COMPUTER LAW & SECURITY REVIEW ■■ (2017) ■■–■■

Available online at www.sciencedirect.com

ScienceDirect

www.compseconline.com/publications/prodclaw.htm

Computer Law & Security Review

Data integration in IoT ecosystem: Information linkage as a privacy threat

Nishtha Madaan [a,*], *Mohd Abdul Ahad* [a], *Sunil M. Sastry* [b]

[a] *Faculty of Engineering and Technology, Jamia Hamdard University, New Delhi, India*
[b] *Sastry & Co., Nagarthpet, Bangalore, India*

Keywords:
Internet of things
Data integration
Information linkage
Privacy
Heterogeneous IoT ecosystem

A B S T R A C T

Internet of things (IoT) is changing the way data is collected and processed. The scale and variety of devices, communication networks, and protocols involved in data collection present critical challenges for data processing and analyses. Newer and more sophisticated methods for data integration and aggregation are required to enhance the value of real-time and historical IoT data. Moreover, the pervasive nature of IoT data presents a number of privacy threats because of intermediate data processing steps, including data acquisition, data aggregation, fusion and integration. User profiling and record linkage are well studied topics in *online social networks (OSNs)*; however, these have become more critical in IoT applications where different systems share and integrate data and information. The proposed study aims to discuss the privacy threat of *information linkage*, technical and legal approaches to address it in a *heterogeneous IoT ecosystem*. The paper illustrates and explains information linkage during the process of data integration in a smart neighbourhood scenario. Through this work, the authors aim to enable a technical and legal framework to ensure stakeholders awareness and protection of subjects about privacy breaches due to information linkage.

1. Introduction

The *Big data* revolution and the emergence of *Internet of things (IoT)* has led to large-scale analyses of data generated from heterogeneous devices in various scientific and governance domains. The data processing tasks include data acquisition, fusion, aggregation and integration (Ahad and Biswas, 2017). Often, in privacy-sensitive domains such as healthcare and smart cities, individual data streams that potentially lead to a privacy threat are identified and anonymized before data processing to prevent any privacy breaches. Consider the scenario where users have been introduced to many online social networks such as Twitter, Instagram, and LinkedIn. Due to diverse functionalities, different online social network platforms attract users for different purposes such as information seeking, sharing and social connection maintenance (Shu et al., 2017). Literature on social network analysis points that *social network anonymization* refers to the process to replace each user's unique identifier (example, username) with a random string, but the network structure remains vulnerable for active and passive analyses (Shu et al., 2017). In addition, network structure from different platforms may be exploited and correlated for record linkage that reveals user's identity. Therefore, often concerns of identification, location tracking and profiling are threats posed by integration of data about users of online social networks and open data released by the government.

In case of healthcare, medical records of a patient are anonymized but when these records are analysed in conjunction with demographic and behavioural data they may uniquely

* *Corresponding author.* Faculty of Engineering and Technology, Jamia Hamdard University, Hamdard Nagar, New Delhi 110062, India.
E-mail address: madaan.nishtha@gmail.com (N. Madaan).
http://dx.doi.org/10.1016/j.clsr.2017.06.007

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

identify the patients and their medical conditions. This is especially critical in case the patient is suffering from depression. On one hand, the information integration tasks are important to gain insights from data analyses and on the other, they may compromise personal privacy of a subject. The autonomous nature of IoT aggravates these privacy threats (Ziegeldorf et al., 2014).

1.1. *Information linkage in an IoT ecosystem*

In case of IoT ecosystems, such as smart homes and cities, data acquired from different streams originating from various devices and subsystems undergo fusion, aggregation and integration to ensure QoS (quality of service). During data integration, attributes of devices and data belonging to different services are correlated and integrated. At times, this reveals information or insights about subjects, their demographic location and activities, which lead to severe privacy concerns. Therefore, user profiling, localization and tracking, and information linkage are some of the critical challenges that need to be addressed for data processing in IoT ecosystems.

A recent study highlights that in the IoT domain, increasingly threats of privacy-violating interactions and presentations, life-cycle transitions, inventory attacks and information linkage arise during data aggregation and standardization phases (Ziegeldorf et al., 2014). This has received limited attention from the research community. The proposed study explains the issue of *information linkage* due to data integration tasks in IoT ecosystems and possible implications on personal privacy. It discusses the technical and legal solutions to reduce the risk of unprecedented information linkage and protection of stakeholder's rights over data-use and sharing.

1.2. *Considerations for privacy concerns*

IoT ecosystems such as healthcare and smart cities comprise a number of heterogeneous devices, services, and stakeholders performing a variety of tasks that utilize data about individuals and their surroundings. Often in such complex systems, the context for which the services are authorized to collect data may vary from the context of actual data-use. Therefore, the sensitivity associated with a device, and context of data collection may not hold for secondary-level data processing, as it is impossible for a single sensor or device to collect all the required data to deliver a service or for a governance purpose. Therefore, in such systems, data is often shared, and re-used for a variety of data analyses. This implies that data sensitivity described by data owners or publishers in one context may not hold for another and it is difficult for them to assess all possible data processing tasks a priori. In addition, with the data sharing and processing tasks possible at various levels of granularity, it becomes more complex to identify where information linkage might occur. For example, in case of an IoT ecosystem, information linkage might occur at device level, metadata level, data-stream level, or while sharing statistics and analytics on the data. For this, further sections draw on related work on privacy preserving data mining in "Big Data" (Xu et al., 2014; Perera et al., 2015).

