

# **A review on differential Privacy for real-world applications: past, present, and future**

Prateek Singh

221218743

## Contents

Topic	Page Number
1. Introduction	3
2. Background of Differential Privacy	3
3. Differential Privacy over other methods	4
4. Differential Privacy in practice	5
5. Tools for privacy analytics	8
6. Future Trends	10
7. Conclusion	10
8. References	11

## 1 – Introduction

Data is the fuel that drives innovation in the domains of Data Analytics and Artificial Intelligence. This statement holds true for various verticals where solutions are delivered using Data Analytics and AI. The efforts to improve our capacity to process and do computations on large volumes of data have led to the birth of hardware capabilities that allow us to find hidden patterns in these datasets. Based on this, there is an ongoing race for data acquisition in corporations and businesses to build products that leverage this huge influx of user data. This leads to a scenario where control of data is one of the major aspects which result in the quality of the solution being delivered. With the advent of these technologies, several concerns which result in a violation of an individual's privacy have been raised.

This results from inappropriate data disclosures which have several underlying implications which can harm individuals as well as corporations' standing in modern-day society. To prevent such scenarios from happening there is a growing consensus for careful use of data in a way such that it does not infringe upon an individual's right to privacy. The traditional methods of data protection are now being replaced with novel methods which further bolster the robustness of data protection and bring forth tangible measures of evaluation of these emerging methods. This review brings to light the shift towards these novel methods like *differential privacy*. This review brings out the flaws and limitations of archaic methods of data protection and discusses the importance of having modern data protection models like *differential privacy* and how they came to be. This review employs discussion on several case studies to show the importance of *differential privacy* across various verticals. The review also discusses the tools (open-source and commercial) that are being used to perform privacy analytics. This review finally summarizes the discussion with implications of employing *differential privacy* as a defence against potential privacy infringements and highlights future trends.

## 2 – Background of differential privacy

[32] defines cryptography as the transforming data or information into a form that is impossible to decode or reproduce without a secret key. Data protection is often ensured by using cryptography i.e., encrypting user data such that only entities with access keys would be able to decrypt the data. [2] – [4]. This method comes with its own set of drawbacks as the computational overhead for these methods is generally huge. Also, ensuring the safe sharing of access keys becomes additionally difficult. As more and more keys are compromised, the differentiation between data protection and data privacy becomes more prominent. The cryptographic methods to ensure data privacy have a single point of failure: *compromised secret key*. This led to adoption of more robust privacy techniques like anonymization and deidentification.

Deidentification brings the promise of anonymized data such that any datapoint could not be linked to any one individual. This moves away from traditional way of data protection and ensures data privacy. One of the ways how data protection is ensured is by removing information from a database that allows an attacker to link the information present in the database with an individual. This process includes using various techniques which result in suppressing the identity of the individual. Generally known as deidentification, the process tries to delink an individual's identity from the information present in the dataset. While the motivation is to suppress the identity, [1] shows how the combination of information present in the de-identified dataset can be used to uniquely link an individual with de-identified information. This defeats the purpose of deidentification as the person can be reidentified from a de-identified dataset, thus posing a severe threat to the privacy of the individual.

Building on top of the concept of deidentification, more robust anonymization techniques like k-anonymity [5] were introduced. This utilizes the concept of storing data in an anonymized way such that no one record can be linked against an individual, much like deidentification, but this method ensures that generalization and suppression are used for quasi-identifiers in groups of k people/rows in the dataset. This ensures that the attacker can know the group but not the individual, although this method does not guarantee privacy protection when the dataset size increases beyond a threshold [6].

These flaws led to the development of a privacy model which addresses these issues and formalizes the concept of data privacy, which is *differential privacy*. Differential privacy works on the underlying principle that two dataset's D and D' should differ only in one row (D: with the individual's record, D': without individual's record), no outputs should become more or less likely [7]. This approach works towards hiding the presence or absence of an individual from a given dataset. This privacy model also allows the control of noise to be added to the dataset to make it differentially private, this gives the user the sense of control of their data while maintaining a healthy trade-off between accuracy and privacy.

### **3 – Differential Privacy over other methods**

The introduction of differential privacy indicates that it is the emerging model which is going to be used for the foreseeable future of data privacy issues [7]. This is because of the controllable parameter *privacy loss* which quantifies the total loss in privacy and conveys the trade-off between accuracy and privacy of the data. Differential privacy is by inheritance applicable to extended groups of data which also gives the control on the loss of privacy on derived groups like families. Apart from providing linkage protection, differential privacy also protects differencing and reconstruction attacks.

