

User-Centric Privacy Engineering for the Internet of Things

Mahmoud Barhamgi
Claude Bernard University
Lyon 1

Charith Perera
Cardiff University

Chirine Ghedira
Jean Moulin University
Lyon 3

Djamal Benslimane
Claude Bernard University
Lyon 1

User privacy concerns are widely regarded as a key obstacle to the success of modern smart cyber-physical systems. This article analyzes, through an example, some of the requirements that future data collection architectures of these systems should implement to provide effective privacy protection for users. Then, the article gives an example of how these requirements can be implemented in a smart-home scenario. An example architecture allows the user to balance the privacy risks with the potential

benefits and make a practical decision determining the extent of the sharing. On the basis of this example architecture, the authors identify a number of challenges that must be addressed by future data-processing systems in order to achieve effective privacy management for smart cyber-physical systems.

We increasingly find ourselves surrounded by smart cyber-physical systems that silently track our activities and collect sensitive information about us. Among the most prominent examples are smart energy grids, smart transportation networks, and smart homes and cities. However, although such systems promise to ease our lives, they raise major privacy concerns for their users. The data collected is often privacy sensitive, such as the location and habits of individuals and patients' vital signs. In fact, the collected data could be misused by the providers of such systems or even sold to interested third parties and exploited for various purposes.^{1,2}

Recent privacy and data protection laws³ have called for more involvement of users in protecting their data by enabling them to control what is collected about them, when, by whom, and for what purposes. However, existing solutions for privacy protection in cyber-physical systems⁴⁻⁷ fall short of that objective.

Most of these solutions are inspired by the old approach employed in databases to ensure the privacy of users.⁸ In that approach, users are required by the system to provide their data. Then, they are prompted to specify their privacy preferences (through a set of variables) as to who can access the data and for what purposes, and to accept a privacy policy specifying a set of rules that might refer to their preferences. Subsequently, the rules are applied to all queries received by the system before returning the result.

Unfortunately, such an approach does not provide effective protection of privacy. In fact, users do not always understand the privacy policy,⁹ which could be incomplete. They also might not necessarily be aware of the direct and indirect risks that might be associated with the disclosure of their data to a given entity in order to correctly specify their privacy preferences.

The successful involvement of users in protecting their data and ensuring their privacy requires two main conditions to be met. First, users should be empowered to understand the privacy risks associated with the disclosure of a piece of data to a given entity and to balance these risks with the potential benefits of the disclosure, to be able to make a meaningful privacy decision regarding whether to disclose, and to what extent. Privacy decisions are intrinsically difficult owing to their delayed and uncertain consequences, which are hard to compare with the immediate rewards of data disclosure.

Second, users should be provided with the necessary tools to implement their privacy decisions by controlling the disclosure level. For example, different data degradation strategies might be used to modify the accuracy of the to-be-disclosed data item to achieve the chosen tradeoff between the risks and the benefits.

In this article, we present a vision of how users of smart cyber-physical systems—e.g., the Internet of Things (IoT)—can be empowered to take a central and effective role in protecting their privacy. We materialize our vision by proposing a reference data-sharing architecture that allows users to make flexible and practical data-sharing decisions that reconcile their privacy requirements with their desire to be rewarded. We validated our vision by implementing the proposed architecture in a smart environment to monitor chronic patients at home.

Compared to similar research that sought to involve users in protecting their data,^{4–7} our solution has the following advantages. First, it allows users to assess the implicit privacy risks that are associated with the release of their data and to compare those risks in a meaningful way with the benefits, to choose the best data protection level. Second, privacy protection is ensured in a pragmatic way, allowing users to take a pragmatic stance between their interests and the risks implied. Third, our solution does not protect privacy in a rigid way. That is, as the context changes, the inferred privacy risks change as well, as does the protection ensured by the solution. This allows for more responsiveness to the surrounding IoT environment.

The remainder of the article is organized as follows. First, we analyze some of the key requirements for effective privacy protection in smart cyber-physical systems. Then, we present an example of an architecture implementing the identified requirements in a smart-home scenario. Finally, we point out future research directions.

A WALKING-THROUGH EXAMPLE

Alice is the owner of a smart home featuring different types of smart appliances, including a refrigerator, stove, and microwave. She is a CDK (chronic kidney disease) patient and has a home hemodialysis machine. The environment also includes smart meters for measuring energy consumption, and different types of sensors, including light, presence, and temperature sensors. These appliances, sensors, and meters generate important data volumes that could be exploited by different entities for different purposes.