A number of organizations are now aligning to "Privacy-by-Design" practices to minimize the privacy concerns during data collection phase. These practices have minimal impact on privacy concerns that arise due to secondary level data analysis such as fusion, integration and aggregation. During the data collection phase, adoption of principles such as data minimization, limited collection and purpose specification reduce the privacy concerns (Cavoukian, 2011). However, the downside of these principles is that some data useful for obtaining insights might be missed. In addition, increasing number of algorithms on data anonymization in streams, privacy-preserving data processing are proposed to address data privacy but these algorithms do not address privacy concerns as a result of re-identification of anonymized data. In case of an IoT ecosystem, integration of metadata, data and services is mandatory for enabling more useful, rapid and efficient development of applicative services (Mainetti et al., 2015). Therefore, it is important to understand the granularity of data integration and its scope to identify and address the privacy concerns in an IoT ecosystem. In addition, understanding ways in which stakeholders can be made aware of the privacy risks their resources may encounter in these IoT ecosystems is critical.

Roadmap. The following Section 2 describes a heterogeneous IoT ecosystem and various data integration tasks performed in it. Next, the challenges to address the issue of information linkage are described. The study describes a hypothetical case of a smart neighbourhood (an aggregation of smart homes) in Section 3. The section also explains data integration among smart home devices and its impact on possible privacy breach of *information linkage*. Further, an in-depth analysis of possible technical solutions is given in Section 3.1. An overview of corresponding legal solutions to protect stakeholders' rights for data-use is described in Section 4.2. Finally, the study concludes with next steps in Section 5.

2. Heterogeneous IoT ecosystem

This section explains the dynamics of a heterogeneous IoT ecosystem, the task of data integration, and concern of information linkage. Heterogeneity of any IoT ecosystem can be defined from different perspectives, as it is not a standalone system. It can be viewed as a layered framework that comprises layers of sensor devices, subsystems that use data from different devices, services that share and re-use data. It may also contain an application layer that assists the subsystems and stakeholders who interact with these complex system-of-systems and a gigantic network of sensors. These layers involve various tasks of complex data management pipeline.

Fig. 1 describes a smart home setup with a number of sensors and IoT devices, such as gas monitor, smoke alarm, temperature and light monitors, and mail notifiers. These devices interact with subsystems such as alarm notification system that in turn generates a notification in form of an emergency call or SMS to the home-owner or service provider. Smart metering system installed in the home generates visualizations for users' usage of utilities that help them optimize their usage based on measurements recorded by other systems such as light and temperature devices. It is important to note that, in some smart home systems, these systems and devices may

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

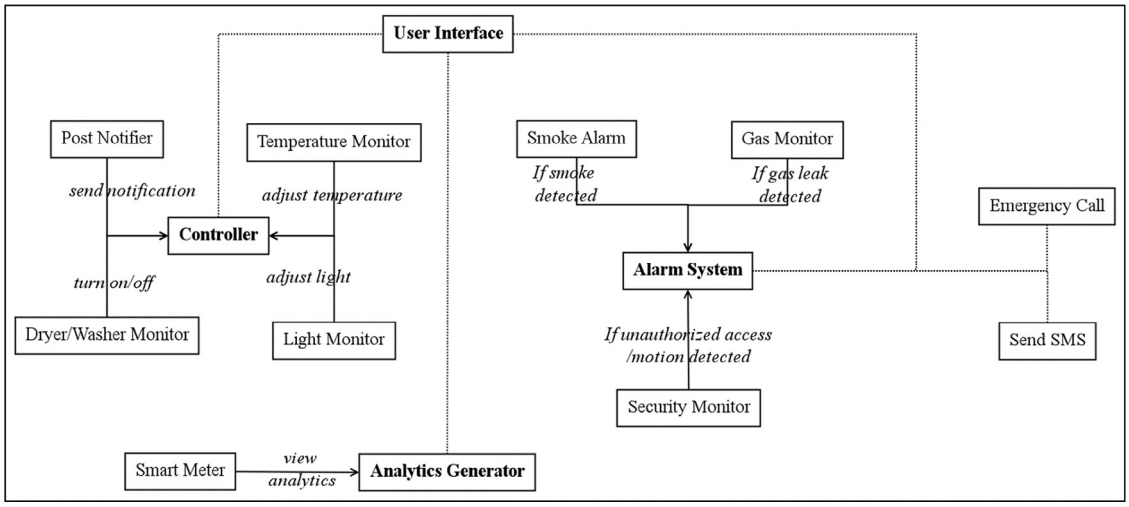

Fig. 1 – An example setup of devices, control systems and other subsystems in a Smart Home.

belong to same manufacturers who provide complete solutions, whereas in others, a homeowner may assemble it using devices from different manufacturers. As described above, the different subsystems share data for optimizing system behaviour. In the next section, a *smart neighbourhood* setup is described where different subsystems of a smart home share their data and analytics for a number of services, and privacy concerns that arise due to these interactions.

2.1. *Data management in heterogeneous IoT ecosystem*

Data management activities are complex in any IoT ecosystem. Since IoT is largely an industry driven technology, a number of protocols and standards for device discovery, data semantics and distribution co-exist. Due to the lack of consensus, data from different devices and within systems remains isolated, making data integration tasks complex. Existing work focuses on the lack of interoperability standards for data with IoT applications being implemented on top of a variety of application-level protocols and frameworks, including the Constrained Application Protocol (CoAP),[1] Representational State Transfer (REST),[2] Extensible Messaging and Presence Protocol (XMPP),[3] Advanced Message Queuing Protocol (AMQP),[4] Message Queue Telemetry Transport (MQTT),[5] MQTT for Sensor Networks (MQTT-SN)[6] and Data Distribution Service (DDS)[7] (Al-Fuqaha et al., 2015). During device-to-device communication occurs, and when access is shared among devices, the data conversion from one format to another in real-time is highly complex. For such horizontal integration, enhanced MQTT protocol (Al-Fuqaha et al., 2015) and regional standards such as Hypercat (Beart et al., 2015) are proposed which do not sufficiently address the problem.

