PETRIoT - A Privacy Enhancing Technology Recommendation Framework for IoT Computing

Fatema Rashid[1], Ali Miri[1] and Atefeh Mashatan[2]
[1]*Department of Computer Science, Toronto Metropolitan University, Toronto, Canada*
[2]*Ted Rogers School of Information Technology Management, Toronto Metropolitan University, Toronto, Canada*

Keywords: Privacy Preserving Data Sharing, IoT Devices, Privacy Enhancing Techniques, Differential Privacy, Federated Learning, Data De-Identification, Homomorphic Encryption, Multiparty Computation, Synthetic Data Generation.

Abstract: Data sharing has become a critical component in any computing domain for organizations of different scales. Governments and organizations often must share their sensitive data with third parties in order to analyze, mine or fine tune data for critical operations. However, this can lead to privacy concerns when dealing with sensitive data. Privacy Enhancing Techniques (PETs) allow data sharing between two or more parties, while protecting the privacy of the data. There are different types of PETs that offer different advantages and disadvantages for specific application domains. Therefore, it is imperative that a careful selection and matching of application domain and PET is exercised. Selection of PETs becomes more critical when it comes to the data generated from Internet of Things (IoT) devices as such devices are becoming more pervasively present in our lives and thus, capturing more sensitive information. In this paper, we design a novel framework in accordance with National Institute of Standards and Technology (NIST) recommendations to select an appropriate PET in different application settings with respect to privacy, computational cost and usability. We design a recommendation system based on a strategy which requires input from data owners and end users. On the basis of the responses selected, the recommendation is made for an appropriate PET to be deployed in a given IoT application.

1 INTRODUCTION

The Internet of Things (IoT) is comprised of different physical objects connected through networks over the internet. For example, different types of sensors gather information and share it across different systems in order to store and analyze the data. It is projected that by 2025 our global data volume will reach 175 zetabytes (Farall, 2021). One of the largest sources of collected data is the Internet of Things (IoT). The amount of data generated by IoT devices is expected to reach 73.1 ZB (zettabytes) by 2025 (Walker et al., 2022). This is equals 42% of the 2019 output, when 17.3 ZB of data was produced (Walker et al., 2022).

Big data analytic tools can be used to process large IoT-related data with speed, efficiency and accuracy. They can provide unique insights when used with IoT devices namely, descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics (Farall, 2021). Analyses can be done to perform a variety of tasks from anomaly detection to locating devices in order to identify how the devices are being used by the end users.

Data sharing is the process of making data available to other individuals or groups within or outside of an organization. Data sharing also enables the distribution of data across different domains. Organizations that share data typically can make better informed business decisions. As mentioned earlier, huge volumes of are being produced everyday. Processing this scale of data - a large portion of which is produced by IoT devices - often necessitates that access to data be be granted to different users within an organization, and often to third-party service providers. Protection of the privacy of sensitive data is often considered a major challenge to this data sharing. Privacy Enhancing Technologies (PETs) offer a possible solution to this challenge. It is important to note that the use of PETs does not guarantee complete privacy protection of the data. It may still be possible for attackers to access the data, depend-

Rashid, F., Miri, A. and Mashatan, A.
PETRIoT - A Privacy Enhancing Technology Recommendation Framework for IoT Computing.
DOI: 10.5220/0012450900003648

In *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, pages 838-844
ISBN: 978-989-758-683-5; ISSN: 2184-4356

ing on the method and the strength of the technology used. The use and the choice of PETs should be considered with a more fulsome decision-making process. The *use*, *level of access*, and *security* are some of the important considerations to take into account. Furthermore, the trade-offs between cost, privacy and usability can have direct implications on the choice of PETs for specific application scenarios, such as IoT systems. Hence, the selection and use of PET must carefully consider costs, privacy and security requirements and computational capacity of the desired application. Typically, the higher the privacy required, the higher the computational cost and lower the usability, whereas a lower privacy requirement can result in a higher usability and a lower computational cost. In this paper, we design a new framework for the selection of an appropriate PET for specific/contexual scenarios commonly arising in IoT use cases through a series of queries based on the National Institute of Standards and Technology (NIST) recommendations (Fagan et al., 2020).

IoT devices are used by different types of end users and organizations. In many cases, IoT devices have limited or no ability to be updated and patched against fast changing threat landscape. Therefore, they represent a major vulnerability to many systems that deploy them. In 2020, NIST released NISTIR 8259, *Foundational Cyber security Activities for IoT Device Manufacturers* (Fagan et al., 2020) so that the manufactured IoT devices provide "necessary cyber security functionality" and provide customers "with the cyber security-related information they need". In this paper, we will devise a set of questions to design our framework and help recommend the most appropriate PET.