The lack of such heuristics for privacy makes the other methods susceptible to failure or prone to privacy infringement and making way for linkage attacks. The introduction of differential privacy brings in a quantifiable measure of privacy which was lacking in other methods, this also lets the user control the amount of noise to be introduced in data to make it differentially private. This leads to a trade-off between accuracy and privacy, a concept lacking from earlier methods.

As the only computation overhead required is to calculate the amount of noise to be added in the database so that it becomes differentially private, this makes differential privacy computationally inexpensive as compared to encryption

Differential privacy comes with an advantage that makes it superior to anonymization techniques, it is the surety that output cannot be distinguished irrespective of the presence or absence of a certain record in the data. This surety makes differential privacy measures more secure and reassuring than anonymizing techniques.

## 4 – Differential Privacy in practice

The following sections discuss the impact of differential privacy across various sectors with the help of case studies and use cases.

### 4.1 – Differential Privacy for internet system

Internet systems are probably the most important space in which privacy infringement can be prevented using differential privacy. We have witnessed a plethora of data leaks that have directly or indirectly affected us. Authors in [9] show the biggest data leaks that have occurred in the recent past, this shows a worrisome trend as private data is being targeted to extract private data of individuals.

#### 4.1.1 – Cambridge Analytica: A case study

Cambridge Analytica scandal has been one of the most infamous data leaks in our recent past, it involved creating a clone of Facebook's database on a private database that could directly be linked with an individual. This information was further used to create a psychometric profile of individuals which was linked to their voting preferences.

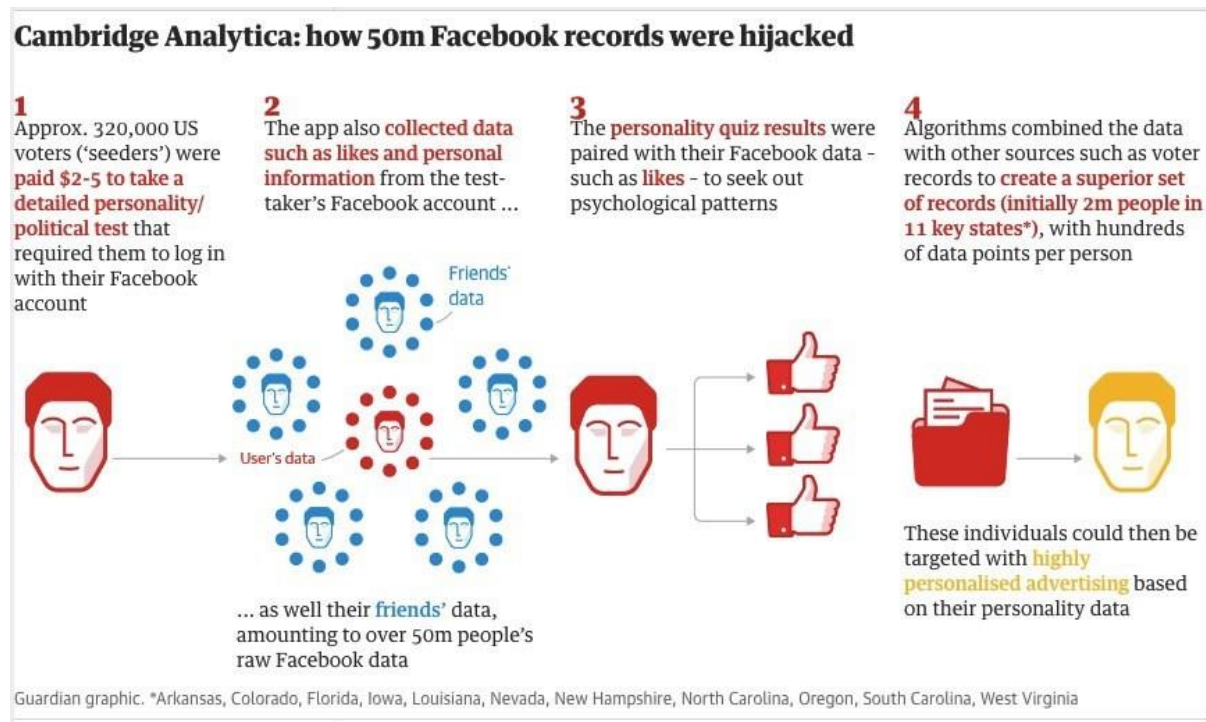


Fig 1: Walkthrough of Cambridge Analytica Scandal

Leaks such as these can be effectively prevented if the platform employs differentially private measures, e.g., in this case, had the data been differentially private when it was sourced, then the leak would not be linked directly with individuals and the impact of leak could be minimized.