A possible data management framework for IoT incorporates a layered and federated paradigm to join independent IoT subsystems in an adaptable, flexible, and seamless data network (Elkheir et al., 2013). For any analyses in an IoT subsystem, data fusion, aggregation and data integration are performed at the device (actuator) level and application level. Data integration creates real value by providing insights on IoT data from a number of streams. Often this data is combined with demographic and social media data for complex services and applications in various IoT enabled systems. As described in the previous section the focus of the given study is to understand the privacy concern of "information linkage" during data integration. Therefore, in the following subsections, data integration and information linkage is described in detail.

2.2. *Need for data integration*

A huge volume of data is collected by the network of billions of IoT devices that requires data integration engine to support decision making in heterogeneous subsystems. According to a recent report from *Gartner and Bit Stew Systems*, around fifty percent of the implementation cost of IoT solutions will be used on integration of various participating entities like devices, databases, and other third party middleware system(s) (Hans, 2017).

As described earlier, due to the pervasive nature of IoT ecosystems, devices and sensors collect sensitive personal information about subjects and their surroundings. Data integration is a necessity in various IoT ecosystems, such as smart cities and its subsystems. Different agencies share data to gain deeper insights into user requirements and design services usable for the citizens. For example, in the healthcare domain, data integration can support emergency preparedness and public safety when public health data is integrated with demographic and social media data. However, due to high sensitivity of data in such collaborative domains, it is critical to develop techniques that enable data integration and sharing without any privacy breaches. Sensitive device data even if anonymized when integrated with other sensor data can help

[1] http://coap.technology/.
[2] http://restfulapi.net/.
[3] https://xmpp.org/.
[4] https://www.amqp.org/.
[5] http://mqtt.org/.
[6] http://mqtt.org/tag/mqtt-sn/.
[7] http://www.omg.org/spec/DDS/.

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

re-identify subjects and may result in unintended information linkage. For instance, a collection of timestamps may not be seen as personal data, but they could contribute to data linkage and influence on privacy risks (Danezis et al., 2014).

Integration is a vital prerequisite before mining the data that is automated through machine learning techniques (Clifton et al., 2004). While this problem is well studied in databases (Xu et al., 2014; Perera et al., 2015) the characteristics of IoT data pose newer and more severe challenges for data integration especially the heterogeneity of data, a single device playing multiple roles in different applications and contexts, stronger time and space correlation. The metadata related to geographic location of data and timestamps associated with data can reveal individuals' coordinates and availability. Several critical challenges, such as user profiling, location and tracking and information linkage need to be addressed for a privacy preserving analysis and data sharing tasks in complex and dynamic IoT ecosystems. The following section focuses on need of information correlation and risk of information linkage in IoT ecosystems.

2.3. *Understanding information linkage*

Information linkage is often referred to linking of different systems such that the combination of data sources reveals information that was previously not disclosed by the individual sources. In addition, these sources did not intend any revelation of such information (Ziegeldorf et al., 2014). This may be possible because of a mismatch between context of data collection and its use. Alternatively, there may be a mismatch between the purpose-of-use for which *consent* was given and the purpose it is used (Ziegeldorf et al., 2014). Among the different privacy threats such as profiling and location tracking, information linkage is complex as it surfaces during secondary data processing activities in contrast to the data collection phase.

According to the study (Ziegeldorf et al., 2014), privacy violations occur due to overlooking of privacy protection mechanisms, resulting in unauthorized access and leaks of private information especially when systems collaborate and combine data sources. The study also stresses on privacy violation due to the linkage of data sources and systems as an increased risk of re-identification of anonymized data. It is important to note that information linkage commonly appears when data from multiple sources is combined resulting in re-identification of subject, user profiling and other critical privacy threats.

In the IoT domain, information linkage is critical due to the following reasons.

(1) **Horizontal integration** is almost inevitable in any IoT domains because of a number of manufacturers and each following different metadata standards (explained in section 2.1). When a user assembles a smart system, he may choose to use devices from different companies. This creates a system-of-systems when delivering new services that no system can provide on its own. Therefore, collaboration will require agile exchange of data and controls between the different parties. However, as horizontal integration contains more local data flows than vertical integration, it could provide a way to enhance privacy (Ziegeldorf et al., 2014).
(2) **Linkage of systems** renders data collection in the IoT less transparent than expected from the predicted passive and un-intrusive data collection by smart things (Ziegeldorf et al., 2014).
(3) Moreover, in case of IoT ecosystems, the **schema or the metadata** of devices keeps changing continuously, resulting in algorithms surpassing possible information linkages.

An analogy can be drawn between horizontal data integration in IoT subsystems and those observed in case of online social networks (OSN) and their applications (Ziegeldorf et al., 2014). In case of latter only two parties are involved the OSN and the third-party application, while the IoT is expected to feature services that depend on the interaction and collaboration of many co-equal systems. Ziegeldorf et al. (Ziegeldorf et al., 2014) describe the most important technical challenges for ensuring systems-of-systems IoT ecosystems:

(1) Transparency in the information exchange among different subsystems and if appropriate user acceptance has been gained for it.
(2) An access control model that considers collaborating linked systems is a critical requirement in the IoT domain.
(3) Modification of anonymization methods to cater to the needs of linked systems and can address against combination of different sets of data.