The NIST guideline presents six activities and a related questionnaire that are focused on two phases: before a device is sent out for sale in the market (pre-sale) and after devise sale (post-sale). Two of these activities are focused on the post-sale phase and four are focused on the pre-sale phase. Using questions from the pre-sale phase, we selected only the questions which suit our PET recommendation IoT computing framework.

Our framework is designed to determine the privacy, network, cost, scalability and authorization requirements of the users (companies or individuals). In our framework, we have prioritized data sensitivity as the most critical element before considering other criteria, such as efficiency and scalability. With the help of the selected queries and user responses, our recommendation system will suggest an optimized PET as per the requirements of the specific application domain or scenario.

The rest of the paper is organized as follows: Section 2 covers the background and the related work highlighting the six selected privacy enhancing techniques used in this work. Section 3 discusses the PETs in the context of IoT computing and how their characteristics can be used to fulfill the requirements of a specific IoT application. Section 4 introduces our recommendation framework, followed by two case studies in Section 5. Section 6 lists the conclusions and possible future work.

2 BACKGROUND

A recent technical report by UK Royal Society investigated the use of PETs in private data sharing(The Royal Society, 2023). One of its key findings was major challenges applications of PETs face in practice. These challenges included a general "lack of knowledge and expertise when it comes to the selection and application of PETs in data sharing"(The Royal Society, 2023). A lack of PET-specific standard of use in the existing market highlights is the motivation for our work that focuses on IoT devices. Leveraging NIST security guidelines for IoT device manufacturers, our framework will attempt to fill in the existing gaps between the end users and manufacturers when it comes to privacy requirements. Based on the input from the users, and considering various factors including privacy, usability and cost, the framework can recommend the most appropriate PET.

Below, we briefly discuss six different PETs used in this paper and provide references for further reading.

Data de-identification is refers to the process of removing all personal identifiers from a dataset, so that modified dataset can be shared with other parties (Binjubeir et al., 2020). One of the main challenges of data de-identification techniques is balancing the need to protect the personal information of the individuals, while providing the opportunity to analyze the characteristics of the raw data.

Differential Privacy (DP) in another popular privacy-enhancing technique (Dwork et al., 2016). DP works by the formation of a layer between queries from users and database itself. It provides provable guarantee that even if the attacker has knowledge of all but one record in the dataset, analyzing the result of any one query would not enable the attacker to identify the presence or absence of an individual user in the dataset.

Synthetic Data is generated from the original data and it is ensured that it exhibits the same underlying data distribution, characteristics and trends as shown

in the original data. The properties which we want be present in the generated synthetic data is that not only the statistical property and the original structure are retained when compared to the original data, but also that no private information is disclosed through the synthetic data (Creswell et al., 2018).

Homomorphic Encryption (HE) allows for certain computations on encrypted data without the need to decrypt it first, which then can be used to protect the privacy of the data (Sun et al., 2018). Related is the **Multiparty Computation (MPC)** techniques that allow different parties to carry out a computation using their private data without revealing their private data to each other (Lindell, 2020). In MPC, a given number of participants each possess a piece of private data. In the next step, the participating parties calculate the value of a public function on the original private data, while keeping their own data share private from other parties (Knott et al., 2021).

Federated Learning (FL) represents the techniques that enables learning from datasets on different machines without the need to share the training data between the machines (Khan et al., 2021), and hence providing a layer of privacy for the data.

To determine which PET work the best for a given environment, privacy requirements have to be balanced with the underlying model of that environment. For example, some PETs are very efficient when used in collaborative learning environments, while others may be better suited with centralized architectures. Some settings may required access to the actual data, where as others may only require access to aggregated or modified data.

3 PETs IN THE CONTEXT OF IoT COMPUTING

In this paper, we formulate a domain-specific PET selection strategy in the IoT device computing domain. We compare the pros and cons of each PET with regards to privacy, efficiency, and usability in the context of IoT devices. There are several factors to consider before opting for a PET, including type of data, domain of data, privacy requirements and available computational resources. In the following, we will discuss their suitability of PETs listed in the previous section in the context of IoT domain, and given the consideration above.