We can see that post such leaks; organizations have employed differentially private methods like RAPPOR (Google) [10] which uses local differential privacy to source data from users. Efforts like these are now more common in big tech which has resulted in increased privacy protection for users.

### 4.2 – Differential Privacy for Healthcare

Healthcare systems are increasingly becoming more and more dependent on cyberspace as there is a growing need for coordination for research. This coordination involves sharing of data that consists of

highly sensitive data of an individual's health and wellbeing. This increase in the digital footprint of sensitive data has resulted in many openings for misuse of patients' data. Another major source of health data comes from smart wearables which include body sensors that continuously monitor data and store it on the cloud.

This sort of data is highly susceptible to inference attacks, keeping this in mind Zhang et al. in [11] have proposed *Re-DProctor*, in this, they analyze the utility and privacy trade-off to make the data differentially private. In DAMSON [12] the data is made differentially private, and a collection of prediction techniques has been explored to establish that a differentially private dataset can give real-world applicable accuracy without violating the rights of participating patients. It also boasts of a query optimization engine that is suitable for real-world analytics while minimizing privacy costs.

We can conclude that by careful data perturbation, a differentially private medical dataset can be considered as the optimal solution to protect an individual's privacy.

#### **4.3 – Differential Privacy for Energy System**

Modern-day energy systems rely heavily on real-world analytics for effective distribution and managing user needs in a personalized way. Real-time sensors are employed to measure the consumption of energy which results in derived statistics such as deciding price in real-time, managing surges in demand, monitoring areas with outlier load characteristics. This data can be linked with an individual and can be used to launch attacks such as non-intrusive load monitoring attacks to analyze user behavior related to power usage and smart appliances.

Although the data is not highly sensitive, therefore data privacy is not a huge concern but still, usage patterns can inform the attacker of an individual's routine which may lead to unwanted scenarios like theft or targeted ads based on the insights from appliance usage patterns. This makes it essential for smart grid providers to employ differential privacy such that user data cannot be fetched from any unsolicited query.

Employing differential privacy measures can also result in an added value proposition because of which customers may decide to switch to a provider which boasts of data privacy. In [13] authors have provided a detailed description of how attackers can launch pre-planned attacks to eavesdrop and analyze traffic to perform spoofing attacks. This makes masking of individuals' data necessary to prevent users from such attacks.

#### **4.4 – Differential Privacy for IoT and IIoT**

IoT systems are at severe risk of data leaks as they are prone to attacks such as false data injection which can cause the system to malfunction. The IoT devices rely heavily on the transmission of data from sensors to the cloud or a master server. The nature of the system infrastructure makes it susceptible to attacks which can lead to privacy infringement and system malfunction, the sensors usually also contain sensitive information which can be related to an individual.

In [14] authors have demonstrated how the location of sensors can be extracted and pose a high threat to the industry. Zhu et al. in [15] have reduced mean absolute error and hence resulted in protecting the privacy of data. Similarly, the authors in [16] have introduced noise in the data to make it differentially private.

Industrial IoT systems have a specific set of challenges associated with them. In [17] the authors have discussed various threats that emerge from a lapse in data protection. Another major challenge is deep learning technologies that are used in this sector which rely heavily on the data collection phase,

which results in an opportunity for information leak which can result in unpredictable problems, [18] authors have proposed GANobfuscator, which is differential privacy based generative adversarial network (GAN). This makes IoT and IIOT a promising area that is observing cutting-edge development in the domain of differential privacy as the stakeholders begin to realize the threats to the system and privacy of user data flowing from sensors.

#### **4.5 – Differential Privacy for transportation system**

With the increasing number of smart city projects sprawling, Intelligent Transport Systems are increasingly replacing legacy transport systems and utilizing the power of data to better cater to the needs of commuters. In such systems vehicles are generally connected with other vehicles or a central server or a cloud to send data related to their surroundings. In [19] authors explain in detail how this is achieved by vehicle-to-vehicle (V2V) and device-to-device (D2D) communications. This communication helps gain insights such as traffic flow rate and helps in solutions such as congestion control by monitoring the rate of flow of traffic.

