in their original forms.

In order to benefit from data mining techniques, Databox apply a transformation which reduces the risk of privacy violations of the underlying data (Figure 1 (d)). Users have different preference in various scenarios: they are very conservative in some situations, and less concern in other cases. For this purpose, we will build a dashboard for users to control their privacy and deploy it on Databox. There are several motivations for publishing personal data gathered by IoT devices (e.g. to what extent user's TV watching hours is different from the average population). When the processing of users' data is completed, users can allow their apps to send the results back to their providers via a privacy-preserving data publishing method (Figure 1 (e)). The main goal of this method is to prevent the reconstruction of original time-series from transformed ones, yet allow to accurately estimate some statistics despite the perturbation.

5. Conclusion

The quantity of personal data people generate is being increased on a daily basis. Analysing such data without any revelation of sensitive information to an unauthorised party is the main concern of the next generation of IoT businesses. In this paper we described a privacy by design solution for privacy-preserving IoT data publishing through the Databox. Authorised third parties interested in the personal information are granted access by this envisioned component. Our proposed framework allows user to benefit from data analysis tools while still being safe from invasion of privacy.

References

[1] Hamed Haddadi, Heidi Howard, Amir Chaudhry, Jon Crowcroft, Anil Madhavapeddy, Derek McAuley, and Richard Mortier. Personal data: thinking inside the box. In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*, pages 29–32. Aarhus University Press, 2015.

[2] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. Sensingkit: Evaluating the sensor power consumption in ios devices. In *12th International Conference on Intelligent Environments (IE'16)*, 2016.

[3] Jan Henrik Ziegeldorf, Oscar García Morchon, and Klaus Wehrle. Privacy in the internet of things: threats and challenges. *Security and Communication Networks*, 7(12):2728–2742, 2014.

[1] `http://sensingkit.org/`