dedication (optional)

Summary

Write your summary here...

Preface

Write your preface here...

Table of Contents

List of Tables

List of Figures

Abbreviations

Symbol	=	definition

Chapter 1

Introduction

1.1 Equations

To write an equation

```
\begin{eqnarray}\label{eq1}
F = m \times a
\end{eqnarray}
```

This will produceasdasd asdf asdf asdf

$$F = m \times a \tag{1.1}$$

To refer to the equation

```
\eqref{eq1}
```

This will produce (1.1).

1.2 Figures

To create a figure

```
\begin{figure}[h!]
  \centering
    \includegraphics[width=0.5\textwidth]{fig/pikachu}
  \caption{Pikachu.}
\label{fig1}
\end{figure}
```

To refer to the figure

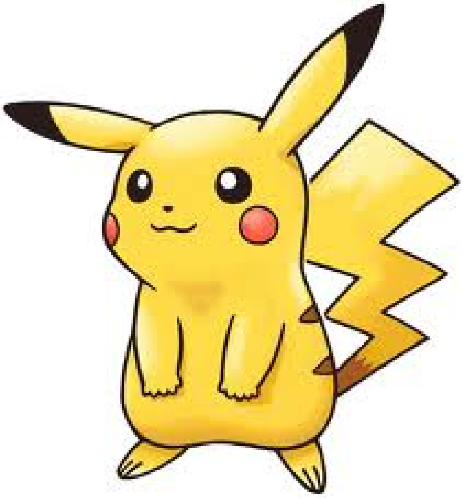

Figure 1.1: Pikachu.

```
\textbf{Fig. \ref{fig1}}
```

This will produce **Fig. 1.1**

1.3 References

To cite references

```
\cite{1,2,3}
```

or

```
\citep{1,2,3}
```

This will produce: [8, 2, 6] or [8, 2, 6], respectively.

1.4 Tables

To creat a table

```
\begin{table}[!h]
\begin{center}
    \begin{tabular}{ | l | l | l | l |}
    \hline
    \textbf{No.} & \textbf{Data 1} & \textbf{Data 2} \\ \hline
     1 & a1 & b1 \\ \hline
     2 & a2 & b2 \\ \hline
    \end{tabular}
\end{center}
\caption{Table 1.}
\label{Tab1}
\end{table}
```

This will produce

No.	**Data 1**	**Data 2**
1	a1	b1
2	a2	b2

Table 1.1: Table 1.

To refer to the table

```
\textbf{Table. \ref{Tab1}}
```

This will produce **Table. 1.1**.

Chapter 2

Literature Review

Chapter 3

Basic Theory

In a world where massive amounts of sensitive personal data are being collected, attacks on the individual's privacy are becoming more and more of a threat. One type of attack is the identification of an individual's personal information from massive data sets, such as people's movie ratings from the Netflix data set[7], and the medical records of a former governor of Massachusetts[1]. These types of privacy breaches may lead to the unwanted discovery of a person's embarrassing information, and could also lead to the theft of an individual's private data or identity.

Many different approaches have been tried by data custodians to privatize the data they hold, such as removing any columns containing Personally Identifiable Information (PII), anonymizing the data by providing k-anonymity protection[9], or perform group based anonymization through l-diversity[5]. All of these methods mentioned have been proved to be susceptible in some way or form to attacks [4]. Motivated by these shortcomings, a researcher at Microsoft came up with a data theoretical framework called differential privacy, which operates off a solid mathematical foundation and have strong theoretical guarantees on the privacy and utility of the released data.

3.1 Differential Privacy

The term ”differential privacy” was defined by Dwork as a description of a promise, made by a data holder to a data subject: ”You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.” [3] In an ideal situation, databases which implement differential privacy mechanisms can make confidential data widely available for accurate data analysis, without resorting to data usage agreements, data protection plans, or restricted views. Nevertheless, the Fundamental Law of Information Recovery states that overly accurate answers to too many questions will destroy privacy in a spectacular way [3], meaning that data utility will eventually be consumed.