- DP and its many extensions have been used in practice in many application domains. However, there is a careful need to balance privacy with utility, fairness and robustness considerations. For example, it has been shown that when DP is applied to data from a large diverse set of nodes, DP will face serious challenges when analysis is focused on statistical properties of data from small subset of nodes with specific attributes (Dwork and Roth, 2014; Jordan et al., 2022). There have been reported possibility of information leakage, when information gained through application of DP on a given dataset is combined with other information from elsewhere (Alvim et al., 2015), as well as challenges of handling time-series data (Rastogi and Nath, 2010).
- HE and MPC can both provide some of the strongest privacy protection, and accuracy. However, they may not be applicable to all scenarios. They are typically complex, and their required operations required can impose large overheads that may limit their use with resource constrained IoT devices (Jordan et al., 2022).
- FL is an important tool for decentralized learning that can not only work with datasets with heterogenous sources of unstructured complex data, but also by design offers an inherent added layer of privacy by allowing each nodes in decentralized supervised learning work on its own unreleased data, and share updates to the model. Although, it does not provide a provable privacy guarantees, typical implementations have reasonable computational and communication overheads for most IoT devices. Recently, a number of possible weaknesses in FL have been identified that can impact data privacy (Lyu et al., 2022). So, it has been suggested that combining FL with DP, MPC, or HE can provide a stronger privacy-protecting solutions (Kairouz and et al, 2021).
- SD generated data can offer an attractive solution for protecting privacy, when the access to the real data is not fully required (El Emam et al., 2020). A common and effective method to generate synthetic data is to use the Generative Adversarial Networks (GANs). Although overhead cost associated with this data generation can be considered manageable, the result may not be suitable for applications with high accuracy requirements.
- Data de-identification has been one of the most common method to handle sensitive data publishing. It is inexpensive in terms of required computational overhead, and it can provide accurate results for non-anonymized attributes of data. However, many studies have shown the ease of re-identification of data by utilizing big data analytic tools and leveraging other related source of data (Scaiano et al., 2016). Therefore, data de-

identification cannot be considered a viable option for protecting IoT systems.

As illustrated in Figure 1 the privacy provision afforded by a given PET is inversely proportional to its computational overhead cost.

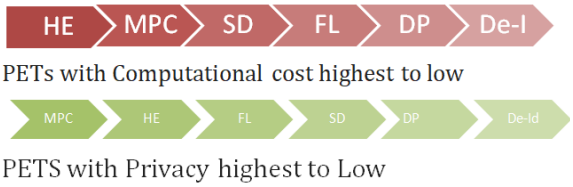

Figure 1: Ranking of PETs with respect to Privacy and Computational Cost.

4 PET RECOMMENDATION FOR PRIVACY PRESERVING IOT COMPUTING

In the following, we leverage IoT best security practices proposed in the NIST's document *Foundational Cyber security Activities for IoT Device Manufacturers* (Fagan et al., 2020) in order to formulate our framework. The purpose of NIST's document was to provide recommendations to the manufacturers on how to improve the security of their IoT devices. Implementing these recommendations would allow IoT devices offer device cyber security capabilities and manage their cyber security risks.

The NIST's document outlines six recommended foundational cyber security conditions that manufacturers should consider in order to meet and enhance the security of their IoT devices. Our framework selection is based on users' answers to questions we have selected with regard to two of the activities namely: *Activity 1: Research End User Cyber Security Needs and Goals* and *Activity 2: Identify Expected End Users and Define Expected Use Cases* as discussed in the NIST's document. The answers can be used to determine the appropriate PET for the IoT device by analyzing the privacy requirements of the device, end users and the data generated by the devices. This preliminary information from the end users can help the users/analysts to determine an appropriate PET based on the application domain. The selected five questions from two different categories with the possible response choices from the users are highlighted in Table 1. The response choices can be modified or expanded as needed.

The user will be asked to select the answers from the provided options only. Each response has defined context and the decision of the recommendation of the PET depends on the selected responses. Once the responses have been recorded, the system will recommend the most optimal PET to the user for the given IoT computing domain.

Table 1: Framework for PET Recommendation.

Question Domain	Question	Options
Identify expected end users	What is the scale of the size of the expected end users of this IoT device?	Personal-SME-Large
Identify expected data types	What type of data, the device is expected to generate?	Tabular-Time series-Images
Identify device cyber security requirements	What types of access the IoT device will be exposed to?	Authorized access-Unauthorized access
Identify data cyber security requirements	What is the nature of the severity of the privacy of IoT device's data?	Extremely sensitive-Moderately Sensitive-Not Sensitive
Identify device cyber security requirements	How the IoT device's cyber security capabilities be obstructed by the device's operational limitations?	Low Bandwidth-High Bandwidth

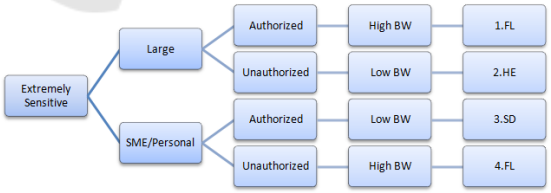

Figure 2: Decision Tree for Extremely Sensitive Data.