##### **4.5.1 – Example use case:**

A smart city project has a submodule Intelligent Transport System which has public buses connected to a central city traffic police server. The system works upon a variety of real-time information for analytics which includes the traffic density, weather conditions, etc. This system works closely with other modules of the smart city project like vehicle surveillance, accident monitoring, and Intelligent traffic monitoring system which collectively work in providing meaningful insights to stakeholders. These modules make the Intelligent Transport System very effective in handling unforeseen scenarios and work in mitigating load from manual agents.

The presented sample use case shows the huge amount of data that flows between such systems, this makes the system prone to attacks like correlation attacks. In [20] authors show how attackers can launch identity attacks to impersonate the identity of a vehicle. This also opens door to knowing the real-time location of an individual.

Such threats make this a required setting to use a differential privacy model to ensure the privacy of the data flowing through the system. Ma et. al [21] use dynamic sampling to source the real-time location of a vehicle and use differential privacy perturbation to mask private details. By doing so, they make the system robust against adversaries and maintain some level of usefulness in the data as well.

#### **4.6 – Differential Privacy for smart farming**

Farmers are increasingly moving to smart farming technologies to assess soil conditions and the status of crops in real-time. This helps farmers to maximize profit by leveraging data analytics, further IoT devices like drones are now being used to plant seeds [22] which increase productivity and allows farmers to reduce required manual effort. While this migration towards smart farming techniques has its advantages, the potential for data breaches and planned attacks also arises due to the use of such technologies.

To employ countermeasures against these attacks, farmers tend to move towards differential private models either by migrating to the global differential privacy model or local differential privacy. In the former method, a central authority introduces differential privacy, and in the latter, the user introduces differential privacy before the data is sent for analytics. Authors in [23] focus on the latter category, they introduce a deep learning architecture that works on local differential private data to build a model for predictive analysis. This establishes the efficacy of differential privacy in deep

learning scenarios for smart farming, where deep learning models can be trained on a dataset that is differentially private and protects the individual privacy of farmers who share their data for better insights.

## 5 – Tools for privacy analytics

Differential privacy is at the forefront when it comes to creating solutions for data privacy. Many companies are coming up with novel solutions for implementing differential privacy, this has resulted in many open source and commercial tools which aid in achieving differential privacy.

Some of these notable tools are discussed below:

**5.1 – RAPPOR [10]:** RAPPOR stands for Randomized Aggregatable Privacy-Preserving Ordinal Response, this is a technology launched by Google which allows users to upload data without the worry of the data being inked back at them. RAPPOR allows the study of the collection of data (forest of user data but not the individual trees which could be linked to an individual). It is particularly useful for crowdsourcing data and using it for predictive analysis without infringing upon an individual's privacy.

**5.2 – OPACUS [24]:** Pytorch is one the most famous framework which is widely adopted in the AI research community. This makes OPACUS one of the leading differential libraries, launched by Meta (Facebook) OPACUS promises high-speed training with differential privacy. The salient features of this library are its speed, safety, flexibility, productivity, and interactivity.

**5.3 – DiffPrivLib [25]:** This is a general-purpose library from IBM, this library comes with its own set of models that incorporate differential privacy in the training process. The salient features are easy integration and abstract APIs which makes development and prototyping very fast.

**5.4 – TensorFlow Privacy [26]:** The TensorFlow equivalent of OPACUS, this library from Google gives TensorFlow optimizers for training differential private models using the TensorFlow framework – one of the most popular frameworks for creating deep learning pipelines.

**5.5 – Differential Privacy [27]:** This is another library by Google that contains different algorithms used for differential privacy building blocks. This library has implementations in C++, Go, and JAVA. This library presents primitive APIs to build a differential private pipeline that could be customized in a platform-agnostic way. The salient feature of this library is a high speed and flexible customization because of primitive APIs.

**5.6 – OpenMined – PyDP [28]:** This open-source library is a python implementation of the differential privacy library by Google.


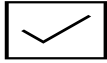
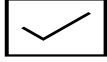
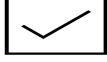
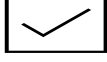
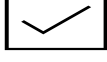
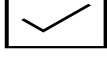
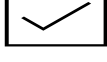
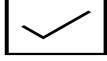
**5.7 – OpenMined – PYSyft [29]:** This open-source library offers to decouple private data from the model training step. The salient feature of this library is that it allows us to write code to process data, which is not owned by the programmer, rather it allows us to process data stored at a remote location like edge devices. This allows us to incorporate local differential privacy effectively as user does have the certainty of data ownership and the only data which leaves the user device is differentially private data.

