**Privacy in the big data era**

Davi Grossi Hasuda

Committee Composition:

|  |  |  |
| --- | --- | --- |
| Prof. | ............................................ | Prof. Dr. Juliana de Melo Bezerra- ITA |
| Prof. | ............................................ | Prof. Dr. Ana Lorena |
|  |  |  |
|  |  |  |

**ITA**

**Resumo**

Diariamente vários *terabytes* de dados são coletados de todos os usuários da internet. E toda essa informação é guardada e processada por várias empresas ao redor do mundo. Todos esses *bytes* são de muito valor para se conseguir entender o mercado, aprender sobre os anseios humanos, construir novos produtos e dar uma direção mais clara aos já existentes. Informação se tornou algo de muito valor e essencial a praticamente toda atividade de produção humana, e no frenesi de se capturar mais e mais dados acaba-se, muitas vezes, invadindo a privacidade das pessoas.

Não raro encontram-se casos em que a privacidade das pessoas foi violada quando utilizando a *internet*. Necessita-se, pois, de ferramentas e processos que permitam, ao mesmo tempo, a captura de dados relevantes para as tecnologias hoje sendo desenvolvidas e a manutenção da privacidade de cada indivíduo. Nunca isso foi uma tarefa simples, mas dada a escala das aplicações que estão em produção na atualidade, é possível desenvolver técnicas que se alavancam de tal escala para que a privacidade seja mantida.

Este trabalho busca, ao seu fim, estudar algumas das técnicas de privatização de dados e oferecer um guia para quem desejar implementá-las. Espera-se, com o encerramento do trabalho, que as principais técnicas de privatização tenham sido validadas empiricamente e que haja diretrizes claras de como utilizá-las.

**Abstract**

Every day an unimaginable amount of data is collected from internet users. All this data is stored and processed by a lot of companies around the world. All these bytes have great value and are essential for a better understanding of the market, learning about human wishes, building new products and building new products. Information has become something truly valuable and imperative to most human activities. In this frenzy of capturing as most data as possible, privacy is very often put at risk.

Frequently there is a new case where someone’s privacy has been violated, and, in some cases, this affects millions of people. Therefore, there is a need for better tooling and processes that enable, at the same time, capturing the relevant data needed for the technologies in development and preserving the privacy of each person. It is not a simple task but leveraging from the scale of the applications that are in production, such tools and processes are viable.

This work aims to study some of the data privatization techniques and build a guide for their implementation. It is expected that, by the end of this work, that such techniques will be validated empirically and there will be clear directives to implement them.

**Contents**

[1 Introduction 1](#_Toc11264334)

[1.1 Motivation 2](#_Toc11264335)

[1.2 Objective 3](#_Toc11264336)

[1.3 Outline 4](#_Toc11264337)

[2 Background 6](#_Toc11264338)

[2.1 A simple DP method (Coin Mechanism) 6](#_Toc11264339)

[2.2 Differential Privacy 8](#_Toc11264340)

[2.3 Online vs Offline Differential Privacy 10](#_Toc11264341)

[2.4 The Laplace Mechanism 10](#_Toc11264342)

[2.5 Related work 12](#_Toc11264343)

[2.6 Chapter Summary 13](#_Toc11264344)

[3 Conclusions 14](#_Toc11264345)

[4 References 15](#_Toc11264346)

1. Introduction

It is noticeable that data is an important asset of the globalized world. As stated in [1], every minute, Google conducts 3.877.140 searches and 4.333.560 YouTube videos are viewed. There is, every moment, a lot of new data being generated by all the users of the internet worldwide, and it is not simply trash data, it is actually useful, and that why many companies invest a lot in storing and processing all the data they can collect.

Amazon.com, for instance, developed its Recommender System using from almost all of its users [3]. In one of Amazon’s first approaches, they searched for users with similar interests, and made suggestions based on this similarity.

Generally, for companies to understand how the user experience is evolving with the product, they have to collect user data. And there are a lot of ways of doing that, and a lot of research about this topic [5]. All of this shows how important data can be and how it can shape the final product delivered to the customer.