For example, Alice's electricity provider would be interested in exploiting the generated data (e.g., energy consumption) to improve energy distribution across the city and avoid service cut-offs. The provider can also use the data to provide Alice with personalized recommendations for reducing her energy consumption and bills.

Related edge services include businesses providing services to energy consumers based on energy consumption data. Examples of services include real-time energy usage monitoring to optimize consumption (e.g., by proposing actions to users such as turning on or off a certain device) and raising energy awareness by allowing consumers to monitor their carbon emissions and compare them to those of friends on social media.

Law enforcement agencies might use the collected data for purposes such as performing real-time (or near real-time) surveillance on suspects by determining whether they are present and their current activities in the home. Police investigators might also screen the energy consumption records of the utility to identify houses in which some illegal activities might be taking place—e.g., potential drug production sites across the city.

Marketers might be interested in determining Alice’s lifestyle to send her targeted advertisements. For example, they might be interested in knowing the appliances she might or might not own or her eating patterns, to send her special offers, etc.

Privacy Concerns

The cited possible data uses might raise several privacy concerns for Alice. First, Alice might be interested in reaping the benefits offered by a utility or some edge services (for instance, usage optimization of her appliances), either by explicitly providing detailed information about the appliances’ usage or by providing energy consumption data that is granular enough to infer the appliances’ usage (through their consumption signatures¹⁰). However, she might not want to disclose that she is a CKD patient (through the use of her hemodialysis machine). The disclosure of such sensitive information, inadvertently or intentionally by one of the entities processing her data, could irreversibly affect her professional and social life.

Second, Alice might not wish to disclose her habits and lifestyle—e.g., her presence–absence, walking, sleeping, eating, or TV-watching patterns. Such information, if misused, could harm her in different ways. For example, she might become a target for housebreakers or get penalized by her employers or her social entourage. Alice might not wish to be under permanent surveillance. This would make her feel uncomfortable and impact her natural behavior.

In this work, we uphold the privacy definition given in “Internet Privacy Concerns Confirm the Case for Intervention,”¹¹ where privacy is described as consisting of four dimensions:

- privacy of personal information,
- privacy of the person (i.e., the integrity of his or her body),
- privacy of personal behavior, and
- privacy of personal communications.

Therefore, we consider as sensitive all the data items that might be used to compromise the privacy of a person in at least one of these dimensions.

Practical Privacy Decisions

Alice is also a pragmatic person and might be willing to accept potential privacy risks to reap the benefits of sharing some of her data. For example, she might agree to release fine-grained energy consumption data (consumption readings with a high sampling frequency) to her electricity provider to help it better optimize its energy distribution, provided she receives some financial benefit (e.g., bill reductions or bonuses). She might find it acceptable for an edge service to gather the data necessary to compute her daily carbon emission and compare the information with that of friends on a social network (to gain some social recognition), but not to infer the list of her appliances. Similarly, she might agree to provide law enforcement services the necessary data to check for illegal activities, but not allow real-time surveillance.

In all of these examples, Alice makes her practical decision after evaluating how trustworthy the data consumer is, for what purposes (that potentially relate to her privacy requirements and preferences) the released data can be exploited, and what the benefits of the data-sharing decision are.

Identified Requirements

On the basis of our simple example, we identify below some of the key requirements for ensuring effective privacy protection for cyber-physical systems.

User-Centric Privacy Protection

As required by data protection regulations, users (data owners) should be enabled to play a central role in protecting their data. This implies that before sharing a data item with a given data consumer, the user should be

- enabled to understand the privacy risks that pertain to his or her subjective vision of privacy, and
- provided with the necessary mechanisms to control the extent of the sharing.

For example, the data-sharing architecture should warn Alice, before she releases her fine-grained energy consumption readings to her electricity provider or an edge service, that these bodies might determine her appliances (on the basis of their signatures) and consequently know that she is a CDK patient. It should also allow Alice to reduce the frequency of the released energy consumption readings to prevent such a privacy breach.