3.1.1 Definition of Differential Privacy

The classic example for explaining a security breach is the case of Mr White: Suppose you have access to a database that allows you to compute the income of all residents in a specified area. If you knew that Mr White was going to move, simply querying the database before and after his relocation would allow you to deduce his income.

Definition 1: a mechanism $\tilde{f}$ is a random function that takes a dataset D as input, and outputs a random variable $\tilde{f}$(D).

Definition 2: the distance of two datasets, d(D, D), denotes the minimum number of sample changes that are required to change D into D.

Formally, differential privacy is defined as follows: A randomized function f gives ϵ-differential privacy if for all data sets D and D' differing on at most one row, and all S⊆Range($\tilde{f}$),

$$Pr[\tilde{f}(D) \in S] \leq \mathrm{e}^{(\epsilon)} \times P[\tilde{f}(D') \in S]$$

What this means is that the risk to an individual's privacy should not be substantially increase as a result of participating in a statistical database, as the risk is bounded by the parameter ϵ. Therefore an attacker should not be able to learn anything about any participant that they would not have learned if the participant had opted out of participating.

3.1.2 Noise Mechanisms

This is guaranteed by applying noise to the result of an query to the dataset, by using the function $f()$.

Definition 3: A query f is a function that takes a dataset as input. The answer to the query f is denoted $f(D)$.

There are many different mechanisms for applying this noise, but the two most common are the Laplace mechanism and the Exponential mechanism.

Laplace Mechanism

The Laplace mechanism involves adding random noise which follows the Laplace statistical distribution. The most common question that needs to be answered before doing research with differentially private data, is how we should define our Laplace random variable, i.e how much noise needs to be added. The Laplace distribution centered around zero has only one parameter, its scale, and this is proportional to its standard deviation. The scale is naturally dependent on the privacy parameter ϵ, and also on the risk of the most different individual having their private information leaked from the data. This risk is called the sensitivity of the query, and is defined mathematically as:

$$\Delta f = \underset{D,D'}{max} ||f(D) - f(D')||_1$$

This equation states that the maximum difference in the values that the query f may take is on a pair of databases that differ on only one row. Dwork proved that adding a random Laplace variable, $(\Delta f/\epsilon)$, to a query you could guarantee ϵ-differential privacy[3].

Exponential Mechanism

The Exponential mechanism is a method for selecting one element from a set, and is commonly used if a non-numeric valued query is used. An example would be: ”What is the most common eye color in this room?”. Here it would not make sense to perturb the answer by adding noise drawn from the Laplace distribution. Instead we would use the aforementioned mechanism and make it exponentially more likely to make high quality outputs.

Definition 4: the sensitivity of score function H is defined as

$$s(H, ||.||) = \underset{(D,D)=1, a \in A}{max} ||H(D,a) - H(D',a)||$$

The exponential mechanism: given a dataset D and a set of possible answers A, if a random mechanism selects an answer based on the following probability, then the mechanism is -differentially private: P(a A is selected) e H(D,a)/2s(H,k.k)

Chapter 4

Experiment

Chapter 5

Analysis

Chapter 6

Conclusion

Bibliography

[1] Barth-Jones, D. C., 2012. The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (June 4, 2012).

[2] Brouwer, D. R., Jansen, J. D., 2004. Dynamic optimization of waterflooding with smart wells using optimal control theory. SPE Journal 9 (4), 391–402.

[3] Dwork, C., Roth, A., 2013. The algorithmic foundations of differential privacy. Theoretical Computer Science 9 (3-4), 211–407.

[4] Ganta, S. R., Kasiviswanathan, S. P., Smith, A., 2008. Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 265–273.

[5] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., 2007. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1), 3.

[6] Muskat, M., 1937. Flow of Homogeneous Fluids. McGraw Hill.

[7] Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, pp. 111–125.

[8] Sarma, P., Chen, W. H., 2008. Applications of optimal control theory for efficient production optimization of realistic reservoirs. In: Proceedings of the International Petroleum Technology Conference. Kuala Lumpur, Malaysia.

[9] Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (05), 557–570.

Appendix

Write your appendix here...