Another important factor of the globalized world is the fact that some of this data is used to train neural networks, and very often this training requires a great amount of data. Face ID, for example, the Apple’s system to recognize someone’s face and authenticate based on that, took over 1 billion images to train its neural network [2].

With all of that in mind, is expected that all collectable data is being collected – or at least there are people trying to collect it. Another example of intense use of data is in Target Corporation Inc, where they wanted to know which customers were pregnant, so they could do focused promotions for them [4]. But, in the end, part of their strategy also surrounds not letting the pregnant customers know that Target knew they were pregnant. It is due to the fact that it would make most of this customer suspicious about how Target got this information, so they would avoid buying at Target. This comes to show that personal data is also sensitive and must not be abused.

The world has come to a point where the focus of getting and processing as much data as possible sometimes blind us about the dangers of making it with no or little responsibility. It begs for the need of updated regulations and processes in order to preserve people’s privacy.

* 1. Motivation

In the midst of the frenzy of collecting data, many times privacy is jeopardized [8]. One of the most emblematic case was the scandal involving Facebook and Cambridge Analytica [6], where Facebook provided the data and Cambridge Analytica used it improperly to influence the presidential run in the United States.

In this case, Facebook API let developers have access not only to data from people that gave them this permission, but also access their friend’s data on Facebook. This way, Cambridge Analytica had access to data from over 50 million people. And all of this was used to leverage Donald Trump’s campaign.

There are other cases of data breach. Another example was with Tanium, a cybersecurity startup, that exposed the network of a client without permission [7].

Concerned about these privacy scandals, there are some efforts emerging in order to preserve privacy. GDPR, for instance, is the General Data Protection Regulation [9] from the European Union that rewrites how data sharing must work on the internet. It’s a government effort and establish constraints and rules when accessing user’s data or sharing it. It also establishes some policies for intervention on the stored data.

Another effort, and the one that will be the focus of this work, is Differential Privacy [12]. This establishes constraints to algorithms that concentrate data in a statistical database. Such constraints limit the privacy impact on individuals whose data is in the database.

Another factor that is important to keep in mind is that some simple anonymization processes can be very ineffective [10, 11]. Take for example the 2006 Netflix Prize, a competition promoted by Netflix where competitors must develop a algorithm to predict ratings from users. For that, Netflix shared a dataset with over 100 million ratings by over 480 thousand users. All the names were removed, and some fake ratings were added. But as shown later, it was not enough, since a de-anonymization process was possible [11] by comparing the Netflix dataset with an IMDb dataset.

So, the process that privatize data must also be linkage attack-proof. Also, it must not compromise the final result of machine learning algorithms and statistical studies where they are used.

Fortunately, Differential Privacy takes that in count. And also provides a way of measuring privacy [13]. After the privatization process, it is expected that the data still useful [14]. But most of the papers are theory focused, and it is not broadly used yet [7, 8]. It is missing more pragmatic ways of implementing DP algorithms. In **Table 1** there is a selection of DP papers and a column distinguishing its focus (whether it has a theorical focus, a practical focus, or there is a balance between both).

* 1. Objective

The goal of this project is to develop a guide for applying Differential Privacy methods. It will start by studying the most used DP methods (Laplacian and Exponential, for instance) and compare them. After the initial study, it is expected to have a practical guideline for some use cases with real world examples of use of such algorithms.

One important aspect of this work is to understand the impact of DP methods in data analysis and be able to identify directives of usage for getting started in data privatization.

There are two main types of privatization: Online (or adaptative or interactive) and Offline (or batch or non-interactive). The first one depends on the queries made and the number of them (which can be limited). The second type of privatization does not make assumptions about the number or type of queries made to the dataset, so all the data can be stored already privatized. The focus of this work will be in the Offline methods of DP.