Pragmatic Privacy Decision Making

Studies have showed that users, in practice, tend to take a pragmatic stance on sharing their private data.¹² That is, they would accept the release of some of their private data in return for some incentives. This has motivated an important drive in the database research field to monetize private data, on the basis of potential privacy risks, and compensate users directly.¹³

On the basis of that observation, privacy protection in cyber-physical systems should not be ensured in a rigid fashion by deciding whether a data item should be shared with a given entity or not. Rather, users (e.g., Alice) should be empowered to assess the privacy risks, balance them against the benefits offered by data consumers, and potentially negotiate with the data consumers before making their sharing decisions. Data owners and consumers will play roles very similar to the roles of sellers and buyers in a free market, in which the sellers and buyers might bargain with each other to reach a suitable deal.

Adaptive Privacy Protection to Cope with Context Evolution

The data-sharing decision might depend on the user's context. For example, Alice might accept, in the general case, to release fine-grained energy consumption data to her electricity provider that would allow the latter to infer her appliances' usage, but not when she uses her hemodialysis machine.

The data-sharing architecture should detect the context changes that would lead to privacy breaches and take the necessary actions to avoid them. For instance, when Alice turns on her hemodialysis machine, the architecture could warn her about the potential leakage of her health condition to her energy provider. It could then suggest that she decrease the sampling frequency of the released energy consumption readings.

A REFERENCE DATA-SHARING ARCHITECTURE FOR USER-CENTRIC PRIVACY PRESERVATION FOR THE IOT

In this section, we describe a data-sharing architecture for a smart home that satisfies the requirements discussed above and gives control over data sharing back to users.

Architecture Overview

We use the term “data owners” to designate the users of cyber-physical systems that generate data by interacting with the systems (e.g., occupants of smart homes and monitored patients). “Data consumers” designates the stakeholders that are interested in collecting and exploiting the data generated, such as electricity companies in smart grids, healthcare providers in intelligent healthcare networks, and third parties that would provide data owners with smart services.

In our architecture (see Figure 1), raw data generated by connected things (IoT objects) is stored in a *personal data store* before being released to data consumers. All dataflows between data owners and consumers pass through the proposed reference architecture. When a data consumer queries a data item from a data owner (either directly or by getting the data owner to use a service or application provided by the data consumer), our architecture processes the received query as follows.

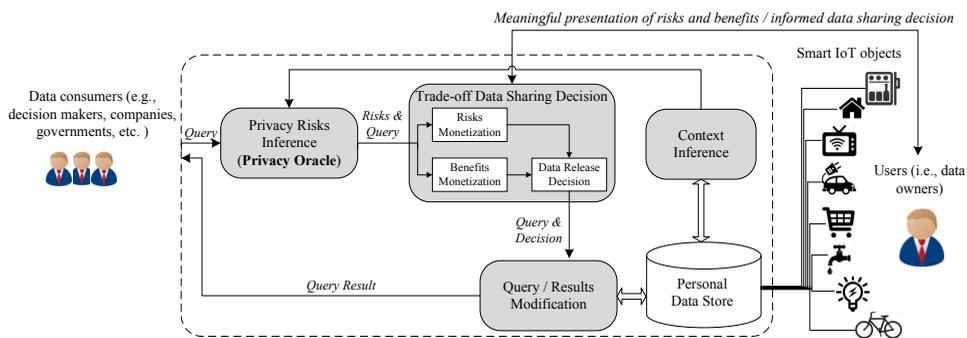


Figure 1. A reference architecture for allowing the users of Internet of Things (IoT) smart systems to control their data.

The architecture assesses, through the *privacy risks inference* component (also called the *privacy oracle*), the risks associated with releasing the requested data to the data consumer. Because privacy is a subjective notion (different people might be concerned with different privacy risks), the privacy oracle takes into account several factors, including the profile of the data owner, his or her context, and his or her trust in the data consumer. Context is a key input to the privacy oracle; the architecture monitors it in a continuous way (by observing the data sent by IoT objects) and exploits it when identifying the risks.

Then, the architecture monetizes (quantifies) the identified privacy risks and the potential benefits using a numerical model, through the *tradeoff data sharing* component, and helps the data owner make a pragmatic decision balancing the two. The decision denotes the data items that can be shared with the consumer, along with their accuracy (precision). Data owners and consumers could also negotiate before the owners make a pragmatic data-sharing decision.

Finally, the architecture modifies, through the *query/result modification* component, the query before applying it to the personal data store, to discard the data items to which the data consumer is not entitled. The architecture also modifies the query’s result to change its accuracy before its release.