In the following section, we describe the challenges that need to be addressed to overcome privacy threat of "Information Linkage" with an example in a smart neighbourhood (an aggregation of smart homes). Smart home is assembled from devices and solutions provided by different manufacturers. It describes privacy concerns and illustrates the technical challenges to build a *privacy profile* in terms of metadata of the devices and data generated by them to prevent information linkage during data integration. Further, it describes the need of a legal policy framework for protecting rights of citizens over use of their data by a variety of service providers in a number of application contexts.

3. A case of smart neighbourhood

In this section, a smart neighbourhood scenario is explained which utilizes data from smart home environments. The metadata of NEST[8] products are considered for understanding linkability of attributes and granularity of the information linkage. Smart neighbourhood can help citizens create ambient living environments by providing smart garbage collection, energy efficient environment (charging electric cars or adjusting thermostats during good weather), and more secure neighbourhood among others. However, when data is collected, aggregated and integrated from individual smart homes, a number of privacy

[8] https://nest.com/uk/.

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

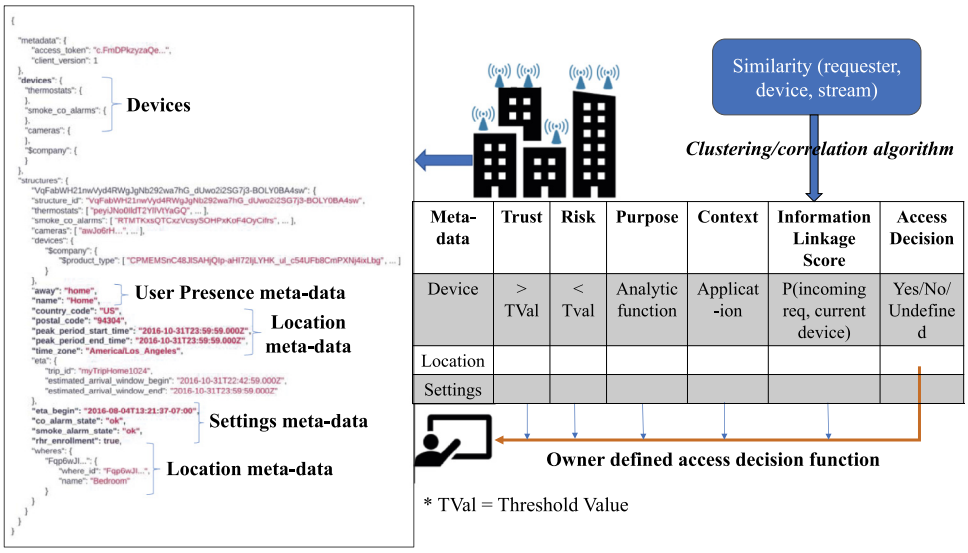

Fig. 2 – An example of Google's nest data model and a smart neighbourhood scenario where a user can share his device data and meta-data based on the considerations of trust, risk, information linkage score, purpose of access or sharing and context of data-use.

concerns also arise, such as availability of a person may be known to unauthorized parties or usage of certain counsel services may be estimated (for example, smart meter data). In literature, *Unlinkability* ensures that privacy-relevant data cannot be linked across domains that are constituted by a common purpose and context. This implies that the processes have to be operated in such a way that the privacy-relevant data are un-linkable to any other set of privacy-relevant data outside of the domain. Mechanisms to achieve unlinkability comprise data avoidance, separation of contexts (physical separation, encryption, usage of different identifiers, access control), anonymization and pseudonymization (Danezis et al., 2014). These solutions now need to be extended to the IoT ecosystems.

Nest Labs is a home automation producer of programmable, self-learning, sensor-driven, Wi-Fi-enabled thermostats, smoke detectors, and other security systems (NEST, 2017). It uses a machine-learning algorithm that requires the users to regulate the thermostat for initial few days to obtain a reference data set (NEST, 2017). In case of Nest, all the devices are associated with a *structure* and any number of devices can be added to this *structure*. Device metadata can be configured through a user-interface. The devices read the data model (of a home) to understand their target settings and adjust the mode accordingly. Data is stored in the Nest service as a shared JSON document (as shown in Fig. 2). When a subscribed data value changes, the data values in the Nest service are updated to reflect the *new state*. The Nest API is represented as a JSON document with top-level attributes of metadata, devices and structures. Every data element in the JSON document is addressable by URL (also known as "data locations") and its sections can be read, written and subscribed to ensure real-time changes to the system.

Consider a scenario, where homeowners share their energy consumption readings with the service provider to obtain competitive pricing. This requires sharing the metadata of "thermostat" along with metadata of *expected_time_arrival* (ETA). Simultaneously, they choose to share their video stream from their camera for additional security arrangements in a block by setting the attribute *is_public_share_enabled* attribute. In such a case, *has_motion* attribute is also shared (as shown in Fig. 2). Further, these two streams may be integrated to deploy community security personnel and equipment. This generates a threat of *attribute linkage* that may reveal subjects' presence and absence at home at different times during a day. Any malicious attack on the system can cause major privacy threat. Over a period of time, when the data and geo-spatial information from these devices are integrated with the geo-data and open data available and published by government agencies, the identity of each of the subjects can be revealed, leading to serious privacy concerns. Therefore, in such scenarios, a fine-grained distributed access control, specification of a privacy profile for the devices and data-streams can support a more secure yet smart neighbourhood in the cyber-physical worlds.