**5.8 – OpenDP – SmartNoise Core Differential Privacy Library [30]:** This library has been developed in the collaborative effort of Harvard and Microsoft. This library offers bindings in Python and Rust. This



library is also loaded with differentially private algorithms and offers tools to calculate the loss of privacy on a given dataset.

**5.9 – Diffpriv [31]:** This open-source library is one of the limited libraries which support differential privacy in R. This library offers tools to calculate sensitivity and other metrics related to differential privacy.

Tools	Organisation	Supported Languages	Open Source
RAPPOR	Google	Python, R	
OPACUS	Facebook	Python	
DiffPrivLib	IBM	Python	
TensorFlow Privacy	Google	Python	
Differential Privacy	Google	C++, Go, JAVA	
OpenMined – PyDP	OpenMined	Python	
OpenMined – PYSyft	OpenMined	Python	
OpenDP – SmartNoise Core Differential Privacy Library	OpenDP	Python, Rust	
Diffpriv	-	R	

**Table 1:** Table comparing discussed tools available for differential privacy.

## 6 – Future Trends

Owing to increased data sharing and leaks associated with it, there has been a growing concern in the tech and non-tech community alike for infringement of individuals' privacy. Various government and private stakeholders are voicing strong opinions related to maintaining strong data protection frameworks. Official bodies like the EU have tabled laws like General Data Protection Regulation (GDPR - 2016) which show a strong proclivity towards maintaining a strong stance for data protection. This strong stance is underscored by heavy fines/penalties associated with the illegal use of data. The law provides a general framework for consensual data sharing and privacy impact assessment while also ensuring the right to access user data. One key aspect of this regulation is the right to be forgotten which states that the data must be erased once the original purpose has been achieved.

Another legislative framework for user data protection is Health Insurance Portability and Accountability Act (HIPAA – 1996), this act focuses on US patient data. This asserts a patient's control over the sensitive information related to one's well-being by utilizing the concept of consent to data sharing. This act covers various stakeholders like health providers, health plans, business associates, etc. This ensures that a patient is immune to the unsolicited targeted advertisement.

One of the recent additions to the ever-growing list of regulations and frameworks to help strengthen the data protection principles is the California Consumer Privacy Act (CCPA – 2018). This act provides consumers control of their data over businesses that source their data for gaining insights from consumer behavior. The strong point of this act is that the user has the right to know how and what data is used. Another aspect of this act allows the users to ask for the removal of their data once it has been sourced and opt-out of future use of their data.

While these laws provide data protection rights to users, these are generally limited to the geography that they are enforced. There are several other laws across different geographies across the globe that follow similar principles of data protection, but the majority affect the residents of the same geography. Nonetheless, these laws indicate a growing global consensus towards employing comprehensive data protection practices and frameworks. With the emergence of these laws, more and more corporations are utilizing concepts like differential privacy to strengthen their data protection standards. Current growth indicates a blossoming future trend that will build on top of differential privacy.

## 7 – Conclusion

This review establishes the importance of differential privacy by emphasizing the importance of data protection in various use cases like the internet, smart healthcare, IoT, farming, energy, transportation, systems. Another key aspect is bringing out the difference in concepts of data security and data privacy, while the traditional method of encryption offers data security, it does not ensure data privacy. This can be seen in the case Cambridge Analytica case study where encrypted data once leaked violated the principles of data privacy. This shows while data security can be one aspect of data privacy, it cannot be considered as a standalone solution for data privacy.

Differential privacy builds upon the principle of anonymity and takes a quantifiable step towards complete data privacy by introducing a privacy budget parameter. This gives control while building AI and analytics pipelines to ensure the amount of privacy being lost (highlighted by the privacy-accuracy trade-off.) Then the review moves to the practical aspect of implementing differential privacy through various libraries available which can be integrated into existing AI pipelines to ensure data protection.

Future trends indicate how legislative actions are driving businesses and the research community to adopt data protection principles.