Privacy risks relate closely to what can be inferred from the collected data. For example, granular readings of smart electricity meters can be analyzed to infer information about the occupants, such as their presence or absence and the possession of specific devices. The disclosure of such information could lead to privacy risks such as the user being subject to discrimination, surveillance, and burglaries. Our architecture identifies the relevant risks, through the privacy-risks-inference component, by rendering explicit the implicit relationship between raw collected data and risks.

Context changes could lead to privacy breaches in previous decisions. Therefore, context is monitored, and risks are analyzed in a permanent way (e.g., old data-sharing decisions might be recomputed when the context changes).

Our architecture would use any data protection scheme (whose efficacy is proved) to anonymize the data once a data-sharing decision is made. For example, differential privacy can be used to alter the precision of released smart-meter readings, or anonymization can be used to anonymize a location.

Privacy Risks–Benefits Tradeoff Model

We present here our model to trade off the privacy risks with the data-sharing benefits. Our model builds on previous work on tradeoff decision models, such as in “Adaptive Sharing for Online Social Networks: A Trade-Off between Privacy Risk and Social Benefit.”¹⁴ The model is shown in Figure 2.

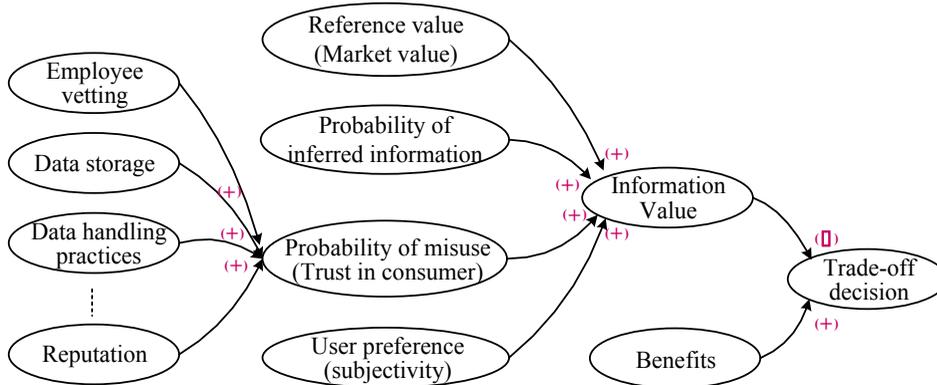


Figure 2. A privacy risks–benefits tradeoff model.

The tradeoff decision is based on two factors:

- the privacy risks associated with answering query q of data consumer d , and
- the benefits generated by the query answering.

The privacy risks factor has a negative impact on the tradeoff decision, whereas the benefits factor has a positive one. The privacy risks are measured on the basis of three factors:

- the sensitivity of the data items requested in q ,
- the trust in the recipient d (the data consumer), and
- the information leakage caused by answering q .

In subsequent subsections, we define all of these factors and explain how they can be computed.

Here we define the tradeoff privacy decision. The decision to answer query q of data consumer d is computed as

$$Decision(q, d) = \begin{cases} \mathbf{Answer}, & \text{if } U_d(q) > 0 \\ \mathbf{Deny}, & \text{otherwise} \end{cases},$$

where $U_d(q)$ is the utility function of answering q . We compute $U_d(q)$ as

$$U_d = (1 - w) * b_d(q) + w * r_d(q),$$

where $b_d(q)$ and $r_d(q)$ are functions quantifying the benefits and the risks that are generated or caused by answering q , respectively. The parameter w can be tuned by the user to bias the benefits–risks tradeoff decision (e.g., if $w = 0.5$, the benefits should be at least equal to the risks). In the following, we detail how $r_d(q)$ and $b_d(q)$ can be computed.

Measuring the Privacy Risks

The privacy risks can be measured on the basis of three factors:

- the sensitivity of the queried data items,
- the trust in the data recipient, and
- the information leakage caused by the release of the data items.

The sensitivity of a data item can be determined on the basis of its direct or indirect relation to a piece of privacy-sensitive information. For example, the electricity consumption readings are not privacy sensitive in themselves. However, because they can be analyzed by some data-mining algorithms¹⁰ to determine the appliances' usage and consequently infer behavior patterns (e.g., waking–sleeping patterns and meal times), they are considered sensitive.

The NIST guidelines for smart-grid cybersecurity have identified the different privacy-sensitive information pieces that could be relevant to a wide range of users in smart-home scenarios and that correspond to our vision of privacy. These include users' personal information, their presence or absence, real-time surveillance, users' habits, and the use of a specific device. We call these information pieces the *privacy parameters* and exploit them to compute the sensitivity of a data item i (e.g., energy consumption data) as follows.