* 1. Outline

In chapter 2, we provide a background for a better understanding of privacy and, more specifically, of Differential Privacy, which is the focus of this work. We do that by first presenting a simple example of a privatization mechanism (we called it the coin mechanism) and then showing that it is a differentially private mechanism.

Along with this example, the formal definition of Differential Privacy is presented, with great efforts to make its understanding as intuitive as possible. To help with that, some variations of the coin mechanism are presented. These variations aim to provide a better understanding of how the parameters in the Differential Privacy are correlated with the privacy of the data.

Also, in chapter 2, there is an introduction to the Laplace Mechanism, which uses a laplacian noise to privatize data. This is an important method to learn before entering the rest of Differential Private mechanisms, since it provides a great and solid base for a better understanding of the core concepts of Differential Privacy Methods.

By the end of the chapter 2, there is a table with a collection of DP related papers that tries to show that most of these papers are theoretically biased and provide little help for implementing DP methods into real world systems.

Finally, in chapter 3, there are presented the next steps of this project. And the main goals we expect to achieve.

1. Background

In this chapter, a DP mechanism is introduced and paths the way for a formal definition of Differential Privacy. After that, there is a glance at some DP methods.

The first privatization method presented in this chapter is a very simple one [30] that had the aim of eliminating evasive answer bias by showing to subjects that their privacy would be preserved regardless of their answer.

When presenting the mathematical definition of Differential Privacy, there is an effort for making it as intuitive as possible and showing how it prevents the privatized data from linkage attacks.

The final part of this chapter glances at some DP methods and brings the path for the continuation of this work.

* 1. A simple DP method (Coin Mechanism)

In reference [30] there is a description of a simple DP method. In this experiment, the goal was to collect data that may be sensitive to people and, because of that, they might be willing to give a false answer, in order to preserve their privacy.

Let’s suppose we want to make a survey to know how many people make use of illegal drugs. It is expected that many people that do use illegal drugs might lie in their answer. But in order to get a clear look at the percentage of people that use illegal drugs, we can use the coin mechanism in order to preserve people’s privacy. It goes like **Figure 1** shows: when registering someone’s answer, first a coin is tossed. If the result of the first toss is “*Heads*”, we register the answer the person gave us. On the other hand, if the result of the first toss is “*Tails*”, we toss the coin again. Being the second result “*Heads*”, we register “yes” (the person do use illegal drugs); being “*Tails*”, “no” (the person do not use illegal drugs).