Fig. 2 describes the basic elements to be considered for specifying the elements of a privacy profile for granting access or share permission to a device or data in a smart home scenario. For the "purpose" for which the data will be used in terms of the analytic functions, visualizations are an important consideration. Next, the "context-of-use" in terms of application domain is critical for the users to grant access over their personal information. The "risk" and "trust" measures can be defined in terms of sensitivity of metadata attributes of participating devices to estimate which information the subjects intend to share. The probability of information linkage between metadata and data attributes is approximated based on specific "analytic function" intended by the requester. These parameters allow subjects to grant access to data from their sensors and devices. A number of similarity based methods, distance based methods (e.g. Euclidean, Minkowski) and unsupervised clustering methods (e.g. K-means, K-median) can be used to determine the probability of information linkage between the meta-data attributes of devices and data points

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

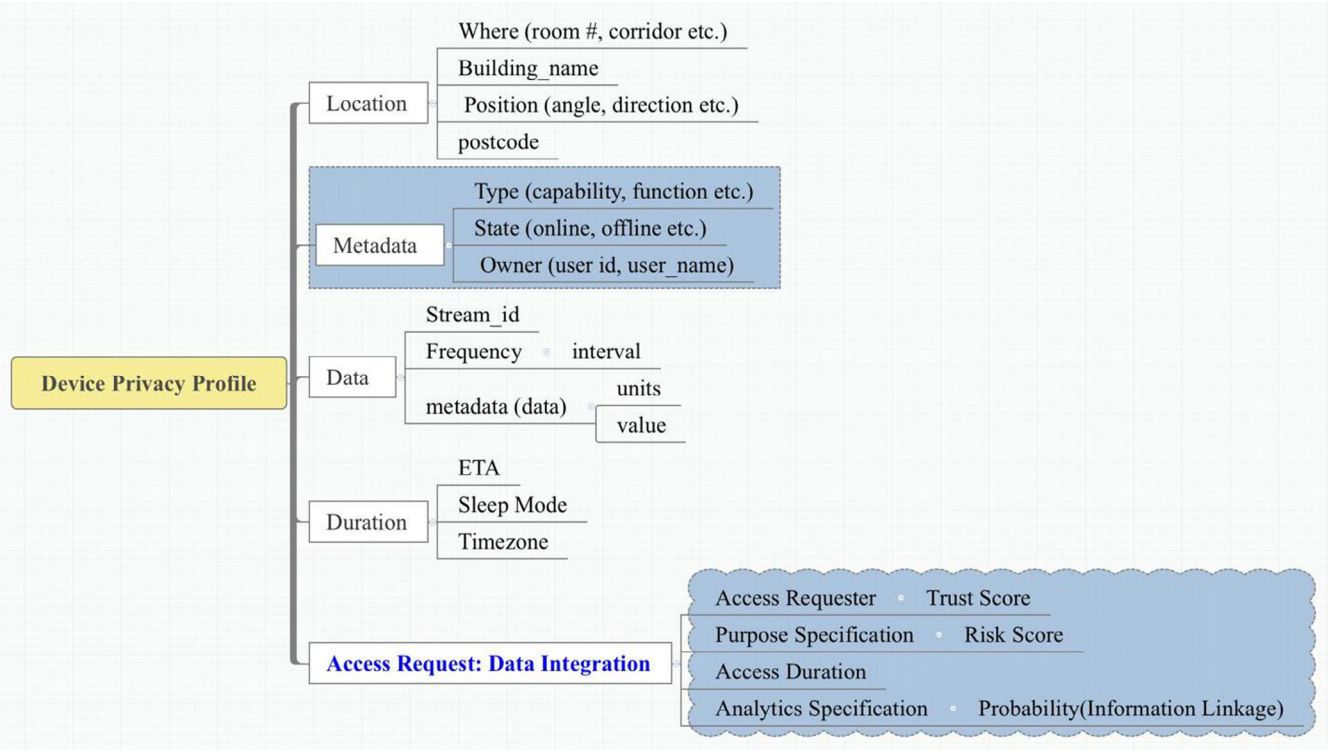

Fig. 3 – An indicative privacy profile of an IoT device to grant data access for the purpose of data integration.

of multiple streams. The main aim of this study is to identify the components of a privacy profile, and other technical solutions to enable stakeholders understand the privacy threat of information linkage.

3.1. *Privacy in distributed data integration: technical solutions*

As described above, information linkage occurs during a number of data integration tasks when data is shared or re-used by different services and stakeholders in an IoT ecosystem. Ziegeldorf et al. (Ziegeldorf et al., 2014) indicate solutions such as distributed access control and anonymization across streams but do not give details how these solutions can be implemented. The proposed study draws from their work and describe the use of "informed consent" with respect to data integration tasks as a solution to bridge the gap between "context" of data collection and use.

3.1.1. *Privacy profile for devices and data*

Metadata of smart sensors and devices can be utilized for defining privacy profile. "Privacy profile" can be defined as device or data preference settings for data sharing and integration purposes. A device can have multiple privacy profiles based on access requesters, and purpose of access. Each privacy profile needs to be associated with a risk function that can be defined in terms of the purpose, duration, analytic function and probability of information linkage between the given data-stream from the device in question and the other streams with which the integration function is invoked. In addition, the device metadata, its location attributes and device activity functions (mentioned as *duration* in Fig. 3) play a crucial role in assessing the risk function for a data integration request. This risk function is evaluated in real-time when an access request for integration is received. The attributes of the privacy profile can be enhanced based on a clustering algorithm to estimate the probability of information linkage among device metadata or metadata of the corresponding data stream.

Fig. 3 provides a schematic privacy profile for a device. As shown in the figure, any access request to the device marked with a purpose as "data integration"‘ is evaluated on the basis of the requester's trust score, the purpose of data integration, analytic function specification and time period the access rights will be retained. The result of the evaluation is stored as an access token within the device, and any change in the request parameters or expiry of the access duration revokes the access permissions from the requester.