The user is prompted to assign an importance weight to each of the privacy parameters $f_j (1 \leq n)$. Then, whenever a data item i is requested by a query, its sensitivity is computed by summing up the weights of all the privacy parameters that relate to it (that can be inferred from it), by

$$\text{Sensitivity}(i) = \sum_{j=1}^n \text{weight}(f_j) * \text{context}(f_j),$$

where the function $\text{context}(f_j) \in \{0, 1\}$ takes the value 1 or 0 depending on whether f_j is relevant or not in the current context.

A user's trust in a data consumer might change depending on the consumer's profile and its history of interaction with users. For example, a law enforcement agency might be trusted more than a publicity company. An electricity provider with which a user has had a long and satisfactory history of interaction might be trusted more than an unknown edge service. In our model, we can rely on any of the trust models¹⁵ that compute the trust in a consumer by aggregating and averaging quantitative feedback ratings of users (e.g., smart-home owners).

Information leakage refers to the knowledge leaked to d about the different privacy-sensitive parameters $f_j (1 \leq n)$ when a data item i is disclosed to d . The information leakage relative to a particular privacy parameter f_j , denoted by L_{f_j} , can be measured by capturing the uncertainty of d about f_j and is dependent on the accuracy level of the released i . L_{f_j} can be measured either analytically or experimentally.

For example, experimental studies showed that releasing energy consumption data with good accuracy (a sampling frequency of 15 seconds) led to determining eating habits with a 59% degree of confidence and appliance use with a 72% degree of confidence.¹⁰ But, when the sampling frequency dropped to 30 minutes, the degree of confidence dropped to nearly 0 for both cases.

Measuring the Benefits

Different types of benefits require different kinds of measurements. For example, usage optimization advice that would lead to lower energy consumption and to bonuses and bill reductions can be quantified by the introduced financial gain. The benefits of sharing fine-grained data with law enforcement agencies to help secure the living environment can be quantified by the user's feeling of self-satisfaction and civic duty. The benefits of sharing carbon emission data on a social network can be quantified by the social recognition the user receives.

REAL-LIFE CASE-STUDY: MONITORING ELDERLY AND CHRONIC PATIENTS

We implemented our reference architecture in a smart environment for monitoring chronic patients. The environment involves wearable sensors for monitoring several vital signs such as heart activity and blood pressure, as well as smart objects and sensors installed in fixed positions in the monitored environment. The study involved 20 real patients of ages 20 to 67.

Figure 3 shows the implementation architecture. Collected raw data are stored in a *personal data vault* (PDV) that provides a simple implementation of the different components of our architecture. Healthcare providers provide patients, through a dedicated mobile app, with personalized healthcare services by consuming collected data. Dataflows between the users (patients) and data consumers (healthcare providers) go through the PDV.

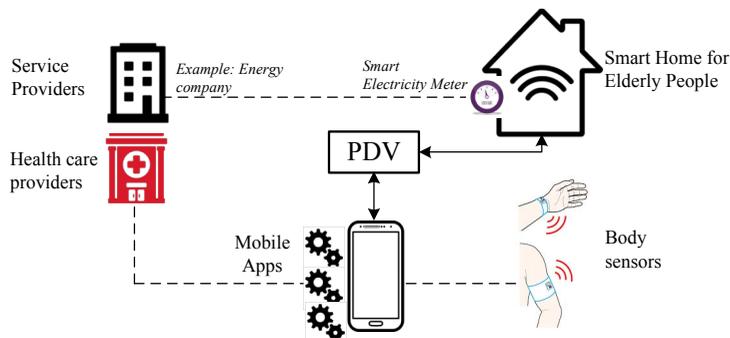


Figure 3. Implementation of the proposed architecture in an IoT environment for monitoring chronic patients.

We implemented the privacy-risk-inference component as a knowledge base. The collected data, context elements, and associated risks are modeled by a domain ontology. The knowledge base models the implicit relationship between the raw data and risks by a set of inference rules expressed in the Semantic Web Rule Language (SWRL).

Figure 4 presents simplified examples of inference rules. Rule 1 states that the use of a device can be inferred from the readings of an energy smart meter (EMR). The terms “Person,” “EMR,” and “Device” are ontological concepts, whereas “hasEMR,” “isShared,” “useDevice,” and “isInferable” are properties. Rule 2 states that the use of a medical device reveals the health conditions for which the device is used. Rule 1 and Rule 2 can be combined to infer that the health conditions could be inferred from the readings of an EMR.