## 8 – References

- [1] L. Sweeney, <http://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html>.
- [2] L. Chen, S. Thombre, K. Jarvinen, E. S. Lohan, A. Alen-Savikko, H. Leppakoski, M. Z. H. Bhuiyan, S. Bu-Pasha, G. N. Ferrara, S. Honkala, et al., "Robustness, security, and privacy in location-based services for future IoT: A survey," *IEEE Access*, vol. 5, pp. 8956–8977, 2017.
- [3] K. Muhammad, R. Hamza, J. Ahmad, J. Lloret, H. H. G. Wang, and S. W. Baik, "Secure surveillance framework for IoT systems using probabilistic image encryption," *IEEE Transactions on Industrial Informatics*, in Print, vol. 14, no. 8, pp. 3679 – 3689, Aug. 2018.
- [4] W. Meng, E. W. Tischhauser, Q. Wang, Y. Wang, and J. Han, "When intrusion detection meets blockchain technology: a review," *IEEE Access*, vol. 6, pp. 10 179–10 188, 2018.
- [5] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.
- [6] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy preservation in big data from the communication perspective — a survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 753 – 778, 2018
- [7] C. Dwork, F. McSherry, K. Nissim & A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis, Proceedings of the Third Theory of Cryptography Conference 265 (2006)", [http://dx.doi.org/10.1007/11681878\\_14](http://dx.doi.org/10.1007/11681878_14). – Differential Privacy
- [8] T. Zhu, P. Xiong, G. Li, W. Zhou, and S. Y. Philip, "Differentially private model publishing in cyber-physical systems," *Future Generation Computer Systems*, in Print, 2018.
- [9] <https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html>
- [10] Ú. Erlingsson, V. Pihur, A. Korolova, "Proceedings of the 21st ACM Conference on Computer and Communications Security", ACM, Scottsdale, Arizona, 2014, pages 1054 – 1067.
- [11] J. Zhang, X. Liang, Z. Zhang, S. He, and Z. Shi, "Re-dpdoctor: Real-time health data releasing with w-day differential privacy," *arXiv preprint arXiv:1711.00232*, 2017.
- [12] M. Winslett, Y. Yang, and Z. Zhang. "Demonstration of damson: Differential privacy for analysis of large data. In 18th IEEE International Conference on Parallel and Distributed Systems," ICPADS 2012, Singapore, December 17-19, 2012, pages 840–844, 2012
- [13] W. Wang, Y. Xu, and M. Khanna, "A survey on the communication architectures in smart grid", *Computer Networks*, vol. 55, no. 15, pp. 3604-3629, 2011.
- [14] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial Internet-of-things," *IEEE Transactions on Industrial Informatics*, in Print, vol. 14, no. 8, pp. 3628 – 3636, 2017.

- [15] M. Li, W. Lou, and K. Ren, "Data security and privacy in wireless body area networks," *IEEE Wireless communications*, vol. 17, no. 1, 2010.
- [16] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," *Future Generation Computer Systems*, in Print, 2018.
- [17] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: Opportunities, applications, and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10430–10451, 2021.
- [18] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "Ganobfuscator: Mitigating information leakage under gan via differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2358–2371, 2019.
- [19] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1858 – 1877, 2018.
- [20] A.-S. K. Pathan, *Security of self-organizing networks: MANET, WSN, WMN, VANET*. CRC Press, 2016.
- [21] Z. Ma, T. Zhang, X. Liu, X. Li, and K. Ren, "Real-time privacy-preserving data release over vehicle trajectory," *IEEE Transactions on Vehicular Technology*, in Press, pp. 1–1, 2019.
- [22] P. Sanjeevi, B. Siva Kumar, S. Prasanna, J. Maruthupandi, R. Manikandan, and A. Baseera, "An ontology enabled internet of things framework in Intelligent Agriculture for Preventing post-harvest losses," *Complex & Intelligent Systems*, 2020.
- [23] R. Udendhran and M. Balamurugan, "Towards secure deep learning architecture for smart farming-based applications," *Complex & Intelligent Systems*, vol. 7, no. 2, pp. 659–666, 2020.
- [24] A. Yousefpour et al., "Opacus: User-friendly differential privacy library in PyTorch," *arXiv [cs.LG]*, 2021.
- [25] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher, "Diffprivlib: The IBM Differential Privacy Library," *arXiv [cs.CR]*, 2019.
- [26] <https://github.com/tensorflow/privacy>
- [27] <https://github.com/google/differential-privacy>
- [28] <https://github.com/OpenMined/PyDP>
- [29] <https://github.com/openmined/pysyft>
- [30] <https://github.com/opendp/smartnoise-core>
- [31] B. I. P. Rubinstein and F. Aldà, "Pain-free random differential privacy with sensitivity sampling," *arXiv [cs.LG]*, pp. 2950–2959, 06--11 Aug 2017.
- [32] G. J. Simmons, "cryptology," *Encyclopedia Britannica*. 17-Aug-2016.

9144068

Assignment9\_1.docx (155.05 KB)  
Turnitin® Submission ID 1755463161

18 %

06 February, 2022 12:33 AM

*N.B. Similarity score including reference list.*