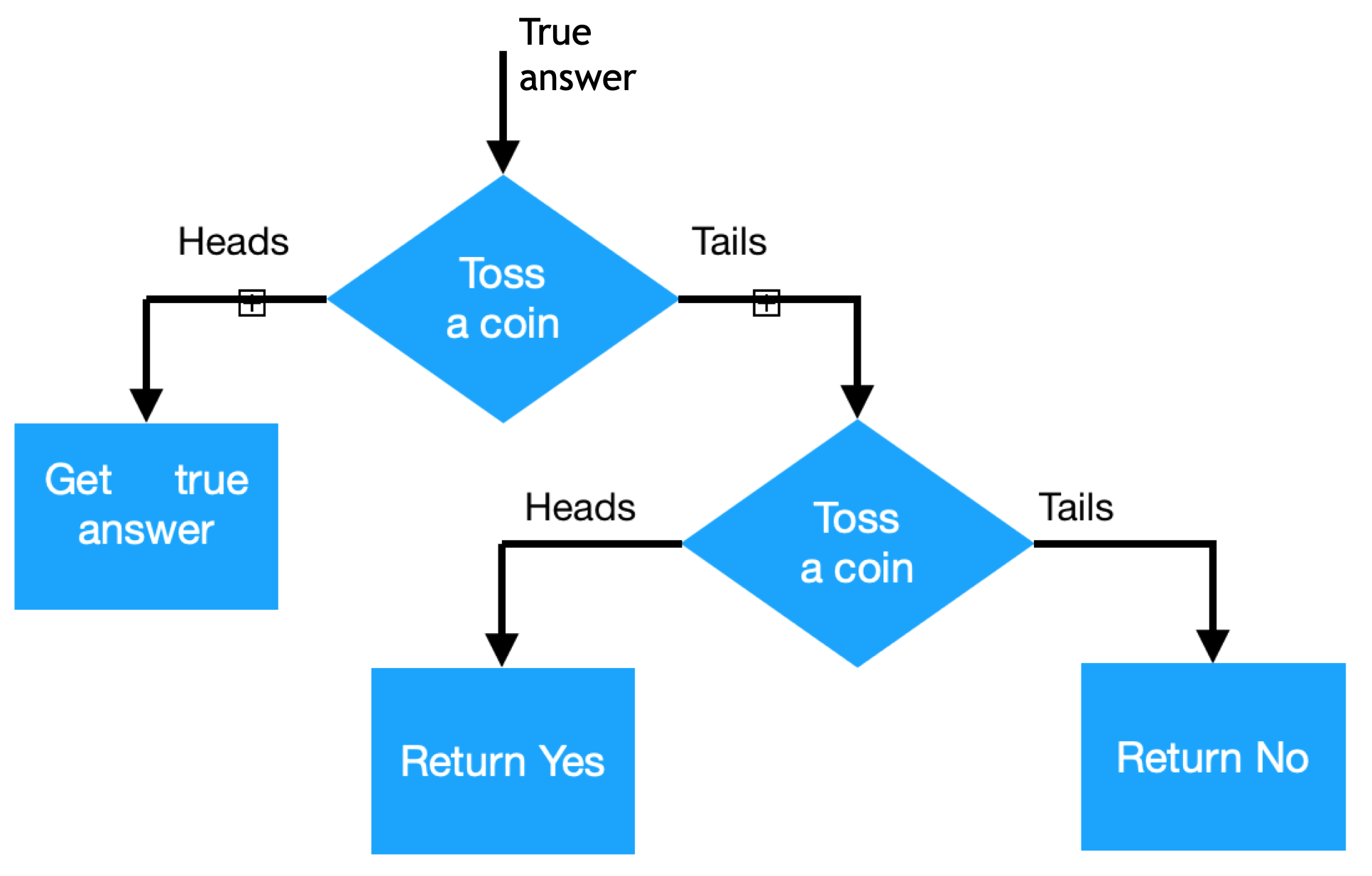


Figure 1. Coin Mechanism schema.

By the end of the experiment, there will be a database with answers from all the subjects, but it is expected that 50% (assuming that the coin has a 50% chance of getting each result) were artificially generated. So if we look at the answer of a single person, there will be no certainty if that was the true answer of his or her.

At the same time, if we subtract 25% of the total answers and which the answer is “yes” and 25% of the total answers and which the answer is “no”, we can have a clear view of the percentage of the population make use of illegal drugs. It was possible, concomitantly, to have a statistically accurate result (assuming that there were enough people involved in the study) and preserve everyone’s privacy.

The amount of privacy given to subjects can also be adjusted by changing the probability of the first coin to have a “*Head*” or “*Tail*” result. The bigger the chance of a “Tail” in the first coin, the bigger the privacy of the subjects.

In the extreme case, where the first coin toss always returns “*Tails*”, we would achieve maximum privacy, but it is important to notice that the final data would have no utility and would provide no statistic result on the population. In the case were the first toss always returns “*Heads*”, there would be no privacy, since all stored answers are exactly the one the person gave.

* 1. Differential Privacy

The basic structure of a DP Method is the one shown in **Figure 2**.