3.1.2. *Distributed access control*

Most of the smart home devices and solutions such as Nest are enabled with OAuth 2.0 protocol that is based on attribute based access control (ABAC) (IETF-OAuth-WG, 2006). When a user agrees to provide access to the requested permission level, Nest authenticates the request and an access token is granted to the requester. That access token can be used to access the Nest API and interact with the user's devices. Granting an access token establishes a trust relationship between a user and a Nest product. However, existing access control mechanisms are largely implemented in a top-down manner and do not consider the data sharing and re-use scenarios across distributed services and subsystems.

A simple solution here could be integration of distributed lightweight access control methods which support single-sign on (SSO) (Single-sign-on, 2013) among devices of different

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

manufacturers. A more sophisticated solution to cater data integration scenarios and data inferences need to be considered in the local access policies (Haddad et al., 2014). Haddad et al. propose a solution for data inferences based on functional dependencies in multiple queries by using a graph-based algorithm (Haddad et al., 2014). Similar methods need to be reinvented for the IoT domain for large-scale integration of both online and offline streams.

3.1.3. *Context of data collection versus data-use*

Informed Consent is an important element for data protection in Information and Communication Technology (ICT) systems as the consent of a data subject (e.g., the citizen) is often necessary for a third party to legitimately process personal data (Neisse et al., 2016). To provide informed consent regarding the use of personal data, the citizens must have a clear understanding on how their data will be used by the system. This may not be an easy task in the upcoming paradigm of Internet of Things (IoT) where personal data can be collected without the full awareness of the user (Neisse et al., 2016). For example, in autonomous cars, a policy-based toolkit may be implemented for a driver to broadcast messages related to road safety, high risk and emergency. However, in certain situations they might choose not to broadcast unimportant messages. Such a complex scenario can be remedied by making users understand data-use and sharing tasks that can be associated with their data and devices. The challenge here is that consent cannot be always accounted at a fine-grained level in complex IoT ecosystem.

4. Discussions

Privacy is a subjective term whose degree and definition often varies from one stakeholder to another, from one context to another and from one domain to another. For example, one homeowner may be willing to share his hourly readings from smart meter to obtain competitive pricing while another homeowner may consider this as breach of privacy. Therefore, in a highly dynamic and large-scale IoT ecosystem a privacy framework needs to be supported by the complementing legal framework (Ziegeldorf et al., 2014). This requires designing solutions that can replicate the privacy protection in smart settings that otherwise users allow in real-world situations. In such solutions, privacy-by-design principles can address some of the concerns by recommending practices of data minimization, purpose specification during the data collection. Moreover, other privacy threats during data integration and aggregation can be addressed by a minimal set of privacy protection mechanisms that are instantiated in different settings.

4.1. *Linkability: pivot of privacy concerns*

While linkability of the data is a critical privacy threat in itself, it also increases the severity of other privacy threats such as user profiling, location and tracking and identification (Aleisa and Renaud, 2016). Granularity of information linkage directly influences the risk of user profiling. Finer granularity at which information linkage can be done increases the risk of user profiling as the attribute correlation increases. Therefore, user profiling has become a more serious concern owing to fine granularity at which IoT collects data for big-data driven business models. With increasing number of different types of devices joining the IoT network, autonomous interactions between the number of privacy violations have increased. The study (Aleisa and Renaud, 2016) re-emphasizes the pervasive nature of personal data with the emergence of IoT and the need to address concerns of identification that arise from connecting an identifier with user identity. In addition, privacy threats such as location and tracking have become severe due to extensive use of GPS, internet traffic and smartphones. For example, the meta-data of devices often has the device location which when integrated with details added to users on different social networking platforms (even though anonymized) can reveal sensitive information about users.

4.2. *Technical caution to legal protection*

The privacy threat of "Information linkage" in IoT ecosystems is complex due to the presence of federated yet tightly coupled entities (devices, services) and subsystems. Horizontal data integration across these devices and subsystems increase the risk of privacy breaches at an unforeseen scale. In the previous sections, the study aims to empower various stakeholders in an IoT ecosystem to understand privacy threat of "information linkage" and its ramifications over their data and devices. In addition to stakeholder's understanding of data-use, this section discusses the legal solutions for protection of stakeholders' rights and ensuring legitimacy of their actions.

To identify the risk instigating elements during data integration that lead to information linkage we consider the existing work on "Unlinkability" (Danezis et al., 2014), legislations and best practices for "big data analytics" (Kemp, 2014; Nancy and Jay, 2016; Mantelero, 2016) and ongoing legal discussions on cybersecurity in IoT (Community, 2016; Weber and Studer, 2016). To the best of our knowledge, data privacy regulations based on "Privacy-by-design" protect rights of stakeholders during data collection phase and none of the existing work considers data processing challenges in IoT ecosystems from a legal perspective. Following is a discussion on how a legal framework can address privacy concerns during data integration across heterogeneous IoT subsystems (described in section 2.3).

- ***Characterization of IoT Ecosystem*** – IoT is an enabling technology for real-world settings, with deeper impact and automation than big data analytics. Therefore, it is important to characterize an IoT ecosystem with respect to its stakeholders – service providers and users. In addition, with respect to administrative domains, membership criteria, liabilities of different parties, data processing tasks and associated privacy risks need in-depth policies that codify "purpose of data collection", "intended versus actual use", "retention period", "informed consent" of data owners' per data-instance basis. In addition, if the architecture of an IoT ecosystem (smart home, neighbourhood, and city) is replicated in another administrative domain, the terminologies and local legislations need to be revisited and integrated.