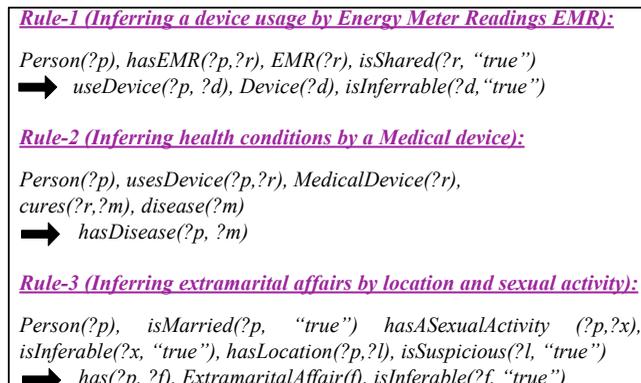


Figure 4. A sample of inference rules for inferring privacy risks.

Rule 3 simply states that the location and heart rate metadata could be combined to infer whether the data owner is having an extramarital affair. (The heart rate can be used to infer whether the data owner is engaged in sexual activity. The location can be exploited to infer whether the data owner is in a suspicious location—e.g., not at home.) The rule uses contextual information such as whether the data owner is married.

Figure 5, window 1, shows the user interface of the PDV. Upon the reception of a new data request, the PDV takes into account the user’s context and shared data (window 2) to provide a description of the associated privacy risks (window 3) and a set of recommended actions (window 4).

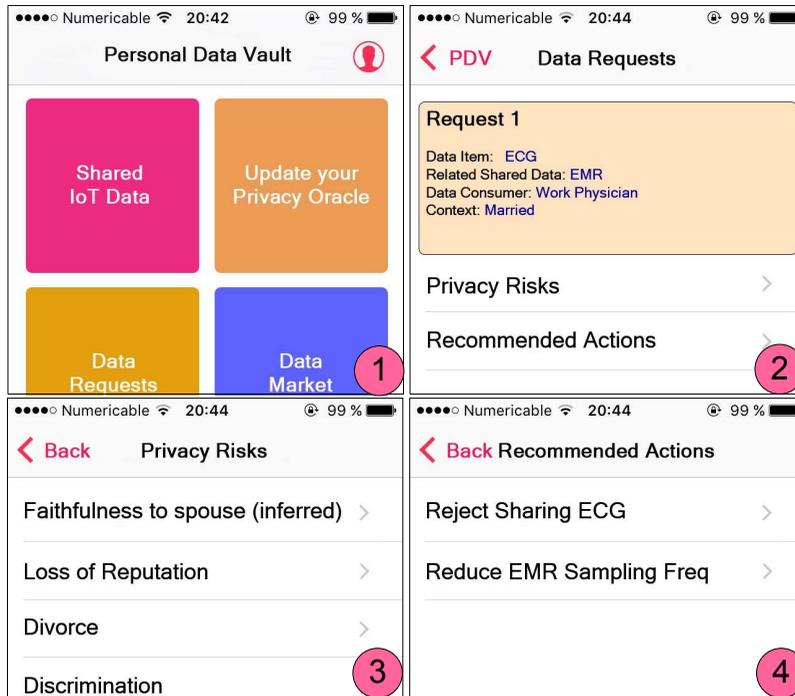


Figure 5. Interaction with users to control the release of data.

CONCLUSION

We conclude the article by identifying three of the key challenges that should be addressed to achieve effective privacy protection in cyber-physical systems.

The first challenge is meaningful data degradation strategies. Smart environments produce a rich set of data elements that are different in nature and require different forms of degradation (to implement chosen tradeoff decisions). For example, data elements such as the address or the age of a person can be degraded by applying well-known data anonymization techniques. Electricity consumption readings can be degraded by reducing the sampling frequency. Research is needed to define, for each type of data element, suitable degradation strategies and data degradation levels that map directly to the elements’ different possible uses (risks) and to data leakage levels.

The second challenge is rich pricing models for privacy-sensitive data. Data consumers might offer different forms of benefits, including financial, social, or societal benefits. Research is needed to devise rich and flexible pricing models (for privacy-sensitive data) that would fit various forms of benefits. Efforts are also needed to help users assess the sensitivity of their private data. Crowdsourcing techniques can be explored for that purpose.