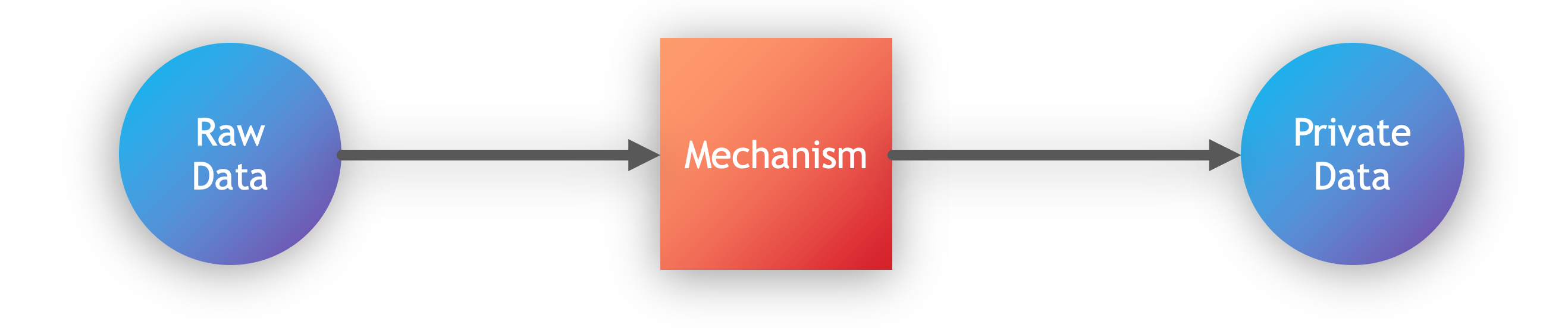


Figure 2. Basic structure of a DP Method.

It consists of a Mechanism that has the not privatized data as input, and outputs the privatized data. Differential Privacy establishes constraints this mechanism must conform to in order to limit the privacy impact of individuals whose data is in the dataset.

A mechanism with domain is -differentially private if for all and for all such that :

In this definition, we have as the probability of a certain event happening and .

What this definition is making is comparing two datasets () that are neighbors () and seeing the probability of the resulted dataset after privatization being alike.

In the case of the Coin Mechanism, explored in topic 2.1, if is a dataset of answers, a neighbor dataset can be, for example, that same dataset , but with another answer from a new person. The only difference between these datasets is the answer of one person (that does not exist in ). Preserving privacy from using a mechanism means that and are expected to be very similar. In other words, one person’s answer should not have a great impact in the final result, in order to preserve this person’s privacy.

Thinking this way, the bigger the parameter , the less private the mechanism is; and the smaller the parameter , the more private the mechanism is.

As a simple exercise, let’s calculate the privacy of the Coin Mechanism. First it is assumed that there is a 50% chance of the coin toss to output “*Heads*”, and another 50% of outputting “*Tails*”. Fix a respondent and we can see neighbor datasets as changing only one person’s true answer. In this line of thought, we can write:

This way, it is possible to see that the Coin Mechanism is ()-differentially private.

Making a little modification to this experiment, assuming that the first coin toss has a 75% of chance of outputting “*Heads*” (and the second coin toss continues with the 50% chance), the new value is:

Therefore, with this modification, the mechanism becomes is ()-differentially private.

Extending this concept to the extremes, it’s possible to calculate that if the first coin toss always outputs “*Heads*” (zero privacy), the result is . If the first coin toss always outputs “*Tails*” (maximum privacy and no utility), we get .

* 1. Online vs Offline Differential Privacy

There are two main ways of privatizing data: online (or adaptative or interactive) and offline (or batch or non-interactive). To better understand these models, let’s first put two definitions: query and curator. A query is a function to be applied to a database, usually to retrieve some data from it. Curator is the one responsible for holding the data from individuals in a database.

The online model permits the data analyst to ask queries adaptively, and this model can impose constraints on the queries made. The queries posed by the data analyst can be based on the response to previous queries. In this model, usually the original (raw) data persists in some database, but it is not accessed directly, the curator adds noise to it based on the query and on the data analyst requesting the data.

The offline model produces a sanitized database once and for all. After collecting the original (raw) data, the curator privatizes it, and there is no need for keeping the original data. In this case there are no true restrictions to the number of queries that can be made to the privatized dataset.