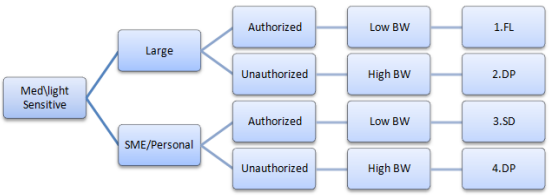

Figure 3: Decision Tree for Medium and lightly Sensitive Data.

Table 2: Reasoning for Recommendations of Extremely Sensitive Data.

1. Reasoning: Since the users are large organizations with authorized access to the IoT data and high bandwidth with extremely sensitive data , the best affordable technique is FL with centralized settings for maximum privacy and minimum computational cost.
2. Reasoning: Since the data is extremely sensitive and the setup is large, HE is affordable. Because the data is accessed by unauthorized users, real data for SD cannot be trusted .
3. Reasoning: SME or Personal organization with low bandwidth can afford to use small quantity of real data to generate SD. Since data is accessed by authorized users only, real data for SD generation can be trusted.
4. Reasoning: Because of the fact that the data is accessed by unauthorized users, SD cannot be used to generate synthetic data but FL can be afforded with High bandwidth. DP cannot be applied because of the extreme sensitivity of the data.

Table 3: Reasoning for Recommendations of Medium and Lightly Sensitive Data.

1.Reasoning: The data is not extremely sensitive and privacy restrictions can be lightened, FL is the best option to go for. The data has not been accessed by unauthorized users and can be trusted for aggregates. Moreover, the less amount of information can be transmitted in low bandwidth unlike SD.
2.Reasoning: With data being med or lightly sensitive and can be accessed by unauthorized users, SD and FL cannot be used since data cannot be completely trusted. DP would be a good option because we have high bandwidth in a larger setup.
3.Reasoning: With less sensitive data, lesser computational power but low bandwidth, SD can be generated from smaller quantity of real data which is being accessed by authorized users and not tempered.
4.Reasoning: With less sensitive data, but access by unauthorized users, data inside cannot be trusted to generate SD. Since the bandwidth is high, DP is the best fit because of the size of the organization. FL would require more computational power than DP.

The reasoning applied on the process of the recommendation of an appropriate PET in the case of medium and lightly sensitive data and extremely sensitive data is explained in Table 3 and Table 2, respectively. In the process of a recommendation, privacy requirements of the data are being analyzed first. Since we are prioritizing privacy on top of all other factors, the decision path starts with the degree of data sensitivity. The next factor we consider is the capability of the users in terms of resources which is determined by the size of the organization. It can be one of the three: small (personal), medium and large. The third factor considered is the type of access to the device and its data (authorized or unauthorized). In the event of unauthorized users, the data should be protected and greater measures should be taken to preserve the privacy of the user's data. The last factor we considered is network bandwidth available to the IoT network of the user. If the bandwidth is low, big volumes of data cannot be transferred. Consequently, an appropriate PET should be assigned that does not generate large data volumes to be transferred. For extremely sensitive data like health and finance and time series data, DP cannot be simply used and should be replaced by the next appropriate PET as indicated in the table.

5 USE CASES

5.1 User Case 1: Smart Watch

Smart Watch is an IoT device with several features commonly used by users. It has personal features such as heart rate monitoring, blood pressure readings, footsteps, temperature, exercise times, sleep patterns, etc. Let us consider the case of a smart watch monitoring the blood pressure readings of a user. These reading must be shared by more than one party for analysis and data mining purposes. The data sharing parties are the family doctor clinic, cardiologist office and the gym used by the end user in order to design exercise patterns. These three parties are not known to each other. Following the paths shown in the decision tree in Figure 2, starting from the fact that the data is extremely sensitive (users health data), the scale of the user is small since it is a single individual (i.e., no heavy computation is possible near the device). Additionally, the smart watch is an IoT device which can be easily lost and therefore, the probability of unauthorized access is high (ruling out synthetic data generation). With good bandwidth available to the end user, we can determine that federated learning is the most appropriate PET in this scenario. Every party involved, processes the data belonging to it and constructs a machine learning model and the model is then consequently shared with a distant centralized server which is accessible to all parties. This central server node will then combine all models to construct a global model which is then returned to the

family doctor, cardiologist office and the gym for independent use. The three parties improve their models by processing knowledge from the distant data sets which are not residing on their servers. Federated learning limits the exposure of information to other parties and therefore protects the privacy of the data. Although federated learning does not guarantee complete privacy protection, the risk is low in severity.