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

Dynamicity challenge. Since IoT ecosystems are highly dynamic, consider that the devices and stakeholders leave and join an ecosystem on a continuous basis. At the policy level, data inferred or data that results from a data integration tasks needs to be covered with comprehensive contextual information to prevent any "in-context conflicts". For critical services, the withdrawal criteria need to be underlined for any participating entity. These are a minimal set of requirements that need to be expanded for specific applications.

- ***Multi-dimensional Regulation for IoT Ecosystem*** – As pointed by Weber and Studer (2016), a polycentric regulation is the way forward in IoT and it can be defined as "enterprise of subjecting human conduct to the governance of external controls, whether state or non-state, intended or unintended". Such a regulation would focus on multi-stakeholder, bottom-up approach towards self-regulation where stakeholders most familiar with an issue devise appropriate rules that can be codified (Weber and Studer, 2016). The model resonates the underlying principle of the given study, where stakeholders willingly participate in a connected heterogeneous smart neighbourhood. It requires addition of dimension of data-centric tasks (collection, integration) which may be performed autonomously by devices, and subsystems within an IoT ecosystem. Moreover, regulating biasness of local moderators, data storage location and purpose of data-use need to be elaborated. Regulations in IoT ecosystems, especially for integration related tasks need to emphasize on accountability and integrity of devices, services, subsystems and stakeholders.
- ***Data Brokers to IoT Data Brokers*** – Data brokers can play a key role in big data analytics (King and Forder, 2016). They are significant for sensitive data discovery and application of analytics; despite this, they are not governed by federal privacy laws in the United States. This is in contrast to the European Union where comprehensive data protection legislation may protect consumers' privacy through laws that apply to big Data processing (Nancy and Jay, 2016). Considering IoT technology is in nascent stage, emphasis on semantics of "personal data" in the IoT context, where devices collect data and share autonomously, is required before rights and functions of data brokers are defined.
- ***Layered Legal Framework for IoT*** – Another approach by Kemp (2014) comprises a layered architecture for big data where IP (intellectual property) rights in relation to data, contracts between parties for data sharing, data regulation and standards, and processes are described. The study emphasizes the enforcement of liabilities through "contract law" which enforces obligations. In addition, liability is based on balance of probabilities. This model intuitively can cater to some IoT domains where enforcement is ideal, such as "Informed consent" in healthcare over personal data becoming more generally applicable. It would be interesting to see how contract laws cater to integration scenarios where "intended purpose" needs to be enforced on the part of requesting parties.

A detailed legal study is essential to address challenges of data integration in IoT. Any such study should consider that most of the IoT devices are owned by a number of stakeholders and their participation in different subsystems often overlaps. Therefore, *accountability*, *liability* and *enforcement* are three critical pillars to support stakeholders' consent against information linkage.

5. Conclusions and future work

The given study presents an overview of privacy threats in emerging Internet of Things (IoT). It describes data management challenges for information and data integration in IoT ecosystems – presence of different metadata standards (interoperability), dynamicity of the data streams and availability of variety of smart devices (heterogeneity). The work presents an in-depth discussion and analysis on the privacy threat of *Information Linkage* during large-scale data integration in IoT ecosystems. It describes an example of a smart neighbourhood where data from various smart homes is integrated, and aggregated for a variety of services. Further, it details how some of these data processing tasks lead to unintended privacy breaches, such as identification of subjects and information linkage between their availability and resource use.

The main contribution of the proposed work is a distributed data integration algorithm for IoT ecosystems and resulting privacy threat of unintended *information linkage*. It then proposes technical and legal solutions to address this threat. The study argues that the scope of best practices of Privacy-by-Design is limited and is able to only partially address individual (and data) privacy. These principles often overlook the privacy breaches that occur during secondary level data processing and sharing tasks that do not directly involve raw data. The study emphasizes that this is an important area of research as most of the privacy threats in automated IoT ecosystems can surface at latter stages of data integration and aggregation where policies associated with raw data objects are of limited use. The given paper highlights the need for stakeholders' understanding of data processing (and integration), its ramifications on their privacy preferences and corresponding legal policies to protect their rights over data-use. As a future work, the authors aim to write a formal technical-legal framework to address privacy concerns during horizontal data integration both at device and data-stream level in heterogeneous IoT ecosystems.

Acknowledgement

The authors would like to thank Dr. Aastha Madaan, University of Southampton, United Kingdom for her valuable feedback on proposed algorithm and proof-reading the paper.

REFERENCES

Ahad MA, Biswas R. Comparing and analyzing the characteristics of Hadoop, Cassandra and Quantcast file systems for handling big data. Indian J Sci Technol 2017;10(8):1–6. http://www.indjst.org/index.php/indjst/article/view/105400.

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007

Al-Fuqaha A, Khreishah A, Guizani M, Rayes A, Mohammadi M. Toward better horizontal integration among IoT services. IEEE Commun Mag 2015;53(9):72–9. doi:10.1109/MCOM.2015.7263375.

Aleisa N, Renaud K. Privacy of the internet of things: a systematic literature review (extended discussion). CoRR 2016;1–10. http://arxiv.org/abs/1611.03340. Volume abs/1611.03340.

Beart P, Jaffrey T, Davies J. Hypercat 2.0 specification. [Online]; 2015. Available from: http://www.hypercat.io/standard.html. [Accessed 14 February 2017].