If the queries that will be made are not known, the online model must be a better implementation, since it must adapt for each new query, but it also comes with severe challenges being able to provide answers to all possible queries. The offline mode is more indicated in cases where the queries are already known in advance, and it should provide the best accuracy.

The focus of this work will rely on the offline model.

* 1. The Laplace Mechanism

One of the most used DP mechanisms is the Laplace Mechanism, and it is based in the Laplacian distribution, shown in **Figure 3**.

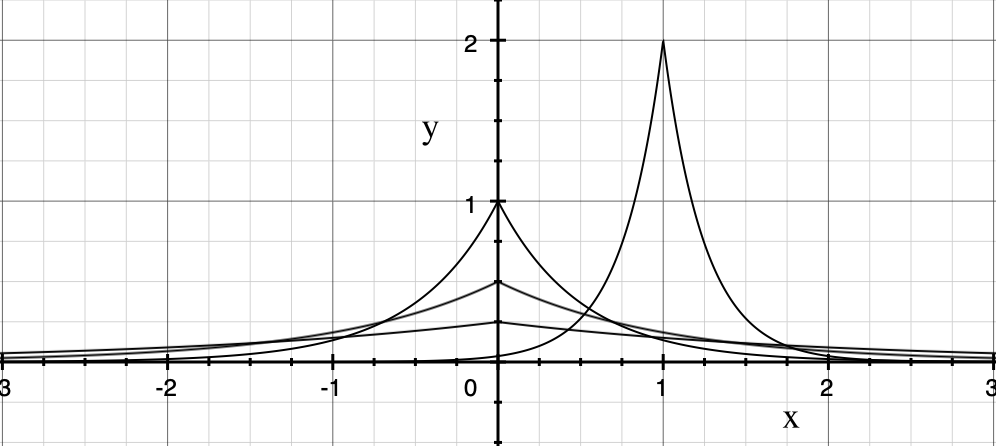


Figure 3. Laplace Distribution examples.

The equation that describes the Laplace Distribution is:

If the value of is increased, for example, the curve becomes less concentrated and more spread. The value is de mean of the distribution. And this distribution can be useful in DP for adding noise to the original dataset. The Laplace Mechanism adds Laplacian noise, as defined in the sequence.

Given any function , the Laplace Mechanism is defined as:

Where are independent and identically distributed random variables drawn from ), being a Laplacian noise with and .

For this definition be complete, it is defined: .

With such constraints imposed to the Laplacian noise, it is possible to demonstrate [13] that this is a -differentially private mechanism. And this is one of the most popular DP method, and has great flexibility for many different types of data.

* 1. Related work

There is some related work in the matter of Differential Privacy, and it is presented here in **Table 1** some of them explicating on what each paper is focused: whether it has a theorical or a practical approach (or if it has a great balance between both).

**Table 1.** Comparison of the focus (practical, theorical or both) of some DP related papers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Paper** | **Authors** | **Year** | **Theorical or Practical** |
| Deep learning with differential privacy. | Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. | 2016 | Practical |
| Private empirical risk minimization: Efficient algorithms and tight error bounds. | Bassily, R., Smith, A. D., and Thakurta, A. | 2014 | Theorical |
| Concentrated differential privacy: Simplifications, extensions, and lower bounds. | Bun, M. and Steinke, T. | 2016 | Theorical |
| Differential privacy. | Dwork, C. | 2006 | Theorical |
| Calibrating noise to sensitivity in private data analysis. | Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. | 2006 | Theorical |
| The algorithmic foundations of differential privacy. | Dwork, C. and Roth, A. | 2014 | Theorical |
| Concentrated differential privacy. | Dwork, C. and Rothblum, G. N. | 2016 | Theorical |
| On the theory and practice of privacy-preserving bayesian data analysis. | Foulds, J. R., Geumlek, J., Welling, M., and Chaudhuri, K. | 2016 | Both |
| (near) dimension independent risk bounds for differentially private learning. | Jain, P. and Thakurta, A. G. | 2014 | Theorical |
| Mechanism design via differential privacy. | McSherry, F. and Talwar, K. | 2007 | Theorical |
| Differential privacy without sensitivity. | Minami, K., Arai, H., Sato, I., and Nakagawa, H. | 2016 | Theorical |
| Renyi differential privacy. | Mironov, I. | 2017 | Theorical |
| Variational bayes in private settings (VIPS). | Park, M., Foulds, J. R., Chaudhuri, K., and Welling, M. | 2016 | Theorical |
| Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. | Talwar, K., Thakurta, A., and Zhang, L. | 2014 | Theorical |
| The complexity of differential privacy. | Vadhan, S. P. | 2017 | Theorical |
| Privacy for free: Posterior sampling and stochastic gradient monte carlo. | Wang, Y., Fienberg, S. E., and Smola, A. J. | 2015 | Theorical |
| Randomized response: A survey technique for eliminating evasive answer bias. | Warner, S. L. | 1965 | Practical |
| Differentially private stochastic gradient descent for in-rdbms analytics. | Wu, X., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. F. | 2016 | Both |
| On the differential privacy of bayesian inference. | Zhang, Z., Rubinstein, B. I. P., and Dimitrakakis, C. | 2016 | Theorical |
| Learning with Privacy at Scale | Differential Privacy Team, Apple | 2017 | Practical |