The third challenge is context modeling and monitoring for triggering the adaptation of privacy decisions. The sensitivity of a data piece might depend on the context. For example, energy consumption data might become sensitive when Alice turns on her hemodialysis machine. Models

and techniques are needed to represent and monitor the context elements that relate to users' privacy and detect context changes that require adapting privacy decisions.

Complex event processing (CEP) techniques can be explored for that purpose. In fact, the interaction of users with their surrounding environment generates various events that can be monitored. The contexts that require adaptation are represented by a set of event combinations that can be tracked by the CEP system.

REFERENCES

1. I.D. Addo et al., "A Reference Architecture for Improving Security and Privacy in Internet of Things Applications," *Proc. IEEE 3rd International Conference on Mobile Services (MS 14)*, 2014, pp. 108–115.
2. S. Sicari et al., "Security, privacy and trust in Internet of Things: The road ahead," *Computer Networks Journal*, vol. 76, 2015, pp. 146–164.
3. "General Data Protection Regulation (GDPR)," European Union, 2016.
4. W. Enck et al., "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," *Communications of the ACM*, vol. 57, no. 3, 2014, pp. 99–106.
5. C. Castelluccia et al., "Enhancing Transparency and Consent in the IoT," *IEEE European Symposium on Security and Privacy Workshops*, 2018, pp. 116–119.
6. A. Ukil, S. Bandyopadhyay, and A. Pal, "IoT-Privacy: To be private or not to be private," *IEEE INFOCOM Workshops*, 2014, pp. 123–124.
7. S. Sicari et al., "A security-and quality-aware system architecture for Internet of Things," *Information Systems Frontiers*, vol. 18, no. 4, 2016, pp. 665–667.
8. K. LeFevre et al., "Limiting Disclosure in Hippocratic Databases," *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, 2004, pp. 108–119.
9. S.W. Smith, "Humans in the Loop: Human-Computer Interaction and Security," *IEEE Security & Privacy*, vol. 1, no. 3, 2003, pp. 75–79.
10. M.A. Lisovich, D.K. Mulligan, and S.B. Wicker, "Inferring Personal Information from Demand-Response Systems," *IEEE Security & Privacy*, vol. 8, no. 1, 2010, pp. 11–20.
11. R. Clarke, "Internet Privacy Concerns Confirm the Case for Intervention," *Communications of the ACM*, vol. 42, no. 2, 1999, pp. 60–67.
12. *Personal data : The emergence of a new asset class an initiative of the World Economic Forum*, World Economic Forum, 2011; <https://www.weforum.org/reports/personal-data-emergence-new-asset-class>.
13. C. Li et al., "A Theory of Pricing Private Data," *ACM Transactions on Database Systems*, vol. 39, no. 4, 2017, pp. 1–28.
14. M. Yang et al., "Adaptive Sharing for Online Social Networks: A Trade-off Between Privacy Risk and Social Benefit," *Proc. IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 14)*, 2014, pp. 45–52.
15. Z. Malik and A. Bouguettaya, "RATEWeb: Reputation Assessment for Trust Establishment among Web services," *VLDB Journal*, vol. 18, no. 4, 2009, pp. 885–991.

ABOUT THE AUTHORS

Mahmoud Barhamgi is an associate professor of computer science at Claude Bernard University Lyon 1. His research focuses on privacy preservation in service-oriented architecture, web, and cloud environments. Barhamgi received a PhD in information and communication technology from Claude Bernard University Lyon 1. Contact him at mahmoud.barhamgi@univ-lyon1.fr.

Charith Perera is a lecturer at Cardiff University's School of Computer Science and Informatics. His research focuses on security and privacy in the Internet of Things. Perera received a PhD in computer science from the Australian National University. Contact him at charith.perera@ieee.org.

Chirine Ghedira is a professor at the School of Management at Jean Moulin University Lyon 3. Her research focuses on service-oriented architecture. Ghedira received a PhD in computer science from the National Institute of Applied Sciences of Lyon. Contact her at chirine.ghedira-guegan@univ-lyon3.fr.

Djamal Benslimane is a professor of computer science at Claude Bernard University Lyon 1. His research focuses on service-oriented architecture, databases, and ontologies. Benslimane received a PhD in computer sciences from Clermont-Ferrand University. Contact him at djamal.benslimane@univ-lyon1.fr.