5.2 Use Case 2: Smart Fridge

Smart fridge is a type of an IoT device where the data collected is medium or slightly sensitive unlike health or financial data. This smart device is commonly used by people in their homes to synchronize the data with their phone, shared with the fridge manufacturers to uncover the usage patterns or with any grocery store for item tracking. Following the path in the decision tree in Figure 3, the data is less private, the end user is a single individual with limited computational capability, assuming that the fridge is more difficult to be stolen and no unauthorized access can occur due to the physical security of the house and limited bandwidth for data transfer is available to the user. Furthermore, the system recommends using synthetic data generation as the most appropriate PET. The data is not shared among multiple parties and no collaboration for analysis is required. The smart fridge transmits a subset of real data only (due to less bandwidth availability) to the manufacturer who can use this small volume of real data to simulate the synthetic data for mining the usage. Synthetic data release ensures the privacy of the user data, since no real data with personal information is released to the manufacturers. Although data privacy is ensured by not sharing the real raw data at all, synthetic data is not an exact representation of the real data and might generate a less accurate data to some extent. However, due to the fact that the data is not very critical nor private, a small deviation from the real data is acceptable in this case.

6 CONCLUSIONS AND FUTURE WORK

Data sharing is unavoidable in the current world of IoT computing because the data belonging to different application domains requires a variety of processing for purposes like anomaly detection, fine tuning, data mining, deduplication etc. Due to the volume of the data, processing of such data typically requires a large computational overhead. Organizations often do not have resources to perform required processing on their premises, and need to outsource this processing. In this scenario, PETs can offer a viable and effective tool to protect the privacy of the data after it leaves the owner's domain. In this paper, we have proposed a novel method based on NIST standards to recommend an optimized PET with regard to privacy, efficiency, cost and scalability parameters of the application domain. The proposed framework is easy to use and it can be adjusted to meet changing needs. We have also presented two different use cases to support our framework and explain the concepts applied behind the selection of a particular PET through the proposed framework. One possible future work direction is to incorporate machine learning models to recommend the most optimized PET. The selection framework can be augmented by adding more dimensions in addition to privacy, prioritizing different factors and using more user responses for recommendations.

REFERENCES

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. (2015). On the information leakage of differentially-private mechanisms. *Journal of Computer Security*, 23(4):427–469.

Binjubeir, M., Ahmed, A. A., Ismail, M. A. B., Sadiq, A. S., and Khurram Khan, M. (2020). Comprehensive survey on big data privacy protection. *IEEE Access*, 8:20067–20079.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

El Emam, K., Mosquera, L., and Hoptroff, R. (2020). *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.

Fagan, M., Megas, K. N., Scarfone, K., Smith, M. N. I. o. S., and Technology (May 2020). Foundational cybersecurity activities for iot device manufacturers. Technical Report NISTIR 8259, National Institute of Standards and Technology (NIST).

Farall, F. (2021). Deloitte insights: Data sharing made easy. https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/data-sharing-technologies.html. Accessed August 1, 2023.

Jordan, S., Fontaine, C., and Hendricks-Sturrup, R. (2022). Selecting privacy-enhancing technologies for managing health data use. *Frontiers in Public Health*, 10:814163.

Kairouz, P. and et al (2021). Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1–2):1–210.

Khan, L. U., Saad, W., Han, Z., Hossain, E., and Hong, C. S. (2021). Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, 23(3):1759–1799.

Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., and van der Maaten, L. (2021). Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34:4961–4973.

Lindell, Y. (2020). Secure multiparty computation. *Communications of the ACM*, 64(1):86–96.

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Yu, P. S. (2022). Privacy and robustness in federated learning: Attacks and defenses. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Rastogi, V. and Nath, S. (2010). Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 735–746.

Scaiano, M., Middleton, G., Arbuckle, L., Kolhatkar, V., Peyton, L., Dowling, M., Gipson, D. S., and El Emam, K. (2016). A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, 63:174–183.

Sun, X., Zhang, P., Liu, J. K., Yu, J., and Xie, W. (2018). Private machine learning classification based on fully homomorphic encryption. *IEEE Transactions on Emerging Topics in Computing*, 8(2):352–364.

The Royal Society (2023). From privacy to partnership. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf. Accessed February 18, 2023.

Walker, M., Torchia, M., Chinta, K., Kotagi, S., Roberti, G., Heriberto, R., and Rotaru, A. (2022). Worldwide internet of things forecast, 2022–2026. Technical report, IDC.