* 1. Chapter Summary

In this chapter, was defined what Differential Privacy is, and how to measure the level of privacy of a DP Mechanism. A simple example of a DP mechanism – the coin mechanism – was presented to illustrate how privacy is preserved as well as the utility of the data.

With all this background, there was an introduction to the Laplace Mechanism, which uses a Laplacian noise to give privacy to the data. And some papers about DP were presented in order to indicate the theorical bias of most publications in this subject.

1. Conclusions

In this final chapter, the next steps will be presented. As the we aim to develop a guideline to assist the implementation of DP methods, we are going to start by better understanding the most popular methods used nowadays.

So, one of the next steps is to list the restrictions for each of the most popular DP Methods. After that, by applying these methods to real world demands, it is expected for us to generate some examples of DP application that can help in more specific cases.

One important task that must also be accomplished is the measure of the impact of the use of DP methods in the final data analysis. So, we must compare the result of the use of data with and without privatization in statistical analysis and in machine learning algorithms.

Some of the AI methods we aim to test how DP influences in the results are Decision Tree, SVM (support-vector machines), k-nearest neighbors, Naïve Bayes Classifier and MLP (multilayer perceptron).

1. References
2. DIGITAL INFORMATION WORLD. **How Much Data Is Generated Per Minute? The Answer Will Blow Your Mind Away [Infographic].** Available at: <<https://www.digitalinformationworld.com/2018/06/infographics-data-never-sleeps-6.html>>. Accessed on: 11 Jun. 2019.
3. **An On-device Deep Neural Network for Face Detection**. Apple Computer Vision Machine Learning Team.Available at: <<https://machinelearning.apple.com/2017/11/16/face-detection.html>>. Accesses on: 11 Jun. 2019.
4. SMITH, BRENT; LINDEN, GREG. **Two Decades of Recommender Systems at Amazon.com**. Available at: <<https://pdfs.semanticscholar.org/0f06/d328f6deb44e5e67408e0c16a8c7356330d1.pdf>>. Accessed on: 11 Jun. 2019.
5. FORBES. **How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did**. Available at: <<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#2f09b8ef6668>>. Accessed on: 11 Jun. 2019.
6. HAVIR, DAVID. **A Comparison of the Approaches to Customer Experience Analysis**. Available at: <<https://www.researchgate.net/publication/320000579_A_Comparison_of_the_Approaches_to_Customer_Experience_Analysis>>. Accessed on: 19 Jun. 2019.
7. THE NEW YORK TIMES. **Facebook and Cambridge Analytica: What you need to know as Fallout Widens**. Available at: <<https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html>>. Accessed on: 19 Jun. 2019.
8. THE WALL STREET JOURLNAL. **Cybersecurity Startup Tanium Exposed California Hospital’s Network in Demos Without Permission**. Available at: <<https://www.wsj.com/articles/cybersecurity-startup-tanium-exposed-california-hospitals-network-in-demos-without-permission-1492624287>>. Accessed on: 19 Jun. 2019.