Cavoukian A. Strong privacy protection – now, and well into the future, A Report on the State of PbD to the 33rd International Conference on Data Protection and Privacy Commissioners. [Online]; 2011. Available from: https://www.ipc.on.ca/wp-content/uploads/Resources/PbDReport.pdf. [Accessed 28 November 2016].

Clifton C, Kantarcioğlu M, Doan A, Schadow G, Vaidya J, Elmagarmid A, et al. 2004. Privacy-preserving data integration and sharing. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (DMKD '04). ACM, New York, NY, USA, 19-26. http://dx.doi.org/10.1145/1008694.1008698.

Community. SCL (Society for Computers and Law), Regulation of the internet of things. [Online]; 2016. Available from: https://www.scl.org/articles/3685-regulation-of-the-internet-of-things. [Accessed 11 May 2017].

Danezis G, Domingo-Ferrer J, Hansen M, Hoepman J-H, Le Métayer D, Tirtea R, et al. Privacy and Data Protection by Design – from policy to engineering; 2014. Available from: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design. [Accessed 11 February 2017].

Elkheir M, Hayajneh M, Ali N. Data management for the internet of things: design primitives and solution. Sensors (Basel) 2013;13(11):15582–612. doi:10.3390/s131115582.

Haddad M, Stevovic J, Chiasera A, Velegrakis Y, Hacid MS. 2014. Access Control for Data Integration in Presence of Data Dependencies. In *DASFAA (2)* (pp. 203-217).

Hans J. Solving IoT integration: vendor landscape. [Online]; 2017. Available from: https://www.rtinsights.com/industrial-iot-companies-integration-platforms/. [Accessed 11 April 2017].

Kemp R. 2014. Legal aspects of managing big data. Comput Law Secur Rev 30(5), 482–91, https://doi.org/10.1016/j.clsr.2014.07.006.

King NJ, Forder J. Data analytics and consumer profiling: finding appropriate privacy principles for discovered data. Comput Law Secur Rev 2016;32(5):696–714.

Mainetti L, Mighali V, Patrono L. An IoT-based user-centric ecosystem for heterogeneous Smart Home environments. London, IEEE, pp. 704–709; 2015. http://dx.doi.org/10.1109/ICC.2015.7248404.

Mantelero A. Personal data for decisional purposes in the age of analytics: from an individual to a collective dimension of data protection. Comput Law Secur Rev 2016;32(2):238–55. https://doi.org/10.1016/j.clsr.2016.01.014.

Nancy JK, Jay F. Data analytics and consumer profiling: finding appropriate privacy principles for discovered data. Comput Law Secur Rev 2016;32(5):696–714. https://doi.org/10.1016/j.clsr.2016.05.002.

Neisse R, Baldini G, Steri G, Mahieu V. Informed consent in internet of things: the case study of cooperative intelligent transport systems. Thessaloniki, IEEE, pp. 1–5; 2016. http://dx.doi.org/10.1109/ICT.2016.7500480.

NEST. Nest developers. [Online]; 2017. Available from: https://developers.nest.com/. [Accessed 14 March 2017].

Perera C, Ranjan R, Wang L, Khan SU, Zomaya AY. Big data privacy in the internet of things era. IT Prof 2015;17(3):32–9. doi:10.1109/MITP.2015.34.

Shu K, Wang S, Tang J, Zafarani R, Liu H. User identity linkage across online social networks: a review. SIGKDD Explor Newsl 2017;18(2):5–17. doi:10.1145/3068777.3068781.

Single-sign-on. How to implement single-sign-on. [Online]; 2013. Available from: https://auth0.com/learn/how-to-implement-single-sign-on/. [Accessed 10 April 2017].

The Internet Engineering Task Force OAuth Working Group (IETF-OAuth-WG). OAuth authorization framework. [Online]; 2006. Available from: https://oauth.net/. [Accessed 10 April 2017].

Weber RH, Studer E. Cyber security in the internet of things: legal aspects. Comput Law Secur Rev 2016;32(5):715–28. http://dx.doi.org/10.1016/j.clsr.2016.07.002.

Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access 2014;2:1149–76. doi:10.1109/ACCESS.2014.2362522.

Ziegeldorf JH, Morchon OG, Wehrle K. Privacy in the internet of things: threats and challenges. Security Comm Netw 2014;7:2728–42. doi= 0.1002/sec.795.

Nishtha Madaan is a final year postgraduate student at Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard (Hamdard University), Delhi, India. Her area of specialization is Information Security and Cyber Forensics. Prior to this, she pursued an undergraduate degree in Information Technology (2011–2015) from Hamdard University. She has published her work in the area of cryptography and is keen to explore the research challenges in data management of emerging Internet of Things.

Mohd Abdul Ahad is working as an Assistant Professor in the Department of Computer Science & Engineering, School of Engineering Sciences and Technology, Jamia Hamdard (Hamdard University), Delhi, India. He has 9 years teaching and research experience in computer science especially Big data, operating systems, IoT, and distributed computing.

Sunil M. Sastry has a B.A.L, LLB from Bangalore Institute of Legal Studies, India. He pursued Masters in Law at University Law College, India, in Constitutional Law in 2000. Thereafter has been practicing at various courts in India, is a legal consultant for many international firms and a guest faculty at University Law College, and Bangalore Institute of Legal studies, India since 2006. He received an MPhil from National Law School of India University and PhD degree in Law from Department of Law, National Law School, Bengaluru, India in 2013. His interests include Constitution of India, banking law and data protection. Sunil has authorized various books on Hindu Law and SAFAESI 2002.

Please cite this article in press as: Nishtha Madaan, Mohd Abdul Ahad, Sunil M. Sastry, Data integration in IoT ecosystem: Information linkage as a privacy threat, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.06.007