9. BUFFERED VPN. **The Biggest Privacy & Security Scandals of 2017**. Available at: <<https://buffered.com/blog/the-biggest-privacy-security-scandals-of-2017/>>. Accessed on: 19 Jun. 2019.
10. GDPR. Available at: <<https://eugdpr.org/>>. Accessed on: 19 Jun. 2019.
11. **Data** **re-identification**. Available at: <<https://en.wikipedia.org/wiki/De-anonymization>>. Accessed on: 19 Jun. 2019.
12. NARAYANAN, ARVIND; VITALY SHMATIKOV, VITALY. **Robust De-anonymization of Large Sparse Datasets**. Available at: <<https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf>>. Accessed on: 19 Jun. 2019.
13. **Differential privacy**. Available at: <<https://en.wikipedia.org/wiki/Differential_privacy>>. Accessed on: 19 Jun. 2019.
14. DWORK, C.; ROTH, A . **The algorithmic foundations of differential privacy**. 2014.
15. ABADI, M.; CHU, A.; GOODFELLOW, I.; MCMAHAN, H. B.; MIRONOV, I.; TALWAR, K.; ZHANG, L**. Deep learning with differential privacy**. 2016.
16. BASSILY, R.; SMITH, A. D.; THAKURTA, A. **Private empirical risk minimization: Efficient algorithms and tight error bounds**. 2014.
17. BUN, M.; STEINKE, T. **Concentrated differential privacy: Simplifications, extensions, and lower bounds**. 2016.
18. DWORK, C. **Differential privacy**. 2006.
19. DWORK, C.; MCSHERRY, F.; NISSIM, K.; SMITH, A. D. **Calibrating noise to sensitivity in private data analysis**. 2006.
20. DWORK, C.; ROTH, A**. The algorithmic foundations of differential privacy**. 2014.
21. DWORK, C.; ROTHBLUM, G. N. **Concentrated differential privacy**. 2016.
22. FOULDS, J. R.; GEUMLEK, J.; WELLING, M.; CHAUDHURI, K. **On the theory and practice of privacy-preserving bayesian data analysis**. 2016.
23. JAIN, P.; THAKURTA, A. G**. (near) dimension independent risk bounds for differentially private learning**. 2014.
24. MCSHERRY, F.; TALWAR, K. **Mechanism design via differential privacy**. 2007.
25. MINAMI, K.; ARAI, H.; SATO, I.; NAKAGAWA, H. **Differential privacy without sensitivity**. 2016.
26. MIRONOV, I. **Renyi differential privacy**. 2017.
27. PARK, M.; FOULDS, J. R.; CHAUDHURI, K.; WELLING, M. **Variational bayes in private settings (VIPS).** 2016.
28. TALWAR, K.; THAKURTA, A.; ZHANG, L. **Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry**. 2014.
29. VADHAN, S. P.  **The complexity of differential privacy**. 2017.
30. WANG, Y.; FIENBERG, S. E.; SMOLA, A. J. **Privacy for free: Posterior sampling and stochastic gradient monte carlo**. 2015.
31. WARNER, S. L. **Randomized response: A survey technique for eliminating evasive answer bias**. 1965.
32. WU, X.; KUMAR, A.; CHAUDHURI, K.; JHA, S.; AND NAUGHTON, J. F. **Differentially private stochastic gradient descent for in-rdbms analytics**. 2016.
33. ZHANG, Z., RUBINSTEIN, B. I. P., AND DIMITRAKAKIS, C**. On the differential privacy of bayesian inference**. 2016.
34. **Learning with Privacy at Scale**. DIFFERENTIAL PRIVACY TEAM, APPLE. Available at: <<https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf>>. Accessed on: 19 Jun. 2019.